Novel Genomic Approaches for the Identification of Virulence Genes and Drug Targets in Pathogenic Bacteria

by

Junaid Gamieldien

UNIVERSITY of the
WESTERN CAPE

Thesis presented in fulfillment of the requirements for the Degree of Doctor Philosophiae at the South African National Bioinformatics Institute, Department of Biochemistry, Faculty of Natural Sciences, University of the Western Cape

December 2001

Supervisor: Prof. Winston Hide

DEDICATION

This thesis is dedicated to the loving memory of my father,

Mogamed Hashim Gamieldien,

who encouraged me to start this journey in the first place.

UNIVERSITY of the
WESTERN CAPE

ABSTRACT

While the many completely sequenced genomes of bacterial pathogens contain all the determinants of the host-pathogen interaction, and also every possible drug target and recombinant vaccine candidate, computational tools for selecting suitable candidates for further experimental analyses are limited to date. The overall objective of my PhD project was to attempt to design reusable systems that employ the two most important features of bacterial evolution, horizontal gene transfer and adaptive mutation, for the identification of potentially novel virulence-associated factors and possible drug targets.

In this dissertation, I report the development of two novel technologies that uncover novel virulence-associated factors and mechanisms employed by bacterial pathogens to effectively inhabit the host niche. More importantly, I illustrate that these technologies may present a reliable starting point for the development of screens for novel drug targets and vaccine candidates, significantly reducing the time for the development of novel therapeutic strategies.

Our initial analyses of proteins predicted from the preliminary genomic sequences released by the Sanger Center indicated that a significant number appeared to be more similar to eukaryotic proteins than to their bacterial orthologs. In order determine whether acquisition of genetic material from eukaryotes has played a role in the evolution of pathogenic bacteria, we developed a system that detects genes in a bacterial genome that have been acquired by interkingdom horizontal gene transfer.. Initially, 19 eukaryotic genes were identified in the genome of *Mycobacterium tuberculosis* of which 2 were later found in the genome of *Pseudomonas aeruginosa*, along with two novel eukaryotic

UNIVERSITY of the
WESTERN CAPE

genes. Surprisingly, six of the *M. tuberculosis* genes and all four eukaryotic genes in *P. aeruginosa* may be involved in modulating the host immune response through altering the steroid balance and the production of pro-inflammatory lipids.

We also compared the genome of the H37Rv *M. tuberculosis* strain to that of the CDC-1551 strain that was sequenced by TIGR and found that the organisms were virtually identical with respect to their gene content, and hypothesized that the differences in virulence may be due to evolved differences in shared genes, rather than the absence/presence of unique genes. Using this observation as rationale, we developed a system that compares the orthologous gene complements of two strains of a bacterial species and mines for genes that have undergone adaptive evolution as a means to identify possibly novel virulence –associated genes. By applying this system to the genome sequences of two strains of *Helicobacter pylori* and *Neisseria meningitidis*, we identified 41 and 44 genes that are under positive selection in these organisms, respectively. As approximately 50% of the genes encode known or potential virulence factors, the remaining genes may also be implicated in virulence or pathoadaptation. Furthermore, 21 *H. pylori* genes, none of which are classic virulence factors or associated with a pathogenicity island, were tested for a role in colonization by gene knockout experiments. Of these, 61% were found to be either essential, or involved in effective stomach colonization in a mouse infection model. A significant amount of strong circumstantial and empirical evidence is thus presented that finding genes under positive selection is a reliable method of identifying novel virulence-associated genes and promising leads for drug targets.

UNIVERSITY of the
WESTERN CAPE

DECLARATION

I declare that "*Novel Genomic Approaches for Identifying Virulence Genes and Drug Targets in Pathogenic Bacteria*" is my own work, that it has not been submitted for any degree or examination in any other university, and that all the resources I have or quoted have been indicated and acknowledged by complete references.

Junaid Gamieldien December 2001

UNIVERSITY of the
WESTERN CAPE

Signed

BIOGRAPHICAL SKETCH

Junaid Gamieldien was born in Cape Town on 19 December 1971. He attended Parkfields Primary School in Hanover Park and matriculated at Oaklands Secondary School in Lansdowne in 1988. Junaid originally registered for a degree in dentistry at the University of the Western Cape but became fascinated with Microbiology and Biochemistry, ultimately completing with a BSc degree in these disciplines in 1992, and graduating with a BSc Honours in Microbiology in 1994. He then started an MSc degree with the Esophageal Cancer Research group at UWC, graduating in 1996. He had also been working as a research assistant in the same year and met Winston Hide who introduced him to bioinformatics. Junaid had done first-year computer science as an extra course during his second year and although he was fascinated by the discipline, biology was his primary focus. The promise of being able to merge the two disciplines through bioinformatics prompted him to register for a PhD degree with Professor Hide at the then fledgling South African National Bioinformatics Institute at UWC in 1997. His research project was done completely *in-silico* and is titled “Novel Genomic Approaches for Identifying Virulence Genes and Drug Targets in Pathogenic Bacteria”.

UNIVERSITY of the WESTERN CAPE

PUBLICATIONS ARISING FROM THIS THESIS

Wagied Davids, Junaid Gamieldien, David A Liberles and Winston Hide. **Positive Selection Scanning Reveals Decoupling of Enzymatic Activities of Carbamoyl Phosphate Synthetase in *H. pylori*.** *Journal of Molecular Evolution*. In press.

Junaid Gamieldien, Andrey Ptitsyn, Winston Hide. **Eukaryotic genes in *Mycobacterium tuberculosis* could have a role in pathogenesis and immunomodulation.** *Trends in Genetics* 2002 18(1): 5-8.

UNIVERSITY of the
WESTERN CAPE

ACKNOWLEDGEMENTS

This study was performed in the South African National Bioinformatics Institute, University of Western Cape.

Special thanks to Prof. Winston Hide for having the courage to introduce bioinformatics to South Africa at a time when the discipline was still so young, and for sacrificing so much of your time and energy to make SANBI a world-leader in this highly competitive field. Your enthusiasm and drive that often goes far beyond the call of duty will be a hard act to follow, but will always have a positive impact on anything I do in the future.

To all my colleagues at SANBI: I would never have been able to achieve what I have without your support, technical and otherwise. Thanks to Alan Christoffels, Ekow Oppon, Andrey Ptitsyn and Chellapa Gopalakrishnan for making SANBI more than just a place of work – we've become good friends.

The guys from Electric Genetics, its been great working alongside you – I've learnt a lot from all of you. Also, thanks for improving the 'fun element' at SANBI.

UNIVERSITY of the
WESTERN CAPE

Thanks all my family and friends for helping me through the trying times and pushing me to keep going. You guys have often led by example.

A heartfelt thanks goes to my parents for your sacrifices and encouragement. Without your unwavering support, I would never have come this far.

To my wife, Kareemah: You and Shaqeel are a major driving force in my life. Thank you for always being there and for your words of encouragement through the thesis-writing process.

Last, but definitely not least, thanks to my Creator. All talents I have come from You, just as You are the fundamental source of all knowledge.

TABLE OF CONTENTS

UNIVERSITY of the
WESTERN CAPE

LIST OF FIGURES

UNIVERSITY of the
WESTERN CAPE

LIST OF TABLES

UNIVERSITY of the
WESTERN CAPE

APPENDIX I: GLOSSARY

accession number

An identifier supplied by the curators of the major biological databases upon submission of a novel entry that uniquely identifies that sequence (or other) entry.

adaptation

Any morphological, physiological or behavioral change that enhances survival, growth and the reproductive success. **Alternative**: Change in an organism resulting from natural selection; a structure that is the result of such selection.

alignment

The juxtaposition of amino acids or nucleotides in homologous molecules to maximize similarity or minimize the number of inferred changes among the sequences. Alignment is used to infer positional homology prior to or concurrent with phylogenetic analysis. **Alleles-** Alternate forms or varieties of a gene.

UNIVERSITY of the
WESTERN CAPE

ancestral Sequence

A hypothesized sequence, reconstructed from the relationships between contemporary sequences.

annotation

A combination of comments, notations, references, and citations, either in free format or utilizing a controlled vocabulary, that together describes all the experimental and inferred information about a gene or protein.

antigen

Any foreign molecule that stimulates an immune response in a vertebrate organism. Many antigens are proteins such as the surface proteins of foreign organisms.

bacteriophage

A virus that infects bacteria.

bootstrap

A method of attempting to estimate confidence levels of inferred relationships. The bootstrap proceeds by resampling the original data matrix with replacement of the characters. After the bootstrap procedure is finished, a majority-rule consensus tree is constructed from the optimal tree from each bootstrap sample. The bootstrap support for any internal branch is the number of times it was recovered during the bootstrapping procedure.

chromosome

A linear end-to-end arrangement of genes and other DNA, sometimes with associated protein and RNA. The form of the genetic material in viruses and cells. A circle of DNA in prokaryotes; a DNA or an RNA molecule in viruses; a linear nucleoprotein complex in eukaryotes.

UNIVERSITY of the WESTERN CAPE

clone

A population of genetically identical cells or DNA molecules.

coding regions

The portion of a genomic sequence bounded by start and stop codons that identifies the sequence of the protein being coded for by a particular gene.

codon

A group of three adjacent nucleotides that encode an amino acid.

codon preference

For amino acids with several codons one or a few are preferred and are used disproportionately.

competent

The ability to take up exogenous DNA and thereby be transformed.

congruence

Agreement between characters or trees.

conjugation

A process whereby two cells come in contact and exchange genetic material. In prokaryotes the transfer is a one-way process. In protozoa it is a two-way process, genetic material is passed between each conjugant.

UNIVERSITY of the
WESTERN CAPE

convergent evolution, convergence

The independent development of similar proteins in different groups; thought to be the result of similar environmental selection pressures on different groups.

duplication

A genetic event that results in a region of DNA being copied to another part of the genome.

DNA polymerase

An enzyme that catalyzes the synthesis of DNA from a DNA template given the deoxyribonucleotide precursors.

eukaryote

A cell or organism with a distinct membrane-bound nucleus as well as specialized membrane-based organelles (see also prokaryote).

evolution

Descent with modification, or change in the form, physiology, and behaviour of organisms over many generations. A general definition is that evolution is a continuous process of change in temporal perspective, long enough to produce a series of transformations.

frameshift

A deletion, substitution, or duplication of one or more bases that causes the reading-frame of a structural gene to shift from the normal series of triplets.

GenBank

A databank of genetic sequences operated by a division of the National Institutes of Health (USA).

UNIVERSITY of the
WESTERN CAPE

gene

The fundamental physical and functional unit of heredity, which carries information from one generation to the next.

gene product

The product, either RNA or protein, that results from expression of a gene.

gene tree

A branching diagram that depicts the known or (usually) inferred relationships among an historically related group of genes or other nucleotide or amino acid sequences.

genome

The complete genetic content of an organism.

genomics

The analysis of the entire genome of a chosen organism.

genotype

The genetic makeup of an individual. Genotype can refer to an organism's entire genetic makeup or the alleles at a particular locus.

Hidden Markov model (HMM)

A joint statistical model for an ordered sequence of variables. Hidden Markov models are valuable in bioinformatics because they allow a search or alignment algorithm to be trained using unaligned or unweighted input sequences; and because they allow position-dependent scoring parameters such as gap penalties, thus more accurately modeling the consequences of evolutionary events on sequence families.

UNIVERSITY of the
WESTERN CAPE

homolog

A gene or morphological character that shares a common ancestry with a different gene or morphological character.

housekeeping genes

Genes that are always expressed due to their constant requirement by the cell.

incongruence

Refers to different characters or trees suggesting different groups of relationships.

macroevolution

Major evolutionary changes in a population's gene pool, occurring over many generations, resulting in the evolution of new species.

Maximum Likelihood (ML)

A method of inferring phylogenetic relationships using a pre-specified (often user-specified) model of sequence evolution. Given a tree (a particular topology, with branch lengths), the ML process asks the question "What is the likelihood that this tree would have given rise to the observed data matrix, given the pre-specified model of sequence evolution?"

Maximum Parsimony (MP)

The principle of simpler solutions being preferred over more complex ones. In relation to phylogeny reconstruction this means that phylogenetic trees that can explain a data matrix by fewer evolutionary events is preferable over a tree that requires more evolutionary events.

UNIVERSITY of the
WESTERN CAPE

microevolution

Small changes in a population's gene pool occurring over a few generations. The accumulation of microevolutionary changes can result in macroevolution.

mismatch repair

A form of excision repair initiated at the sites of mismatched bases in DNA.

multiple sequence alignment

A matrix of data, in which the individual columns represent homologous characters.

mutability

The ability to change.

mutant

An organism or cell carrying a mutation. An alternative phenotype to the wild-type; the phenotype produced by a non wild-type allele.

mutation

An inheritable alteration to the genome that includes genetic (point or single base) changes, or larger scale alterations such as chromosomal deletions or rearrangements.

mutator mutation

Mutation of DNA polymerase that increases the overall mutation rate.

natural selection

An evolutionary mechanism that occurs when some individuals of a population are better able to adapt to their environment and, subsequently, produce more offspring.

UNIVERSITY of the
WESTERN CAPE

nonsynonymous substitutions

Those substitutions that change the identity of the encoded amino acid.

origin of replication

The point of specific sequence at which DNA replication is initiated.

ortholog

Genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. (See also Paralogs.)

paralog

Genes related by duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.

phenotype

The observable or detectable characteristics of an individual organism; the detectable expression of a genotype.

phylogeny

The historical relationships among lineages of organisms or their parts (e.g., genes).

point mutation

A mutation in which a single nucleotide in a DNA sequence is substituted by another nucleotide.

UNIVERSITY of the
WESTERN CAPE

polymorphism

The existence of a gene in a population in at least two different forms at a frequency far higher than that attributable to recurrent mutation alone.

Positive Darwinian Selection

This is a phenomenon whereby there is a selective pressure favoring change. It is usual to think of natural selection as a process of editing genetic change so that only a small number of mutational events are retained in a population. With positive selection, the retention of mutations is much closer to the rate of mutation. Detection of positive Darwinian selection is usually carried out by estimating the rate of synonymous to non-synonymous substitutions. If they are equal or if non-synonymous substitutions appear to be occurring more frequently than synonymous substitutions, then we can surmise that positive selection is acting on the protein sequence.

prokaryote

An organism or cell that lacks a membrane-bounded nucleus.

protein families

Sets of proteins that share a common evolutionary origin reflected by their relatedness in function, which is usually reflected by similarities in sequence, or in primary, secondary or tertiary structure. Subsets of proteins with related structure and function.

query (sequence)

A DNA, RNA of protein sequence used to search a sequence database in order to identify close or remote family members (homologs) of known function, or sequences with similar active sites or regions (analogs), from whom the function of the query may be deduced.

UNIVERSITY of the
WESTERN CAPE

selection

process which favors one feature of organisms in a population over another feature found in the population. This occurs through differential reproduction -- those with the favored feature produce more offspring than those with the other feature, such that they become a greater percentage of the population in the next generation.

similarity search

Detection of significant extended sequence similarity to a protein of known structure, or of a sequence pattern characteristic of a protein family.

speciation

The evolution of new species from older ones.

structural gene

Gene that encodes a structural protein.

synonymous substitutions

Those substitutions that do not change the identity of the encoded amino acid.

transformation

A genetic alteration to a cell as a result of the incorporation of DNA from a genetically different cell or virus; can also refer to the introduction of DNA into bacterial cells for genetic manipulation.

wild-type

Form of a gene or allele that is considered the "standard" or most common.

UNIVERSITY of the
WESTERN CAPE

CHAPTER 1

General Introduction and Literature Review

1.1. Background

The effort to completely sequence bacterial genomes is rapidly gaining momentum, with 54 projects completed and a further 240 underway at the time of writing, and genome sequence data will soon be available for most bacteria of medical significance. While preliminary analysis of these genome sequences has provided valuable insight into the biology of the targeted organisms and numerous new ‘classical’ virulence factors e.g. toxins, adhesins and invasins have been identified, many novel virulence-associated genes still await discovery. Furthermore, very little impact has been made on antibacterial drug discovery process to date.

UNIVERSITY of the WESTERN CAPE

This PhD project was initiated in 1997 and was intended to focus on an African problem that would be particularly relevant to South Africa. In the same year, the Sanger Center started releasing raw genomic sequences for a laboratory strain of *Mycobacterium tuberculosis* (H37Rv), and as South Africa has the highest incidence of tuberculosis worldwide, I decided to embark on a computational biology project focusing on tuberculosis. The initial specific objectives of this project were; assembly of the raw genomic sequences into contiguous segments (contigs), prediction of accurate open reading frames, and annotation of the predicted genes. This data would then be used for further analysis in an attempt to decipher the organism’s biology from its genetic makeup.

As part of the evaluation of the various gene-finding algorithms, sequence similarity searches were done against a complete non-redundant gene database to determine their efficiency in identifying known genes and their correct start and stop codons. Also, through comparison of the predicted H37Rv genes with those predicted from the preliminary genomic sequences provided by The Institute for Genomic Research (TIGR) for another *M. tuberculosis* strain (CDC-1551) that had been isolated during an outbreak in the United States, SANBI was one of the first institutions to suggest that *M. tuberculosis* may use an impressive repertoire of repeat-rich proteins as a means of immune evasion (unpublished observations).

A more interesting observation that arose from the similarity searches was that a significant number of the predicted proteins appeared to be more similar to eukaryotic proteins than to bacterial counterparts. Particularly intriguing was the fact that many of these proteins scored best against *Homo sapiens*, especially since there had been some speculation that the pathogen, due to its obligate intracellular lifestyle, may have acquired genes from its host (Valerie Mizrahi and Albert Beyers, private communications). While we acknowledged that microbial genomes were under-represented in the genetic databases at that point in time, we developed a system that identifies bacterial genes that are foreign to the bacterial kingdom from an evolutionary perspective, in order to ascertain whether acquisition of eukaryotic genes by played a role in the patho-adaptive evolution of *M. tuberculosis*.

UNIVERSITY of the
WESTERN CAPE

In order to compare two closely related strains of the same pathogen whose level of virulence may be quite different, TIGR opted to sequence the CDC-1551 strain, which was demonstrated to be more virulent in mice than the H37Rv strain. We found, however, that the organisms were virtually identical with respect to their gene content and hypothesized that the differences in virulence may be due to evolved differences in shared genes, rather than the absence/presence of unique genes. Using this observation as a rationale, we developed a system that compares the orthologous gene complements of two strains of a bacterial species and mines for genes that have undergone adaptive evolution as a means to identify possibly novel virulence –associated genes. While this system did not work well with *M. tuberculosis*, due to its high level of gene sequence conservation, it was successfully applied to two other organisms.

Insights gained from the preliminary analyses of the *M. tuberculosis* genome sequences have thus prompted the development of two computational strategies for identifying candidate virulence genes from genomic sequence data. These technologies detect two of the three most important evolutionary phenomena that enable benign bacteria to successfully invade and survive within the host niche, namely, acquisition of new genetic material and adaptive mutagenesis of existing genes. While gene loss, the third mechanism of evolution, plays an important role in the pathoadaptation of virulent bacteria, its occurrence is not readily detectable even when closely related species are compared as the converse case for horizontal gene transfer can always be argued.

UNIVERSITY of the
WESTERN CAPE

1.2. Evolution of bacterial virulence

Bacterial virulence appears to be a recently acquired, or purely accidental, secondary property that facilitates adaptation to a host environment that is novel from an evolutionary perspective (Jain *et al.* 1999). The phenotypic properties acquired during the evolution of pathogenic bacteria include the ability to:

- invade the host niche
- preferentially colonize a specific host organ/tissue i.e. strategies for tissue tropism
- effectively consume available nutrients
- evade host antibacterial defenses
- produce damage to the host

The adaptive evolution of benign bacteria into pathogens usually involves the acquisition of foreign genes encoding for specific virulence factors (Jain *et al.* 1999). This '*gain-of-function*' evolutionary mechanism has the capability of introducing fully functional, complex metabolic activities at the very moment of introduction of the foreign genetic material. An alternate or additional evolutionary mechanism employed by pathogenic bacteria is the subtle modification of an existing gene or genes through point mutations that confers a selective advantage in the virulence niche. These pathoadaptive (Sokurenko, 1999) or '*change-of-function*' point mutations have the ability to facilitate the directed evolution of pathogenic bacteria during their growth in diverse host environments.

UNIVERSITY of the WESTERN CAPE

In this chapter, the involvement of horizontal gene transfer and pathoadaptive mutations in the adaptive evolution of bacterial pathogens and computational methods for their detection will be further reviewed.

1.3. 'Macroevolution' - Horizontal Gene Transfer

The comparative analysis of a number of bacterial genome sequences has provided evidence that the horizontal transfer of genetic material between even evolutionarily disparate microbial lineages is a very common event. Evidence of horizontal gene transfer between prokaryotes (Koonin *et al.* 1997, Kroll *et al.* 1998, Lawrence and Ochman 1998, Woese 1998, Gogarten *et al.* 1999), from bacteria to eukaryotes (Bork 1993, Doolittle 1998), and even from eukaryotes to bacteria has been presented previously (Stephens *et al.* 1998, Wolf *et al.* 1999, Wolf *et al.* 2000, Gamieldien *et al.* in press). The process of horizontal gene transfer allows instantaneous delivery of fully functional, complex metabolic capabilities to a bacterial lineage, and thus represents a powerful mechanism by which the outcome of a bacteria-host interaction can be permanently altered. A number of studies have discovered the presence of horizontally acquired genomic segments, known as pathogenicity islands, that play a major role in the virulence processes of many bacterial pathogens and can convert a benign bacterium into a pathogen upon incorporation (reviewed by Groisman and Ochman 1996).

UNIVERSITY of the WESTERN CAPE

1.3.1. Mechanisms of horizontal gene transfer

In bacteria, the acquisition of foreign genetic material occurs by three major mechanisms:

i. **Conjugation** is the transfer of genetic material from one bacterial cell to another by means of specialized cell-to-cell contact and has been implicated in the rapid dissemination of antibiotic resistance genes among human and animal bacterial pathogens. Many antibiotic resistance genes are found on plasmids or transposons, which are the primary vehicles of gene transfer in bacteria. These can be disseminated among various bacteria by conjugation. In pathogenic bacteria, most plasmids contain genes that encode virulence factors, which include toxins, factors involved in immune evasion, and proteins that sabotage the host cellular machinery (Jain *et al.* 1999). These 'pathogenic' plasmids, in many cases, carry the backbone of genetic information that makes a given bacterium a pathogen. For example, the phenotypic changes that are induced by the Ti plasmid of *Agrobacterium* and the virulence plasmids of *Yersinia*, *Shigella*, and *Escherichia coli* are so drastic that, in essence, a 'new' bacterial species is created by their acquisition (reviewed by Groisman and Ochman 1996).

UNIVERSITY of the
WESTERN CAPE

ii. **Transduction** is the bacteriophage-mediated transfer of genes from one bacterium to another, which is limited to closely related species, due to the high degree of specificity of bacteriophage invasion. Genes introduced by bacteriophages are probably more likely to produce a more permanent adaptation to new environments since they are integrated into the host chromosome and mutation of the excision sequences may cause the phage to become ‘trapped’ in its host. The toxin of *Corynebacterium diptheriae*, for example, is carried by a temperate (dormant) phage that is integrated into the host chromosome (Hall and Collis 1995).

iii. **Natural Transformation** is the introduction of extracellular DNA into a ‘competent’ bacterium. This natural ability of many bacteria to takes up DNA is a major route for the horizontal spread of genes within the bacterial kingdom (reviewed by Dubnau 1999). Large segments of DNA can be taken up by the bacterium by this process and although mechanisms that limit integration of DNA from other species exist, the introduction and integration of some foreign DNA into the bacterial chromosome is inevitable. It has been observed that operational genes that are involved in housekeeping or supplementary functions e.g. pathogenesis are more likely to horizontally transferred than informational genes that are involved in replication, transcription, translation, etc (Jain *et al*. 1999).

UNIVERSITY of the
WESTERN CAPE

Molecular interactions between mobile virulence elements may lead to an ordered, stepwise progression of change. In *Vibrio cholera*, for example, the entry of the filamentous phage (*CTXf)* that carries the cholera toxin gene is dependent on the prior lysogenization of another phage that encodes the intestinal colonization factor which is the binding receptor for CTXϕ (Waldor and Mekalanos 1996).

1.3.2. Genomic islands and the origin of new bacterial species

A genomic island is a large region of foreign DNA that has been acquired by horizontal gene transfer into a bacterial genome. In pathogenic bacteria, these loci are called pathogenicity islands and are thought to be recent insertions into the chromosome and encode functions relevant for bacteria-host interactions. Many pathogenicity islands still show remnants of the mobile elements that led them there, which suggests that chromosomes are a genetic 'graveyard', or temporary residence, of previously mobile genes. Therefore, a significant proportion of the genome of any bacterium probably contains fragments of DNA from various origins, providing clues as to how new bacterial species, or quasi-species arise.

1.3.3. Examples of HGT in the evolution of pathogenic bacteria

1.3.3.1. Enteropathogenic *Escherichia coli* strains

UNIVERSITY of the
WESTERN CAPE

It has been predicted that 17% (~800kb) of the *E. coli* genome has been introduced by horizontal gene transfer during the past 100 million years (Lawrence and Ochman 1998). Similarly, a collection of *E. coli* strains has been found to vary greatly in the size and macro-organization of their chromosomes (reviewed by De La Cruz and Davies 2000). Recently, phylogenetic analysis of a number of disease-causing strains of *E. coli* has revealed evidence that the gain and loss of mobile virulence elements has occurred several times, frequently *in parallel* in separate lineages (Reid *et al*. 2000), which suggests that there is a selective advantage favoring the build-up of *specific* virulence that facilitates the establishment and transmission of new virulence clones. Furthermore, it is proposed that some lineages, due to defective mismatch repair systems, may have an enhanced ability to recombine and thus to acquire foreign DNA. The enteropathogenic *E. coli* strain, O157:H7, is known to contain more than 20 potential virulence genes clustered in several mobile elements: the large plasmid pO157, the lamboid temperate phage 933W, another prophage and the pathogenicity island LEE (reviewed by De la Cruz and Davies 2000). However, the genome of this strain is also 700kb larger than that of *E. coli* laboratory strain K-12, which suggests that the genome of O157:H7 contains additional regions of foreign origin that most likely harbor additional virulence genes. Acquisition of foreign genes by horizontal gene transfer is thus the major force driving the evolution of pathogenic *E. coli*.

UNIVERSITY of the
WESTERN CAPE

1.3.3.2. *Helicobacter pylori* and its pathogenicity island, *cag*

A review by Covacci *et al.* (1999) illustrates that *Helicobacter pylori* has evolved elaborate mechanisms to enable it to persist in the harsh acidic environment of the host stomach, where it causes acute and chronic inflammation. In 20 to 30% of cases, the infection may cause duodenal ulcer, gastric ulcers, adenocarcinoma of the distal stomach and gastric mucosa-associated lymphoid tissue (MALT) lymphoma. The major disease-associated, genetic difference in *H. pylori* isolates is the presence or absence of a pathogenicity island, named *cag*. The GC content of this 40-kb region is different than that of the rest of the genome, suggesting an evolutionarily recent acquisition event, and contains 31 genes that encode for a type IV secretion system. While *cag*+ strains induce visible gastric damage in a mouse model, *cag*- strains resemble commensal bacteria more than pathogens. Epidemiological studies indicate that 60 to 70% of human isolates are *cag*+, except in isolates from Korea and Japan where nearly 100% are *cag*+. Even more interesting is the observation that both cag-positive and -negative isolates with identical DNA fingerprints are found in nearly all patients. It appears that a dynamic equilibrium of *cag*-positive and -negative subpopulations exist, with alternating periods of disease and remission mediated by the overgrowth of *cag* positive and negative strains respectively. Excision of the pathogenicity-island is probably modulated by host factors, which may explain the low frequency of *cag*- isolates in Japan and Korea. This is a classic example of how a single DNA acquisition event can drive the instantaneous evolution of a primarily commensal bacterium to a quasispecies that has the ability to produce a large array of disease phenotypes.

UNIVERSITY of the
WESTERN CAPE

1.3.4. Methods for the detection of horizontal transfer events in genomic sequences

Horizontal gene transfers are difficult to prove and usually a combinatorial approach is taken to provide the necessary evidence. Three of the most commonly used methods for identifying horizontal gene transfer candidates are reviewed here.

1.3.4.1. Identification of regions with unusual base compositions

Genomes display significant compositional uniformity and as such, a foreign gene can often be detected by its unusual nucleotide composition or codon usage pattern (Mrazek and Karlin 1999). The major advantage of this approach is that it only requires the genomic sequence of the organism of interest. However, while the compositional method has been successfully used to identify pathogenicity islands in a number of bacterial pathogens, it has its limitations. The most obvious is that it is sensitive to date of transfer since in ancient transfers, the foreign DNA loses its original 'fingerprint' and conforms to the genome it resides in. Similarly, gene transfers between species with similar compositional bias will be overlooked. Furthermore, other native factors e.g. direction of transcription relative to the origin (Lafay *et al*. 1999) also cause differences in composition.

UNIVERSITY of the
WESTERN CAPE

1.3.4.2. Similarity Search based methods

Here, proteins from the organism of interest are searched against a protein database and the detected 'hits' are classified according to their taxonomic origin. Protein sequences that show stronger similarity to a distant taxon than to known close relatives, based upon a selected cutoff (score or expected value), are selected as candidates for horizontal transfer. Best-match methods, however, do not take into account evolutionary rate variation. True homologs in closely related taxa may thus not share a high degree of similarity well if they have evolved rapidly. Furthermore, multidomain proteins and gene loss can also mislead these best-match methods. The advantages of the similarity-based methods are their speed and automatability, which make them valuable as a ‘first-pass’ for the preliminary identification of candidates to be confirmed using more sophisticated and accurate methods.

1.3.4.3. Phylogenetic Incongruence

The most rigorous criteria for establishing the occurrence of ancient horizontal gene transfer involves the comparison of the phylogenetic tree constructed with regard to a specific protein or gene from a number of distantly related organisms with the known phylogeny of those species. If incongruence is seen between the gene tree and the known species tree, a strong case can be made for horizontal gene transfer, especially if sufficient statistical support can be demonstrated for the gene tree, and more than one method for the construction of the phylogenies is used. The gene tree should, however, be very similar to the expected tree, except for the unexpected placement of one member or group. The method is therefore, entirely dependent on the availability of gene or protein sequences from numerous and evolutionarily distant organisms for accurate construction of the ‘gene tree’. A similarity-based method was used in combination with phylogenetic incongruence to demonstrate the interkingdom transfer of eukaryotic genes to the genomes of Rickettsia prowazekii and Chlamydia trachomatis (Wolf et al. 1999).

UNIVERSITY of the
WESTERN CAPE

Eukaryote-to-prokaryote horizontal gene transfer was identified by use of a a combinatorial approach that employs a similarity-search based method as an initial screen together with rigorous phylogenetic testing.

1.4. Microevolution through pathoadaptive mutations

Many of the attributes that make certain bacterial lineages more virulent than others are conferred through virulence factors encoded by horizontally transferred foreign genes, pathogenicity can also be enhanced through the mutational change of existing genes. This evolutionary mechanism depends on genetic mutations that confer selective advantage to the bacterium in an otherwise hostile environment, and is especially important for the evolutionary success of those organisms that do not readily exchange DNA under natural conditions. For example, development of de novo resistance to the antibiotics streptomycin and isoniazid in Mycobacterium tuberculosis, to flouroquinones in Staphylococcus aureus and the development of an extended spectrum of β-lactamases in Gram-negative bacteria, is the result of single or multiple genetic mutations in existing genetic material (reviewed by Morris *et al*. 1998).

UNIVERSITY of the
WESTERN CAPE

The neutral theory of evolution postulates that most evolutionary change in proteins occurs due to neutral mutations and random genetic drift (Kimura 1977). On the other hand, Darwinian evolution is proposed to be due to selection for advantageous characteristics and a great deal of controversy exists regarding these hypotheses. Recently though, genomic-scale tests has presented substantial evidence that positive Darwinian selection has inflated the rate of protein evolution above what is expected from random genetic drift (reviewed by Fay and Wu 2001).

Mutation is an important stratagem for short-term adaptation of bacteria to adverse conditions and rapid adaptive evolution has been linked to numerous functions associated with virulence (Metzgar and Willis 2000). In virulent bacteria, these mutations can enable a pathogen to enter and/or survive in a virulence niche and have been referred to as pathoadaptive mutations (Sokurenko *et al.* 1999). Since many adaptive mutations may alter the original function of a gene, they can be expected to be detrimental to the bacterium in its ancestral niche, and will be selected against in the ancestral habitat; it is plausible that some clones will evolve a means to exist in both niches. Alternatively, if these bacterial pathogens somehow acquire a mechanism to translocate directly from one host to another, the dependence on the ancestral niche would be greatly reduced, or as in the case of obligate intracellular bacterial pathogens, lost completely. While gene acquisition and gene modification through point mutation are very different evolutionary mechanisms, they can be complementary. In essence, a specific virulence gene located in a pathogenicity island could mutate to become more effective at producing the pathogenic phenotype in a specific host, or a pathoadaptive mutation could be transferred among clones by recombination (Sokurenko *et al.* 1999).

UNIVERSITY of the WESTERN CAPE

1.4.1. Sources of mutations

Point mutations arise from random replication errors as well as through ineffective repair of DNA damage inflicted by endogenous and exogenous mutagens. Most bacteria, however, have a variety of DNA-repair mechanisms, which removes almost all of the DNA damage before replication. These mismatch repair systems that enforce genetic stability by suppressing recombination between imprecise homologies are, however metabolically expensive (Metzgar and Willis 2000), while mutation is essential for evolution and adaptation to diverse environments. The disruption or downregulation of these repair systems could therefore transform a clone into one that is hypermutable, with an increased potential of producing advantageous mutations under conditions of stress, although deleterious mutations may be fatal.

The genetic factors involved in DNA replication and repair may themselves be subject to mutation, selection and adaptive evolution (Wang *et al.* 1999), indicating that organisms may evolve mechanisms that optimize their mutation rates under specific environmental conditions. Selection acting upon mutator loci, and, in turn produces advantageous mutations in other loci important for survival, has been termed second-order selection (Weber 1996) by which organisms ‘evolve to evolve’. The genetic factors that increase an organism’s mutation rate can be divided into ‘global mutators’ and ‘contingency loci’, where the former causes a genome-wide change in types and rates of mutation while the latter affects specific loci (Field *et al.* 1999).

UNIVERSITY of the
WESTERN CAPE

i. **Global mutators** are associated with the DNA mismatch repair pathways and genetic changes affecting these factors can be heritable or transient (Rosenberg *et al.* 1998). In *Mycoplasma pneumoniae*, for example, the entire methyl-directed mismatch repair (MMR) system is absent and the organism shows genome-wide elevated rates of mutation when compared to closely related nonpathogenic species (Woese 1984). Similarly, the most virulent strains of *Yersinia* spp. intrinsically have the highest rates of mutation (Najdenski 1995, Iteman 1995) and it has been reported that some natural isolates of pathogenic *E. coli* and *Salmonella* spp. are mismatch repair deficient, implying a selection for mutability in the host niche (LeClerc 1996). Likewise, naturally occurring mutator strains of *Pseudomonas aeruginosa* have been demonstrated to be responsible for increased antibiotic resistance in lung infections of cystic fibrosis patients (Oliver *et al.* 2000). Hypermutable subpopulations have been demonstrated to exist within *E. coli* laboratory cultures during the stationary growth-phase (Torkelson *et al*. 1997) and it has been proposed that mismatch repair activity is not maintained at a constant level. Rather, a transient downregulation occurs in response to environmental signals, which increases genome-wide mutation rates (Radman 1999, Wang *et al*. 1999), which improves the chances of acquiring advantageous mutations, without compromising the fitness of future generations. Error-prone DNA polymerases IV and V, associated with the SOS system in *E. coli*, which has functional components similar to that of

UNIVERSITY of the WESTERN CAPE

the MMR system, have been shown to be inducible under stressful conditions and appear to introduce errors into DNA (Wagner *et al.* 1999). Furthermore *E. coli* strains that are naturally more mutable have been shown to outcompete low mutating strains *in vitro* (Chao and Cox 1983).

ii. **Contingency loci** are stretches of repetitive sequence, either simple repeats or microsatellites, within specific genes that are involved in producing a specific virulence-phase state or antigenic phenotype in pathogenic bacteria (Moxon 1994). These repetitive tracts are extremely prone to mutation, due to their increased propensity for causing 'slippages' during replication. These loci allow genes to be switched 'on' in one round of replication and 'off' in another due to frame-shifts (Levinson and Gutman 1987, Direnzo *et al*. 1994). While direct selection acts on the gene products altered by strand-slippage mutations, indirect second-order selection acts on the sequence characteristics of the contingency repeats themselves.

In this study, strategies for identifying genes that have undergone adaptive selection through the effects of global mutators, rather than contingency loci, were investigated with the aim of identifying novel virulence-associated genes.

UNIVERSITY of the WESTERN CAPE

1.4.2. Pathoadaptation in *Helicobacter pylori*

In a review by Wang *et al*. (1999), it was illustrated that *Helicobacter pylori* displays higher levels of allelic diversity than any other organism tested to date. It was also suggested that, while horizontal gene transfer has been a major force in its evolution, it appears as if the pathogen prefers de novo DNA mutation as its strategy to adapt to changing environments. The fact that all clinically observed antibiotic resistance was found to be due to mutations of chromosomal genes, rather than through the acquisition of foreign resistance genes via mobile genetic elements, even though the organism is naturally transformable, was also presented as evidence of the organism's preference for mutation in its adaptive evolution. Similarly, a study by Akopyants *et al*. (1995) demonstrated that a *H. pylori* strain that had initially grown weakly in the stomachs of gnotobiotic piglets rapidly adapted to growth in this environment. Genomic fingerprint analysis of the adapted derivative indicated that subtle adaptive mutations were the driving force for evolution.

UNIVERSITY of the
WESTERN CAPE

1.4.3. Statistical methods for identifying adaptive evolution in genes

Proteins are generally highly conserved and positive functional selection therefore has to be the major constraint by which amino acid substitutions occur. Detection of proteins that are under positive selection may enable the identification of novel virulence genes or mechanisms, as well as the genetic events that underlie the variable forms of disease caused by the same organism. Synonymous mutations do not cause amino acid substitutions while nonsynonymous substitutions do. As the number of synonymous substitutions per site can be assumed to be the background mutation rate in coding regions, genes that display a nonsynonymous substitution rate higher than background, can be assumed to be under positive selection (Kimura 1977). Although this is a very strict criterion for inferring positive selection as most proteins are slow evolving relative to the background rate, a large number of proteins have been reported to conform to this criteria (Endo *et al.* 1996).

UNIVERSITY of the WESTERN CAPE

The numerous methods for estimating the rates of synonymous (K_S) and nonsynonymous (K_A) substitutions can be divided into two main classes, *approximate* and *maximum-likelihood* methods. The latter methods, while considered to be more reliable as they employ, amongst others, explicit models of codon substitution and transition/transversion rate ratios, are however, computationally costly and often prohibitive for large analyses. Application on real data, however, has shown that one approximate method (Nei and Gojobori 1986) produces similar results as maximum-likelihood methods (Yang 1998), and we have thus opted for a modification of this method (Endo *et al* 1996). The chosen method also uses a windowing approach to identify specific gene/protein regions undergoing adaptive evolution and is a significant improvement since it avoids the scenario where locally elevated K_A would be 'masked' by globally higher K_S when an averaging method is used. In this study, interspecies comparisons were performed and the genes tested were thus evolutionarily very close, and very few substitutions, especially nonsynonymous ones, were expected.

1.5. Study Aims and Objectives:

The overall objective of this thesis was to determine whether the aforementioned evolutionary phenomena, i.e. horizontal gene transfer and adaptive evolution, could be applied to engineer reusable technologies that may aid in the identification of virulence processes employed by pathogenic bacteria and; to apply these systems to genomic sequences of selected pathogens elucidate novel virulence processes and possible drug targets.

UNIVERSITY of the
WESTERN CAPE

Specific primary objectives were:

A. To determine whether the obligate intracellular parasite, *Mycobacterium tuberculosis* has acquired genes from its human host or other eukaryotes via horizontal gene transfer.

B. To determine whether searching for genes of pathogenic organisms that are under positive selection may be a useful means of identifying novel virulence-associated genes and drug possible targets.

1.6. References:

Akopyants NS, Eaton KA, Berg DE (1995) Adaptive mutation and co-colonization during *Helicobacter pylori* infection of gnotobiotic piglets. *Infect Immun*. 63(1), 116-121.

Bork P (1993) Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile molecules that cross phyla horizontally? *Proteins.* 17, 363-374.

Chao L, Cox EC (1983) Competition between high and low mutating strains of *Escherichia coli*. *Evolution*. 37, 125-134.

Covacci A, Telford JL, Del Guidice G, Parsonnet J, Rappuoli R (1999) *Helicobacter pylori* virulence and genetic geography. *Science*. 284, 1328-1333.

De La Cruz F, Davies J (2000) Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol*. 8(3), 128-133.

DiRenzo A, Peterson AC, Garza JC, Valdez AM, Slatkin M, Freimer NB (1994) Mutational processes of sequence repeat loci in human populations. *Proc Natl Acad Sci USA*. 91, 3166-3170.

Doolittle W F (1998) You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends. Genet.* 14: 307-311.

UNIVERSITY of the
WESTERN CAPE

Dubnau D (1999) DNA uptake in bacteria. *Ann Rev Microbiol*, 53, 217-244.

Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. *Mol Biol Evol*. 13(5), 685-690.

Fay JC, Wu C (2001) The neutral theory in the genomic era. *Curr Opin Genet Dev*. 11, 642-646.

Field D, Mancuso MO, Moxon ER, Metzgar D, Tanaka MM, Wills C, Thaler DS (1999) Contingency loci, mutator alleles and their interactions. *Ann NY Acad Sci.* 870, 378-382.

Gogarten JP, Murphey RD, Olendzensky L (1999) Horizontal Gene Transfer: pitfalls and promises. *Biol Bull*. 196, 359-361.

Grantham R, Gautier C, Gouy M (1980) Codon frequencies of 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res*. 8, 1893-1912.

Groisman EA, Ochman H (1996) Pathogenicity islands: bacterial evolution in quantum leaps. *Cell*. 87, 791-794.

UNIVERSITY of the WESTERN CAPE

Hall RM, Collis CM (1995) Mobile genetic cassettes and integrons: capture and spread of genes by specific recombination. *Mol Microbiol.* 15, 593-600.

Iteman I, Nadjenski H, Carniel E (1995) High genomic polymorphism in *Yersinia pseudotuberculosis*. *Contrib Microbiol Immunol*. 13, 106-111.

Jain R, Rivera MC, Lake JA (1999) Horizontal transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA*. 181, 3801-3806.

Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*. 267, 275-276.

Koonin E, Mushegian AR, Walker DR (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin of the archaea. *Mol Microbiol*. 25, 619-637.

Kroll JS *et al*. (1998) Natural Genetic Exchange between *Haemophilus* and *Neisseria*: intergeneric transfer of chromosomal genes between major human pathogens. *Proc Natl Acad Sci USA*. 95, 12381-12383.

Lafay B, Lloyd AT, McClean MJ, Devine KM, Sharp PM, Wolf KH (1999) Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res.* 27, 1642-1649.

UNIVERSITY of the WESTERN CAPE

Lawrence JG, Ochman H (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci USA*. 95, 9413-9417.

LeClerc JE, Li B, Payne WL, Cebula TA (1996) High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. *Science*. 274, 1208-1210.

Levinson G, Gutman GA (1987) Slipped-strand mispairing a major mechanism for DNA sequence evolution. *Mol Biol Evol*. 4, 203-221.

Morris A, Kellner JD, Low DE (1998) The superbugs: evolution, dissemination and fitness. *Curr Opinion Microbiol*. 1, 524-529.

Moxon ER, Rainey PB, Nowak MA, Lenski RE (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr Biol*. 4(1), 24-33.

Mzarek J, Karlin S (1999) Detecting alien genes in bacterial genomes. *Ann NY Acad Sci*. 870, 314-329.

Nadjenski H, Iteman H, Carniel E (1995) The genome of *Yersinia enterocolytica* is the most stable of the three pathogenic species. *Contrib Microbiol Immunol*. 13, 281-284.

UNIVERSITY *of the* WESTERN CAPE

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*. 3, 418-426.

Oliver *et al.* (2000) High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. *Science* 288, 1251-1254.

Radman M (1999) Enzymes of evolutionary change. *Nature*. 401, 866-869.

Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whitham TS (2000) Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature*. 406, 64-67.

Rosenberg SM, Thulin C, Harris RS (1998) Transient and heritable mutators in adaptive evolution in the lab and in nature. *Genetics*. 148, 1559-1566.

Sokurenko EV, Hasty DL, Dykhuizen DE (1999) Pathoadaptive mutations: gene loss and variation in bacterial pathogens. *Trends Microbiol*. 7(5), 191-5.

Stephens RS *et al*. (1998) Genome Sequence of an Obligate Intracellular Pathogen of Humans: *Chlamydia trachomatis*. *Science*. 282, 754-759.

Wagner J, Gruz P, Kim S, Yamada M, Matsui K, Fuchs RPP, Nohmi T (1999) The *dinB* gene encodes a novel *E. coli* DNA polymerase, DNA polIV, involved in mutagenesis. *Mol Cell*. 4, 281-286.

UNIVERSITY *of the*
WESTERN CAPE

Waldor KM, Mekalanos JJ (1996) Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science*. 272, 1910-1914.

Wang G, Humayun MZ, Taylor DE (1999) Mutation as the origin of genetic variability in *Helicobacter pylori*. *Trends Microbiol.* 7,488-493.

Weber M (1996) Evolutionary plasticity in prokaryotes: a pan-glossian view. *Biol Phil*. 11, 67-88.

Weber M (1996) Evolutionary plasticity in prokaryotes: A panglossian view. *Biol Phil*. 11, 67 –88.

UNIVERSITY *of the*
WESTERN CAPE

Woese C (1998) The Universal Ancestor. *Proc Natl Acad Sci USA*. 95, 6854-6859.

Wolf YI *et al.* (1999) Rickettsiae and Chlamydiae, evidence of horizontal gene transfer and gene exchange. *Trends Genet.* 15, 173-175

Wolf Y.I. *et al.* (2000) Interkingom gene fusions. *Genome Biology* 1(6), research00131.1-00131.13

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*. 15, 568-573.

CHAPTER 2

Eukaryotic genes in the genome of *Mycobacterium tuberculosis* may play a role in pathogenesis and immunomodulation.

2.1. ABSTRACT

Acquisition of new genetic material through horizontal gene transfer has been an important feature in the evolution of many pathogenic bacteria. Here, we report the presence of 19 genes of eukaryotic origin in the genome of *Mycobacterium tuberculosis,* some of which may be unique to the *M. tuberculosis* complex. These genes, having been retained in the genome through selective advantage, most probably have key functions in the organism and in mammalian tuberculosis. We explore the role these genes might have in manipulation of the host immune system by altering the balance of steroid hormones.

UNIVERSITY of the WESTERN CAPE

2.2. INTRODUCTION

Mycobacterium tuberculosis is a gram–positive bacterium that causes approximately three million deaths annually and infects an estimated one third of the world's population, making it the most successful human pathogen. The primary site of infection is the lung, where it is initially ingested by pulmonary macrophages in the lower lobes and undergoes intracellular multiplication.

Many bacterial pathogens have evolved the capacity to produce virulence factors that are directly involved in infection and disease. Changes in the genetic repertoire, occurring through gene acquisition and deletion, are the major events underlying the emergence and evolution of bacterial pathogens (Ochman *et al*. 2001). While horizontal transfer of virulence determinants between bacteria is the most common mechanism for acquisition of new genetic material (Ziebuhr *et al.* 1999, Hacker *et al*. 2000, Morschauser *et al.* 2000, Ochman *et al.* 2000), the genomes of two obligate intracellular pathogens, *Chlamydia trachomatis* and *Rickettsia prowazekii*, have been reported to harbour a number of eukaryote-like virulence genes, of which a number are shared between the two organisms (Wolf *et al.* 1999). *M. tuberculosis* has recently been reported to have the highest number of eukaryotic-prokaryotic interkingdom gene fusions of all the sequenced bacterial genomes (Wolf *et al.* 2000). The individual 'fused' genes, however, do not offer immediate clues with respect to the organism's virulence mechanisms.

We have identified 19 eukaryotic-like genes in the genome of *M. tuberculosis* using a system similar to that reported by Wolf *et al.* (2000) and provide suggestive evidence that acquisition of these genes may have been a critical event in the evolution of ancient saprophytic mycobacteria to become

UNIVERSITY of the
WESTERN CAPE

pathogens, while proposing mechanisms by which the gene products may play a role in the development of the distinct disease phenotype produced by *Mycobacterium tuberculosis*.

2.3. MATERIALS AND METHODS

2.3.1. Sequence data

The proteins predicted from the genome sequence of *M. tuberculosis* H37Rv (as at 10 June 1998) were downloaded from the ftp site of the Sanger Center (ftp.sanger.ac.uk). The non-redundant GenBank protein dataset (as at 10 January 2001) was downloaded from the NCBI site (www.ncbi.nlm.nih.gov) and divided into eukaryotic and prokaryotic sub-databases to be used for comparative similarity searches.

UNIVERSITY of the WESTERN CAPE

2.3.2. Identification of candidate horizontally transferred genes by comparative similarity searching

In order to determine whether the acquisition of eukaryotic genes is an important feature in the evolutionary history of the pathogenic mycobacteria, we have developed a system of stepwise elimination that identifies eukaryotic-like genes in a bacterial genome. The system, similar to that employed by Wolf *et al.* (2000), compared *BLASTP* (Altschul *et al.* 1997) E-values for each *M. tuberculosis* predicted protein against bacterial and eukaryotic subsets of GenBank as a preliminary screen for horizontal transfer candidates. Proteins that scored higher against eukaryotes than bacteria, with at least 10 orders of magnitude difference in E-values, were selected as possible candidates for horizontal gene transfer. All mycobacterial protein sequences were removed from the bacterial subset, which would allow us to identify more ancient gene acquisitions, as well as those unique to *M. tuberculosis*.

2.3.3. Confirmation of absence of bacterial orthologs by more sensitive sequence similarity searches

Candidate proteins were compared against a complete non-redundant protein database using the NCBI *PSI-BLAST* (Altschul *et al.* 1997) search engine (www.ncbi.nlm.nih.gov/blast), to ensure that those proteins matching eukaryotic proteins exclusively in the first step truly had no bacterial orthologs. Those candidate proteins that matched eukaryotes exclusively were classified as putative horizontal transfers, as the more primitive members of the eukaryotic kingdom are poorly represented in the genome databases, while the genomes of many members of the bacterial and archaeal kingdoms have been completely sequenced.

UNIVERSITY of the
WESTERN CAPE

2.3.4. Phylogenetic analyses

In order to confirm horizontal transfer in those cases where bacterial sequences were identified by *PSI-BLAST*, representative protein sequences from the three kingdoms were aligned using *CLUSTALW* (Thompson *et al.* 1994) and subjected to phylogenetic analyses, using the neighbor-joining and protein-parsimony methods of the *PHYLIP* package (Felselstein 1996). Candidates that presented a non-congruent phylogeny, *i.e*. grouping within eukaryotic sequences, with bootstrapping support ≥ 70%, were classified as horizontal transfers, examples of which are shown in Figure 2.1 and Appendix I.

2.3.5. Predicting the order of acquisition

In order to determine whether a possible link exists between a stepwise acquisition of eukaryotic genes and the evolution of mycobacterial pathogenicity, we performed *BLAST* searches against the shotgun sequenced genome data of *M. bovis* (www.sanger.ac.uk); *M. avium* and *M. smegmatis* (www.tigr.org) to identify orthologous genes. The rationale used was that if the eukaryotic genes acquired by *M. tuberculosis* truly have a role in pathogenesis and human or mammalian tuberculosis, then *M. smegmatis* should have the least number of orthologs as it is primarily a saprophyte and only very mildly pathogenic. *M. avium* would, by the same assumption, have less orthologs than *M. bovis* since the latter is more virulent and causes a very similar disease in cattle as *M. tuberculosis* causes in human. Also, since none of the candidates that had been confirmed by phylogenetic analysis appears to have been acquired by *M. tuberculosis* from its human host, we expected all of these genes to be present in *M. bovis*.

UNIVERSITY of the WESTERN CAPE

2.4. RESULTS AND DISCUSSION

2.4.1. Acquisition of eukaryotic genes by *M. tuberculosis* may have occurred in a stepwise fashion

We have identified 19 *M. tuberculosis* genes (Table 2.1) that may have been acquired by horizontal gene transfer from eukaryotes. Of these, 8 could be confirmed by phylogenetic analyses while the remainder have no significant bacterial or archaeal orthologs, as confirmed by *PSI-BLAST* searches. All 19 candidates have homologs in *M. bovis*, while 7 may be evolutionarily recent acquisitions that play important roles in human and bovine tuberculosis, since they are not present in either *M. smegmatis* or *M. avium* (Table 2.1). A further 2 genes have no orthologs in *M. smegmatis* and may represent a key point in the evolution of the mycobacteria into mammalian pathogens. Because there is, however, no way of determining which eukaryotic species were the sources of the transferred genes, it is not possible to estimate the time of horizontal transfer using DNA sequence divergence as a measure.

UNIVERSITY of the
WESTERN CAPE

2.4.2. Are these eukaryotic genes involved in the biology and pathogenesis of *M. tuberculosis*?

Focusing on the group of genes involved in steroid metabolism (Table 2.1, A), we explore their possible roles in tuberculosis. Elimination of a *M. tuberculosis* lung infection requires a T helper 1 (Th1) cytokine balance, tumor necrosis factor alpha and activated macrophages. An inappropriate shift to a Th2 immune response, caused by subtle endocrinological changes brought about by, or in response to, the organism itself, is an important feature of pulmonary tuberculosis and promotes a necrotizing immunopathology that is accompanied by low mycobactericidal activity (Rook *et al.* 1994). It is interesting to note that two candidates, with apparently different eukaryotic ancestry, appear to encode type-IV 17-beta-hydroxysteroid dehydrogenases (17-beta-HSD-4) as determined by sequence similarity, suggesting that this functionality is under strong selective pressure. In mammals, this protein has a high affinity for estradiol and other sex steroids, which it transports into the cell and inactivates by catalyzing their oxidation (de Luanoit *et al.* 1999). Preliminary 2D-PAGE gel data made available by the Max Plank Institute for Infection Biology (www.mpiib-berlin.mpg.de/2D-PAGE) shows that one of these enzymes (Rv0148) is present in the cytoplasmic fraction of cultured *M. tuberculosis*. The mammalian 17-beta-HSD-4 is a multi-domain protein, which is N-terminally cleaved into a steroid dehydrogenase and a sterol transporter (de Luanoit *et al.* 1999). While all of the mycobacterial 17-beta-HSD-4's only have the steroid dehydrogenase domain, another of our candidate genes,

UNIVERSITY of the WESTERN CAPE

Rv2790c, is remarkably similar to experimentally confirmed eukaryotic non-specific sterol transporters (Figure 2.2), and is expressed *in vitro* (www.ssi.dk/en/forskning/tbimmun/Protein_database/protein_database.htm). The presence of both the dehydrogenase and the transporter functions suggests that the overall functionality of inactivating steroid hormones through 17-beta-HSD-4 activity is intact in the pathogenic mycobacteria. Only very weak orthologs of Rv2790c exist in Archaea and Bacteria, while the evidence for its acquisition by horizontal transfer from eukaryotes is very strong (Figure 2.1). Furthermore, only *M. avium* has a homolog to this gene, while *M. smegmatis* does not, which may be added evidence for a possible role in pathogenesis. In addition, another candidate, Rv1373, exclusive to the *M. tuberculosis* complex, is similar to many mammalian steroid-inactivating sulfotransferases which is further evidence that the overall functionality of inactivating host steroid hormones may be under strong positive selection pressure in the tubercle bacilli.

UNIVERSITY of the
WESTERN CAPE

2.4.3. What role could inactivation of host steroids play have tuberculosis?

A striking feature of tuberculosis is that an inappropriate and ineffective T helper 2 (Th2), rather than a protective Th1 immune response, is recruited against the tubercle bacillus. The Th1 immune response is sensitive to inactivation by glucocorticoids (GC), whereas antiglucocorticoids (sex steroids) like DHEA has been shown to be protective in a mouse TB model. It has been shown that the absence of endogenous estradiol is responsible for the development of *M. avium* pulmonary disease in postmenopausal women (Tsyuguchi *et al* 2001). Conversely, ovarian hormones positively affect macrophage antimycobacterial activity *in vitro* (Hernandez-Pando *et al.* 1998) and the formation of protective granulomas in the lung (Shirai *et al.* 1995). The group of horizontally acquired genes involved in steroid metabolism may play an important role in *M. tuberculosis* pathogenesis, driving the balance to higher levels of GC steroids through the importation and inactivation of host sex steroids, causing a sustained Th2 immune response and a prolonged infection.

UNIVERSITY of the
WESTERN CAPE

Table 2.1. *Mycobacterium tuberculosis* eukaryotic genes, grouped into functional classes, based on their putative co-involvement in metabolic pathways.

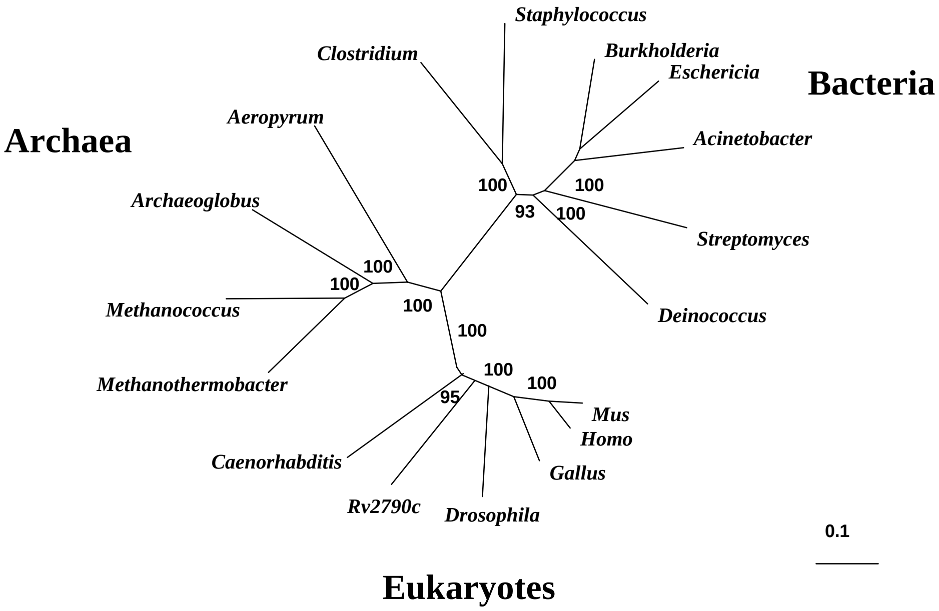
Gene name	Putative functional annotation	Support for horizontal gene transfer
A.Steroid metabolism		
Rv3548	17-beta hydroxyestradiol dehydrogenase	Phylogenetic analysis
Rv0148	17-beta hydroxyestradiol dehydrogenase*	Phylogenetic analysis
Rv1106c	steroid dehydrogenase*	Phylogenetic analysis
Rv2790c^{smegmatis}	sterol carrier protein*	Phylogenetic analysis
Rv0764c	cytochrome-P450 lanosterol demethylase	Phylogenetic analysis
Rv1373$^{avium, smegmatis}$	hydroxysteroid sulfotransferase	Only present in the tubercle bacilli and eukaryotes
B. Lipid metabolism		
Rv3451	cutinase	No bacterial orthologs.
Rv3452	cutinase	"
Rv1758	cutinase	"
Rv1984c	cutinase*	"
Rv2483c	lysophosphaditic acid acyltransferase	Phylogenetic analysis
Rv2590	long-chain-fatty-acid-CoA ligase	Phylogenetic analysis
C.Signal Transduction		
Rv0386$^{avium, smegmatis}$	adenylate cyclase	Only present in the tubercle bacilli and eukaryotes.
Rv0891c$^{avium, smegmatis}$	adenylate cyclase	"
Rv1358$^{avium, smegmatis}$	adenylate cyclase	"
Rv1359$^{avium, smegmatis}$	adenylate cyclase	"
Rv2488c$^{avium, smegmatis}$	adenylate cyclase	"
D. Other		
Rv1834$^{avium, smegmatis}$	Epoxide hydrolase	Phylogenetic analysis
Rv2577^{smegmatis}	purple acid-phosphatase	No orthologs in *M. smegmatis* and other bacteria.

UNIVERSITY of the WESTERN CAPE

* = Demonstrated to be expressed *in vitro* (see text)

^{avium} = No homolog in *M. avium*

^{smegmatis}=No homolog in *M. smegmatis*



UNIVERSITY of the
WESTERN CAPE

Figure 2.1. Phylogeny illustrating the clustering of a mycobacterial protein (Rv2790c) with the eukaryotic sterol transporters. The tree was constructed using the Neighbor Joining method of the PHYLIP package[6]. Numbers indicate the percentage of bootstrap replications out of 1000 replications, in which the given node was observed. Very similar tree topologies were obtained using maximum likelihood and parsimony-based methods. (Additional phylogenies in Appendix I).

```
 Mus                 ----MPSVALKSPRLRRVFVVGVGMTKFMKPGGENSRDYPDMAKEAGQKALEDAQIPYSA
Homo                ----MSSSPWEPATLRRVFVVGVGMTKFVKPGAENSRDYPDLAEEAGKKALADAQIPYSA
Gallus              ARVCEERGGPAAIMQRRVFVVGVGMTKFAKP-SENSVDYPDLAKEAGQKALADAGIPYSA
Rv2790c             --------MPNQGSSNKVYVIGVGMTKFEKPGRREGWDYPDMARESGTKALRDAGIDYRE
Caenorhabditis      ----------MTPTKPKVYIVGVGMTKFCKPGSVPGWDYPDMVKEAVTTALDDCKMKYSD
                                    :*:::******* **    . ****:..*:  .** *. : *

Mus                 VEQACVGYVYGDSTSGQRAIYHSLGLTGIPIINVNNNCSTGSTALFMAHQLIQGGLANCV
Homo                VDQACVGYVFGDSTCGQRAIYHSLGMTGIPIINVNNNCATGSTALFMARQLIQGGVAECV
Gallus              VEQACVGYVYGDSTCGQRAIYHGLGLTGIPIINVNNNCATGSTALFMSRQLVEGGLADCV
Rv2790c             VEQGYVGYVYGESTSGQRALYELG-MTGIPIVNVNNNCSTGSTALYLGAQAIRGGLADCV
Caenorhabditis      IQQATVGYLFGGTCCGQRALYEVG-LTGIPIFNVNNACASGSSGLFLGKQIIESGNSDVV
                    ::*. ***::* : .****:*.   :*****.**** *::**:.*::. * :..* :: *

Mus                 LALGFEKMERGSIG---TKFSDRTTPTDKHIEVLIDKYGLSAHPITPQMFGYAGKEHMEK
Homo                LALGFEKMSKGSLG---IKFSDRTIPTDKHVDLLINKYGLSAHPVAPQMFGYAGKEHMEK
Gallus              LALGFERMAKGSLA---SGFSDRTNPMDKHLEIMINKYGLASAPITPQMFANAGKEHMEK
Rv2790c             LALGFEKMQPGALG---GGADDRESPLGRHVKALAEIDEFGFP-VAPWMFGAAGREHMKK
Caenorhabditis      LCAGFERMAPGSLENLAAPIDDRALSVDKHISVMSETYGLEPAPMTAQMFGNAAKEHMEK
                    *. ***:*  *::       .**  . .:*:. : :   :    ::. **. *.:***:*

Mus                 YGTKVEHFAKIGWKNHKHSVNNTYSQFQDEYSLEEVMKSKPVFDFLTILQCCPTSDGAAA
Homo                YGTKIEHFAKIGWKNHKHSVNNPYSQFQDEYSLDEVMASKEVFDFLTILQCCPTSDGAAA
Gallus              YGTNPEYFAKIAWKNHSHSTNNPYSQFQKKYTLDEVLQSRKVFDFLTVLQCCPTSNGAAA
Rv2790c             YGTTAEHFAKIGYKNHKHSVNNPYAQFQDEYTLDDILASKMISDPLTKLQCSPTSDGSAA
Caenorhabditis      YGSKREHYAKIAYKNHLHSVHNPKSQFTKEFSLDQVINARKIYDFMGLLECSPTSDGAAA
                    **:. *::***.:*** **.:*. :** .:::*:::: :: : * :  *:*.***:*:**

Mus                 AILSSEEFVQQYG-LQSKAVEIVAQEMMTDLPSTFEEKSIIKVVGYDMSKEAARRCYEKS
Homo                AILASEAFVQKYG-LQSKAVEILAQEMMTDLPSSFEEKSIIKMVGFDMSKEAARKCYEKS
Gallus              AILASEDFVKRHK-LQPQAVEILAQVMATDYPSTFEENSCMKMVGYDMTKKAAEKCFKKA
Rv2790c             VVLASEDYLANHN-LAGRAVEIVGQAMTTDFASTFDG-SARNIIGYDMTVQAAQRVYQQS
Caenorhabditis      AVLVSEKFLEKNPRLKAQAVEIVGLKLGTDEPSVFAENSNIKMIGFDMIQKLAKQLWAET
                    .:* ** :: .   *  :****:.  : ** .* *   *  :::*:**  : *.: : ::

Mus                 GLTPNDVDVIELHDCFSVNELITYEALGLCPEGQGGTLVDRGDNTYGGKWVINPSGGLIS
Homo                GLTPNDIDVIELHDCFSTNELLTYEALGLCPEGQGATLVDRGDNTYGGKWVINPSGGLIS
Gallus              GLKPTDVDVIELHDCFSVNEFITYEALGLCPEGKACDLIDRGDNTYGGKWVINPSGGLIS
Rv2790c             GLGPKDFGVIELHDCFSANELLLYEALGLCGPGEAPELIDDNQTTYGGRWVVNPSGGLIS
Caenorhabditis      KLTPNDVQVIELHDCFAPNELITYEAIGLCPVGQGHHIVDRNDNTYGGKWVINPSGGLIS
                     * *.*. ********: **:: ***:***  *:.  ::* .:.****:**:********

Mus                 KGHPLGATGLAQCAELCWQLRGEAGKRQVPGAKVALQHNLGLGGAVVVTLYRMGFPEAAS
Homo                KGHPLGATGLAQCAELCWQLRGEAGKRQVPGAKVALQHNLGIGGAVVVTLYKMGFPEAAS
Gallus              KGHPLGATGLAQSAELCWQLRGLAGRREVGGARRALQHNLGLGGAVVVTLYAMGFPGAAS
Rv2790c             KGHPLGATGLAQCAELTWQLRGTAEARQVDNVTAALQHNIGLGGAAVVTAYQRAER----
Caenorhabditis      KGHPIGATGVAQAVELSNQLRGKCGKRQVPNCKVAMQHNIGIGGAGVVGLYRLGFPGAAQ
                    ****:****:**..**  ****  .  *:*  .   *:***:*:*** **  *  .
```

UNIVERSITY of the
WESTERN CAPE

Figure 2.2. CLUSTALW multiple sequence alignment, illustrating the high degree (>63%) of similarity of Rv2790, a predicted mycobacterial sterol transporter, to mouse (AK004860), human (AAB41286), chicken (AAA02488) and nematode (AL023847) non-specific sterol transporters. Identical amino acids are represented by an asterisk (*), conservative substitutions by a colon (:) and semiconservative substitutions by a period (.).

2.5. CONCLUSION

Our findings suggest that a sequential acquisition of eukaryotic genes is an important feature in the evolutionary history of the pathogenic mycobacteria. The genes described here, having been retained in the genome through selective advantage, most likely play a key role in the biology of *M. tuberculosis*. The strong selection for some of the functions that these transferred genes encode may provide clues regarding virulence of the bacterium and the inappropriate and impotent immune response to *Mycobacterium tuberculosis* lung infections.

UNIVERSITY of the
WESTERN CAPE

2.6. REFERENCES

Altschul S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25, 3389-3402.

de Luanoit Y. *et al.* (1999) Unique multifunctional HSDb4 gene product: 17-β-hydroxysteroid dehydrogenase and D-3-hydroxyacyl coenzyme A dehydrogenase/hydratase involved in Zellweger syndrome. *J Mol Endocrinol*. 22, 227-240.

Felsenstein J. (1996) Inferring phylogenies from protein sequences by parsimony, distance and likelyhood methods. *Methods Enzymol*. 266, 418-427.

Hacker J. *et al.* (2000) Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol*. 54, 641-79.

Hernandez-Pando R. *et al.* (1998) The effects of androstenediol and dehydroepiandosterone on the course and profile of tuberculosis in BALB/c mice. *Immunology*. 95(2), 234-241.

Morschhauser J. (2000) Evolution of microbial pathogens. *Philos. Trans. R. Soc. Lond. B. Biol. Sci*. 355(1397), 695-704.

UNIVERSITY of the
WESTERN CAPE

Ochman H. *et al.* (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*. 405(6784), 299-304.

Ochman H. *et al.* (2001) Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science*. 292(5519), 1096-1099.

Rook, G.A. *et al*. (1994) T cell helper types and endocrines in the regulation of tissue-damaging mechanisms in tuberculosis. *Immunobiology*. 191(4-5), 478-492.

Shirai M. *et al.* (1995) The influence of ovarian hormones on the granulomatous inflammatory process in the rat lung. *Eur Respir J.* 8(2), 272-7.

Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673-4680.

Tsyuguchi K. *et al.* (2001) Effect of oestrogen on *Mycobacterium avium* complex pulmonary infection in mice. *Clin Exp Immun.* 123(3), 428-434.

Wolf Y.I. *et al.* (1999) Rickettsiae and Chlamydiae, evidence of horizontal gene transfer and gene exchange. *Trends Genet.* 15, 173-175.

Wolf Y.I. *et al.* (2000) Interkingom gene fusions. *Genome Biology* 1(6), research00131.1-00131.13.

UNIVERSITY *of the*
WESTERN CAPE

Ziebuhr W. *et al.* (1999) Evolution of bacterial pathogenesis. *Cell. Mol. Life Sci.* 56(9-10), 719-28.

UNIVERSITY of the
WESTERN CAPE

CHAPTER 3

Similarities between the immunopathology of *Mycobacterium tuberculosis* and *Pseudomonas aeruginosa* lung infections may be partially due to shared genes of eukaryotic ancestry

3.1. ABSTRACT

We have previously reported the presence of eukaryotic-like genes in the genome of *Mycobacterium tuberculosis* that may play a role in pathogenesis through modulation of the host immune reponse. Since lung infections by *M. tuberculosis* and *Pseudomonas aeruginosa* lead to the development of similar immunopathological profiles, we performed an analysis to determine whether the organisms share eukaryotic-like genes. Using a system that compares BLAST2 scores for bacterial and eukaryotic datasets against proteins of interest, we have identified two novel eukaryotic-like genes in the genome of *P. aeruginosa*, as well as two eukaryotic-like genes shared between *M. tuberculosis* and *P. aeruginosa*. We provide suggestive evidence that these genes may play a role in the development of the distinct immunopathology produced by lung infections by these organisms.

UNIVERSITY of the
WESTERN CAPE

3.2. INTRODUCTION

Pseudomonas aeruginosa is a Gram-negative bacterium that may be described as the epitome of an opportunistic human pathogen. It is the major cause of mortality in cystic fibrosis (CF) patients and since it is naturally resistant to a wide array of antibiotics and disinfectants, infections are very difficult to eradicate. Although a T helper 1 cytokine balance may be beneficial in the eradication of a *P. aeruginosa* lung infection, the immune response in CF patients is predominantly of a counter-productive Th2 type (Moser et al. 1997, Moser et al. 1999, Moser et al. 2000). The organism has an exceptionally large genome and is genetically complex, which has been proposed to reflect evolutionary adaptations, most likely via horizontal gene transfer, that allow the organism to thrive in diverse ecological niches (Stover et al. 2000).

UNIVERSITY of the WESTERN CAPE

Mycobacterium tuberculosis is a Gram–positive bacterium that causes approximately three million deaths annually and infects an estimated one third of the world's population, making it the most successful human pathogen. The primary site of infection is the lung, where it initially is ingested by pulmonary macrophages in the lower lobes. The bacterium multiplies intracellularly and the pathogen is highly dependent on its ability to infect and survive within host macrophages. Activation of a protective cytokine response, however, can stimulate infected macrophages to kill its resident pathogens. As with *P. aeruginosa* lung infections, elimination of a *M. tuberculosis* infection also requires a Th1 cytokine balance, while Th2 dominance greatly enhances susceptibility to TB (Beyers *et al.* 1998). In mice, even a minor Th2 component abrogates immunity and leads to an immunopathology that mimics human tuberculosis (Rook *et al.* 1987). We have previously identified a number of eukaryotic-like genes in the genome of *M. tuberculosis* that may be involved in modulation of the host immune system (Gamieldien *et al.*, 2002). Due to the similar immunological profiles elicited by the host in response to lung infections by these two pathogens, we have performed a study to determine whether any of these genes are shared between *M. tuberculosis* and P. *aeruginosa*.

The discovery of 2 unique eukaryotic-like gene in the genome of *P. aeruginosa* involved in the production of proinflammatory lipid metabolites, as well as 2 eukaryotic-like genes that are shared with *M. tuberculosis* is reported here.

3.3. MATERIALS AND METHODS

3.3.1 Protein sequences

The proteins predicted from the genome sequences of *M. tuberculosis* H3Rv and *P. aeruginosa* PAO1 were downloaded from the ftp sites of the Sanger Center (ftp.sanger.ac.uk, as at 10 June 1998) and the Institute for Genome Research (ftp.tigr.org, as at 30 September 2000), respectively. Non-redundant bacterial and eukaryotic protein subsets of GenBank (as at 10 January 2001) were downloaded from the website of the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov).

3.3.2. Identification of horizontal transfer candidates

In order to identify eukaryotic genes that are shared by *P. aeruginosa* and *M. tuberculosis*, we have used a modification of our previously described system (Gamieldien *et al.*, 2002; Chapter 2, this document), which compares BLAST2

UNIVERSITY of the
WESTERN CAPE

(Altschul *et al.* 1997) search-result scores, obtained for bacterial and eukaryotic subsets of GenBank against each predicted protein of the organisms of interest. All mycobacterial and *P. aeruginosa* sequences were removed from the bacterial dataset to facilitate the identification of eukaryotic-like genes that are unique to, and shared by, the organisms of interest. Proteins that scored higher against eukaryotes than bacteria, with at least 10 orders of magnitude difference in E-values, were selected as possible candidates for horizontal gene transfer. Candidate proteins were compared against a complete non-redundant protein database using the NCBI *PSI-BLAST* (Altschul *et al.* 1997) search engine (www.ncbi.nlm.nih.gov/blast), those candidates proteins that matched eukaryotes exclusively were classified as putative horizontal transfers. In order to confirm horizontal transfer in those cases where bacterial sequences were identified by *PSI-BLAST*, representative protein sequences from the three kingdoms were aligned using *CLUSTALW* (Thompson *et al.* 1994) and subjected to phylogenetic analyses, using the neighbor-joining and protein-parsimony methods of the *PHYLIP* package (Felselstein 1996). Candidates that presented a non-congruent phylogeny, *i.e.* grouping within eukaryotic sequences, with bootstrapping support ≥ 70%, were classified as horizontal transfers (Figure 3.1).

UNIVERSITY of the WESTERN CAPE

3.4. RESULTS AND DISCUSSION

3.4.1. Shared eukaryotic genes were most likely first acquired by mycobacteria with subsequent transfer to *P. aeruginosa*

We have identified 2 unique eukaryotic-like gene in the genome of *P. aeruginosa* and also 2 eukaryotic-like genes that are shared with *M. tuberculosis* (Table 3.1). The shared genes were most likely acquired via horizontal gene transfer from eukaryotes by an ancient mycobacterium and then transferred to *P. aeruginosa*, since the genes are present in *M. bovis, M. avium, M. paratuberculosis* and *M. smegmatis* and are not present in any other actinomycetes. The eukaryotic genes that are unique to *P. aeruginosa* were probably acquired independently.

UNIVERSITY of the
WESTERN CAPE

3.4.2. Shared eukaryotic-like genes may be involved in immunomodulation

A striking feature of tuberculosis and CF patients with chronic *P. aeruginosa* infections is the recruitment of an inappropriate T helper 2 (Th2) immune response. A Th1 response, along with activated macrophages and TNF-alpha, is required for effective elimination of both these infections. The Th1 immune response is sensitive to inactivation by glucocorticoids (GC), whereas antiglucocorticoids (anti-GC) like dehydroepiandosterone (DHEA) has been shown to be protective in a mouse TB model (Hernandez-Pando *et al.* 1998) and in *P. aeruginosa* infections in mice (Ben-Nathan *et al.* 1999). An increase in active TB has also been observed in rheumatic patients on moderate to high doses of corticosteroids (Kim *et al.* 1998), as well as in asthma patients treated with inhaled corticosteroids (Shaikh 1992), while methylprednisolone has similarly been demonstrated to negatively effect the phagocytic and bactericidal activities of human granulocytes against *P. aeruginosa* (Baltch *et al.* 1986). The activity encoded by the shared eukaryotic 17-beta-hydroxyestradiol dehydrogenase type IV (Rv0148/PA3425) has previously been shown to exist in *P. aeruginosa* (Rowland *et al.* 1992) and binds estradiol, estrone, dihydrotestosterone, estriol, testosterone, progesterone and promegestone, in order of decreasing specificity. The enzyme may play a central role in endocrine manipulation in both organisms, driving the steroid hormone balance to higher levels of GC steroids through the importation and inactivation of host sex steroids, leading to a switch from a Th1 to Th2 immune response. We have previously shown that this functionality appears to be under strong positive selection in *M. tuberculosis*, since the organism has acquired 2 genes from apparently different eukaryotic sources as well as a sterol transporter (Rv2790c) that has a surprisingly high level of sequence identity to its mammalian counterparts (Gamieldien *et al.*, in press; Chapter 2, this manuscript).

UNIVERSITY of the
WESTERN CAPE

3.4.3. Unique eukaryotic-like *P. aeruginosa* genes may be responsible for extreme inflammation and persistence in lung infections

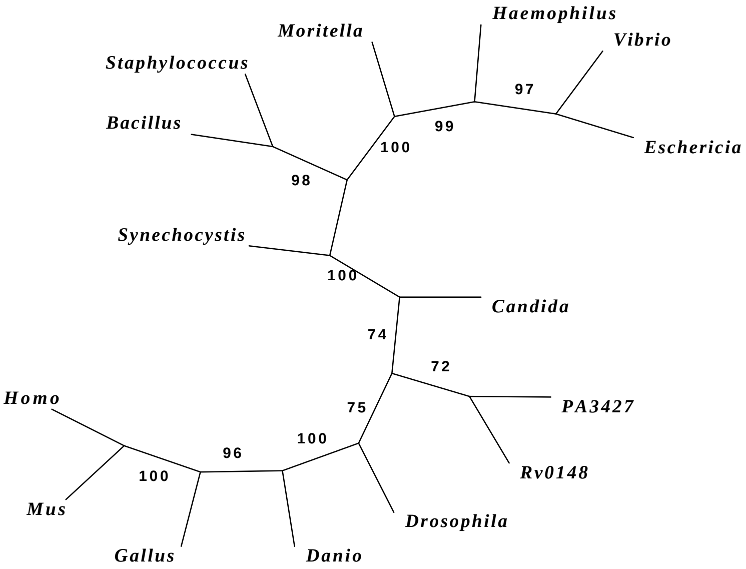
In addition to the eukaryotic genes shared with *M. tuberculosis*, we have identified a eukaryotic-like arachidonate lipoxygenase (PA1169) and a phospholipase D (PA3487) that is unique to *P. aeruginosa*. PA1169 may further promote the deleterious inflammatory response in CF patients through the production of a leukotriene, a potent stimulator for the recruitment of inflammatory cells. Increased urinary levels of leukotrine B4 have been observed in cystic fibrosis patients with chronic *P. aeruginosa* lung infections (Greally et al. 1994) and leukotriene B4 has been shown to have a pathogenic role in the lung damage in CF (Muller and Sorrell 1992). There is also evidence that *P. aeruginosa* metabolizes phagocyte-derived arachidonate during phagocytosis and modulates the potential pro-inflammatory effects of arachidonate metabolites (Sorrel et al. 1992). Furthermore, *P. aeruginosa* produces pyocyanin, an inhibitor of the omega-oxidation of leukotriene B4 (Muller and Sorrel 1992), which further supports the potential importance of the eukaryotic arachidonate lipoxygenase in pathogenesis and the long-term inflammation seen in cystic fibrosis. Gene knockout experiments have recently implicated the phospholipase D gene (PA3487) in the ability of *P. aeruginosa* to persist in chronic lung infections (Wilderman *et al.* 2001), which is tangible evidence that the acquisition of eukaryotic genes played an important role in the evolution of the organism's virulence mechanisms.

UNIVERSITY of the WESTERN CAPE

Table 3.1. *M. tuberculosis* and *P. aeruginosa* eukaryotic-like genes grouped into classes, based on their co-involvement in metabolic pathways. The original published gene names are used e.g. Rv3458 for *M. tuberculosis*, and PA0001 for *P. aeruginosa*. The best eukaryotic and prokaryotic match for each gene as determined by BLAST2, are also displayed. Genes present in both organisms are shown in boldface.

Gene name	**Best Eukaryotic match**	**Best Bacterial Match**	**Predicted functional annotation**
Rv0148/PA3427	*Sus* ($8E^{-81}$)	*Caulobacter* ($1E^{-81}$)	**17-beta hydroxy-estradiol dehydrogenase**
Rv1834/PA3429	*Danio* ($2E^{-35}$)	*Synechocystis* ($3E^{-06}$)	**epoxide hydrolase**
PA1169	*Bos* ($6E^{-59}$)	*None*	arachidonate 15-lipoxygenase
PA3487	*Homo* ($1E^{-26}$)	*None*	phospholipase D

UNIVERSITY of the WESTERN CAPE



UNIVERSITY of the
WESTERN CAPE

Figure 3.1. Phylogeny illustrating the co-clustering of *M. tuberculosis* and *P. aeruginosa* proteins, Rv0148 and PA3427 respectively, with the eukaryotic 17-β-hydroxysteroid dehydrogenases. The tree was constructed using the Neighbor Joining method. Numbers indicate the percentage of bootstrap replications out of 1000 replications, in which the given node was observed.

3.5. CONCLUSIONS

Our findings suggest that the acquisition of eukaryotic genes have been important events in the evolution of virulence of both *Mycobacterium tuberculosis* and *Pseudomonas aeruginosa*. The acquired genes may represent novel markers for important processes in virulence, which can be tested in the laboratory, and possible clues as to the virulence mechanisms employed by the bacteria, especially with respect to the inefficient and destructive immune reaction elicited in response to infection by *M. tuberculosis* and *P. aeruginosa*.

UNIVERSITY of the
WESTERN CAPE

3.6. REFERENCES

Altschul S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25, 3389-3402

Baltch Al, Hammer AC, Smith RP, Bishop MB, Sutphen NT, Egy MA, Michelsen PB (1986) Comparison of the effect of three adrenal corticosteroids on human granlulocyte function against *Pseudomonas aeruginosa*. *J Trauma*. 26(6), 525-533.

Ben-Nathan D, Padgett DA, Loria RM (1999) Androstenediol and dehydroepiandrosreone protect mice against lethal bacterial infections and lipopolysaccharide toxicity.*J Med Microbiol*. 48(5), 425-31.

Beyers A D, Van Rie A, Adams J, Fenhalls G, Gie R, Beyers N (1998) Signals that regulate the host response to *Mycobacterium tuberculosis*. *Novartis Found Symp*. 217, 145-159.

Felsenstein J (1996) Inferring phylogenies from protein sequences by parsimony, distance and likelyhood methods. *Methods Enzymol.* 266, 418-427.

UNIVERSITY of the
WESTERN CAPE

Gamieldien J, Ptitsyn A, Hide W (2002) Eukaryotic genes in *Mycobacterium tuberculosis* could have a role in pathogenesis and immunomodulation. *Trends Genet.* 18(1), 5-8.

Greally P, Cook AJ, Sampson AP, Coleman R, Chambers S, Piper PJ, Price JF (1994) Atopic children with cystic fibrosis have increased urinary leukotriene E4 concentrations and more severe pulmonary disease. *J Allergy Clin Immunol*. 93(1), 100-107.

Hernandez-Pando R, De La Luz Streber M, Orozco, H, Arriaga K, Pavon L, Al-Nakhli SA, and Rook GA (1998) The effects of androstenediol and dehydroepiandosterone on the course and profile of tuberculosis in BALB/c mice. *Immunology*. 95(2), 234-241.

UNIVERSITY *of the* WESTERN CAPE

Kim HA, Yoo CD, Baek HJ, Lee EB, Ahn C, Han JS, Kim S, Lee JS, Choe KW, Song YW (1998) *Mycobacterium tuberculosis* infection in a corticosteroid-treated rheumatic disease patient population. *Clin Exp Reumatol*. 16(1), 9-13.

Moser C, Hougen HP, Song Z, Rygaard J, Kharazami A, Hoiby N (1999) Early immune response in susceptible and resistant mice strains with chronic *Pseudomonas aeruginosa* lung infection determines the T-helper-cell response. *APMIS*. 107(12), 1093-1100.

Moser C, Johansen HK, Song Z, Hougen HP, Rygaard J, Hoiby N (1997) Chronic *Pseudomonas aeruginosa* lung infection is more severe in Th2 responding BALB/c mice compared to C3H/HeN mice. *APMIS*. 105(11), 838-842.

Moser C, Kjaerjaard S, Pressler T, Kharazami A, Koch C, Hoiby N (2000) The immune response to chronic *Pseudomonas aeruginosa* lung infection in cystic fibrosis patients is predominantly of the Th2 type. *APMIS*. 108(5), 329-335.

Muller M, Sorrel TC (1992) Leukotriene B4 omega-oxidation by human polymorphonuclear leukocytes is inhibited by pyocyanin, a phenazine derivative produced by *Pseudomonas aeruginosa*. *Infect Immun*. 60(6), 2536-2540.

Rook GA, Tavern J, Leveton C, Steele J (1987) The role of gamma-interferon, vitamin D3 metabolites and tumour necrosis factor in the pathogenesis of tuberculosis. *Immunology*. 62(2), 229-234.

UNIVERSITY of the
WESTERN CAPE

Rowland SS, Falkler WA, Bashirelahi N (1992) Identification of an estrogen-binding protein in *Pseudomonas aeruginosa*. *J Steroid Biochem Mol Biol*. 42(7), 721-727.

Shaikh WA (1992) Pulmonary tuberculosis in patients treated with inhaled beclomethasone. *Allergy*. 47(4 pt 1), 327-330.

Sorrel TC, Muller M, Sztelma K (1992) Bacterial metabolism of human polymorphonuclear leukocyte-derived arachidonic acid. *Infect Immun*. 60(5), 1779-1785.

Stover CK *et al.* (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature*. 406, 959-964.

Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673-4680

Wilderman PJ, Vasil AI, Johnson Z, Vasil ML (2001) Genetic and biochemical analyses of a eukaryotic-like phospholipase D of Pseudomonas aeruginosa suggest horizontal acquisition and a role for persistence in a chronic pulmonary infection model. *Mol Microbiol*. 39(2), 291-303.

UNIVERSITY of the
WESTERN CAPE

Wolf Y.I. *et al.* (1999) Rickettsiae and Chlamydiae, evidence of horizontal gene transfer and gene exchange. *Trends Genet.* 15, 173-175.

CHAPTER 4

Detection of bacterial genes that are undergoing adaptive evolution: a promising means of identifying novel drug targets and vaccine candidates

4.1. ABSTRACT

Previous studies have shown that certain virulence factors of pathogenic bacteria e.g. adhesins and outer membrane proteins, may be under positive selection. We have performed an intraspecies search for genes where the rate of nonsynonymous (K_A) substitution is greater than the rate of synonymous substitution (K_S) between orthologous genes in pairs of strains of *Helicobacter pylori* and *Neisseria meningitidis*, to ascertain whether new virulence-associated genes may be identified in this way. A total of 85 genes undergoing pathoadaptive evolution was identified of which a large proportion (41/85) code for known or potential virulence genes. Interestingly, it appears that specific cellular processes in *N. meningitidis* may be under strong selection, as we have detected multiple genes involved in iron acquisition and DNA-repair that have acquired adaptive mutations. Furthermore, of 21 *H. pylori* knockout mutants 5 had decreased colonization efficiency and 8 were classified as being 'putative essential', as knockouts were fatal. Due to the demonstrated ability of our system to identify known and potential virulence factors, and the fact that 61% of *H. pylori* genes

UNIVERSITY of the
WESTERN CAPE

tested in gene-knockout experiments were shown to play important roles in the organism’s biology, we conclude that it may be an important tool for novel drug and vaccine targets.

4.2. INTRODUCTION

Natural selection is one of the mechanisms of evolution by which the relative frequencies of genotypes change within a population based upon their relative fitness in the current niche. Nonsynonymous substitutions have a greater potential of contributing to change of protein function than synonymous substitutions, since the latter is mostly neutral, or nearly neutral at a whole-organism functional level. Most nonsynonymous substitutions are, however, deleterious and are consequently eliminated from the population by purifying selection (Kimura 1983). Advantageous mutations are maintained through positive selection and genes/gene regions on which positive selection operate have a characteristic signature, where the rate of nonsynonymous substitutions (K_A) is higher than the rate of synonymous substitutions (K_S) i.e. $K_A/K_S > 1$.

UNIVERSITY of the WESTERN CAPE

While many attributes that make some bacterial lineages more virulent than others are conferred by virulence factors encoded by horizontally transferred foreign genes, pathogenicity can also be enhanced by mutational change of existing genes. Pathoadaptive mutations (Sokurenko 1999) depend on the occurrence of random, often subtle, genetic mutations that confer a selective advantage to the bacterium in an otherwise hostile environment and are especially important for the evolutionary success of those pathogens that do not readily exchange DNA under natural conditions. For example, clones of *Haemophilus* *influenzae* that cause meningitis are rare mutants that arise from the benign population that colonizes the nasopharynx (Moxon 1978). Similarly, a *H. pylori* strain that had demonstrated initial weak growth in the stomachs of gnotobiotic piglets adapted rapidly to vigorous growth in this environment (Akopyants 1995). Genomic fingerprint analysis of the adapted derivative indicated that this adaptation was not associated with deletion or acquisition of genomic segments and was thus most likely due to subtle pathoadaptive mutations. In a large-scale search for genes under positive selection in international DNA databanks, 17 of ~3500 groups of homologous sequences had regions where $K_A/K_S > 1$, demonstrating adaptive mutations (Endo *et al.*, 1996). Nine groups were the surface antigens of parasites and viruses, probably under selection pressure from the host immune system. Three candidates encoded known bacterial virulence determinants, which suggests that searching for genes on which positive selection may operate may be a powerful tool for identifying novel virulence-associated factors in pathogenic bacteria. Those genes undergoing pathoadaptive evolution that code for products that do not interact with the host immune system would be particularly interesting as they probably involved in virulence functions other than antigenic variation, and may represent promising drug targets. With the advent of genome sequencing projects, the genomes of more than one strain of some pathogenic bacteria have been sequenced and the raw data to perform intra-species searches for genes under positive selection has therefore become available.

UNIVERSITY of the
WESTERN CAPE

Helicobacter pylori displays higher levels of allelic diversity than any other organism tested to date, and while horizontal gene transfer has been a major force in its evolution, it appears as if the pathogen prefers *de novo* DNA mutation as its strategy to adapt to changing environments. For example, although *H. pylori* is known to be naturally easily transformable, all clinically observed antibiotic resistance has been found to be due to mutations of existing chromosomal genes, rather than through the acquisition of resistance genes via mobile genetic elements (Wang 1999). *Neisseria meningitidis* has a largely commensal lifestyle and lives on the mucosa of the nasopharynx. It has been shown to have comparable synonymous and nonsynonymous substitution rates to *H. pylori* (Suerblaum and Achtman 1999), a hint that it too may favor mutation as the major origin for genetic diversity. We identified genes undergoing pathoadaptive evolution in both organisms by intraspecies comparisons of the genome sequences of the two *H. pylori* isolates, 26995 (Tomb *et al.* 1997) and J99 (Alm *et al.* 1999), and the two *N. meningitidis* strains, Z2491 (Parkhill *et al.* 2000) and MC8 (Tettelin *et al.* 2000). We present strong circumstantial and empirical evidence that identifying genes under positive selection in pathogenic bacteria is a reliable means of identifying novel virulence-associated genes and novel drug targets.

UNIVERSITY of the WESTERN CAPE

4.3. MATERIALS AND METHODS

4.3.1. Sequence data

Predicted gene and protein sequences for two *H. pylori* isolates, 26995 (Tomb *et al.* 1997) and J99 (Alm *et al.* 1999), and the two *N. meningitidis* strains, Z2491 (Parkhill *et al.* 2000) and MC8 (Tettelin *et al.* 2000) were downloaded from Genbank (www.ncbi.nlm.nih.gov). *H. pylori* 29975 was isolated in the

United Kingdom before 1987 from a gastritis patient and had a history of subculturing before being sequenced, and the J99 strain was isolated from patient with a duodenal ulcer in the United States and was only subjected to minimal subculturing before being sequenced. *N. meningitidis* MC58 is a serogroup B strain isolated from a case of invasive infection, while Z2491 is a serogroup A strain, which is responsible for most of the morbidity and mortality associated with meningococcal meningitis.

4.3.2. Construction of accurate alignments of orthologous genes

Protein sequences were used to produce orthologous gene pairs for the 2 genomes of interest, using the FASTA2 similarity search program (Pearson 1990). As we compared two strains of the same species, we were confident that true orthologs were identified. To prevent unnecessary processing, a working dataset was constructed using those pairs that were less than 100% identical, as K_A for genes encoding identical proteins is obviously zero. Protein alignments were then generated for each orthologous pair, using the CLUSTALW program (Thompson *et al*, 1994) and used as a 'scaffold' for making accurate *in*-frame alignments of the coding sequences. The generation of in-frame alignments was extremely important to ensure that the correct codons were compared, which is necessary for the accurate calculation of mutation rates.

UNIVERSITY of the WESTERN CAPE

4.3.3. Identification of genes under positive selection

The rate of synonymous substitutions (K_S) per site can be assumed to represent the background mutation rate within a specific gene or genome and the average rate of nonsynonymous substitution (K_A) can be expected to be orders of magnitude less than K_S due to the action of purifying selection (Kimura 1977). Therefore, those genes that display a nonsynonymous substitution rate (K_A) that is higher than the background can be assumed to be under positive functional selection (Kimura 1977). To effectively identify genes containing regions under positive selection, we used the *wina* program (Endo et al. 1996), which calculates K_S and K_A in 20 codon windows of a gene. The windowing approach was used to identify gene regions under positive selection, where the averaged K_A for the gene would still be less than K_S, thus missing interesting genes. We considered this highly significant in our study, as the organisms tested are evolutionarily very close, and very few substitutions, especially nonsynonymous ones, were expected. Furthermore, to avoid the saturation effect of nucleotide substitutions (Endo et al. 1996) only windows where $K_S<1$ were considered. Each preliminary candidate was then searched against the entire protein database of each isolate using *BLASTP* (Altschul *et al.* 1997) to determine whether it had been duplicated in one or both genomes. In cases where gene duplication has occurred, it is not possible to determine whether any diversifying mutations observed is due to the action of positive selection or relaxed selection on the duplicates.

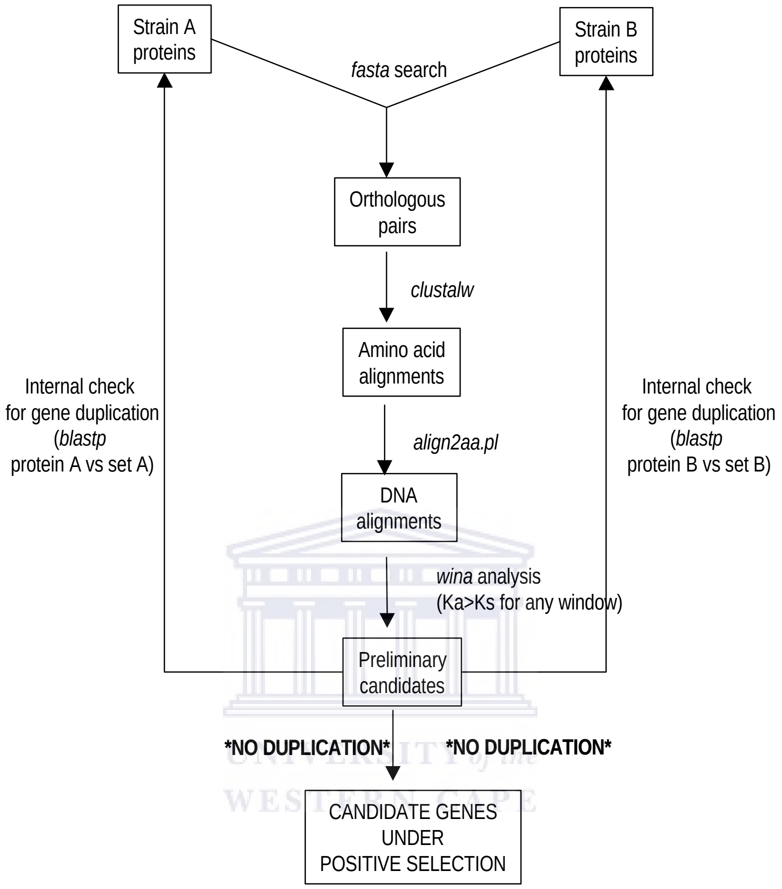
(The entire process is summarized in Figure 4.1).

UNIVERSITY of the
WESTERN CAPE

4.3.3. Selection of *H. pylori* candidates for assessment of their involvement in colonization by gene knockout

To test our hypothesis that searching for genes on which positive selection may operate may be a powerful tool for identifying novel virulence-associated factors in pathogenic bacteria, 21 *H. pylori* candidate genes under positive selection were chosen for knockout studies by site-directed mutagenesis. The resulting mutants were tested for relative colonization efficiency in a mouse model (Guo *et al.*, manuscript in preparation). The list of selected candidates was presented to the Mekalanos research group at Harvard University, which performed the knockout experiments. The subset of candidates was chosen based on each gene's speculative involvement in colonization, virulence, higher levels sequence divergence and/or accumulation of chemically non-similar amino acid substitutions (Table 4.3).

UNIVERSITY of the
WESTERN CAPE



Strain A proteins
Strain B proteins
fasta search
Orthologous pairs
clustalw
Amino acid alignments
align2aa.pl
DNA alignments
wina analysis (Ka>Ks for any window)
Preliminary candidates
Internal check for gene duplication (*blastp* protein A vs set A)
Internal check for gene duplication (*blastp* protein B vs set B)
NO DUPLICATION
NO DUPLICATION
CANDIDATE GENES UNDER POSITIVE SELECTION

UNIVERSITY of the WESTERN CAPE

Figure 4.1. Diagram summarizing the method used to identify genes under positive selection.

4.4. RESULTS AND DISCUSSION

4.4.1 Numbers of Genes on which Positive Selection May Operate

Using a window-based search for genes on which positive selection may operate, we have determined that 41/1334 (2.55%) and 44/2081 (1.83%) genes that have acquired pathoadaptive mutations in *H. pylori* and *N. meningitidis*, respectively (Tables 3.1 and 3.2). Examples illustrating gene regions undergoing adaptive evolution are presented by plots of K_S and K_A (Figures 4.2 and 4.3) and protein alignments (Figures 4.4 and 4.5). Two genes, pantoate-β-alanine ligase and lipopolysaccharide biosynthesis protein, are present in both organisms, implicating their involvement in a common pathoadaptation/virulence strategy. As the system has the potential to identify common genes that have a role in pathoadaptation in related organisms e.g. Gram negative pathogens, it bears the promise of exposing targets for the development of novel broad-spectrum antimicrobial drugs, should they be determined to be indispensable by knockout mutagenesis experiments.

UNIVERSITY of the
WESTERN CAPE

4.4.2. A large proportion of candidates encode known or potential virulence factors

Surprisingly, a large number of candidate genes (41/85) encode known virulence factors, or have been shown to be involved in pathogenesis in other organisms (Tables 4.1, A and 4.2, A). In *H. pylori*, 2 of the most important known virulence factors, *cagA* and *babA*, which code for the immunodominant antigen and a blood-group-antigen-binding adhesin respectively, are under positive selection (Table 4.1, A). These genes have been positively correlated with the ability of *H. pylori* to cause gastric ulcers and adenocarcinomas (Gerhard *et al*. 1999). A third factor, vacA, which is also normally associated with the development of these diseases, was also detected in the screening process. However, this gene was not reported as it has undergone gene duplication in the 26695 strain, and it is impossible to determine whether the mutations seen is due to selection pressure or relaxed selection induced by the duplication event. Other known *H. pylori* virulence factors that appear to under positive selection are type IV secretion system component, virB4; a cytotoxin secreting protein; a urease accessory protein; and an iron-regulated outer membrane protein. *N. meningitidis* known virulence factors under selection include genes involved in invasion, a hemolysin, pilus assembly and modification proteins, an immunoglobulin A protease, a glutathione synthase, and many coding for proteins involved in host iron acquisition (Table 4.2, A). Furthermore, two genes encoding known virulence factors, pantoate beta-alanine ligase and a LPS biosynthesis protein, are under positive selection in both organisms.

UNIVERSITY of the
WESTERN CAPE

Many of the candidates identified by our system are transmembrane or outer-membrane proteins, or are secreted. They are therefore most likely under selective pressure for diversification to allow escape from the host immune system. These proteins may thus also be highly immunogenic, which make them promising candidates for recombinant vaccines. Interestingly, several are not exposed to the host immune system and the nonsynonymous mutations seen here may represent selective mechanisms for the functional adaptation of factors that are essential for virulence.

4.4.3. Specific processes appear to be under positive selection

The process of iron acquisition, known to be critical for intrahost survival, appears to be under strong selection in *N. meningitidis*, with 7/38 candidates belonging to this class alone. These genes may highlight a critical process that could be exploited for the development of novel therapeutic interventions. Furthermore, 6 *N. meningitidis* enzymes involved in DNA mismatch repair have accumulated diversifying mutations (Table 4.2, B) which, at first glance, appears to allude to strong selection on processes that rectify insults on the organisms' genetic material caused by the initial host immune reponse like the 'oxidative-burst'. However, mutations in genes of this functional class have been shown to increase frequency of genome-wide point mutations from a few fold to over 1000-fold in *E. coli* (reviewed by Wang *et al.* 1999). It therefore appears that relaxation of replication fidelity is important to pathoadaption in *N. meningitidis*, as this improves the chances of acquiring beneficial mutations, a phenomenon that has been observed in other bacterial pathogens. The most virulent strains of Yersinia spp., for example, intrinsically have the highest rates of mutation (Najdenski 1995, Iteman 1995). While only one of the *H. pylori* candidates belong to this it lacks the entire *mutHLS* mismatch repair system (reviewed by Wang *et al*. 1999), and may in fact be hypermutable and both organisms appears to be 'evolving to evolve'. As the resulting global-

UNIVERSITY of the WESTERN CAPE

mutator phenotype proposed here causes genome-wide nonspecific mutations, the positive selection of the nonsynonymous mutations in specific genes suggests that they may be involved in pathoadaptation and virulence.

4.4.4. *H. pylori* gene knockouts implicate genes undergoing pathoadaptive mutation as potential drug targets

Knockout mutants for 21 *Helicobacter pylori* genes were prepared and tested in a mouse infection model for colonization efficiency. Of these, 5 displayed a reduced or attenuated stomach colonization phenotype. A further 8 were tentatively categorized as being essential, with the knockout mutagenesis proving fatal to the organism. 61% of all candidates tested demonstrate important roles in the organism’s biology. Identifying genes under positive selection is thus a powerful tool for identifying novel virulence-associated genes, drug targets and possible vaccine candidates. Of 8 hypothetical genes that were knocked out, 2 were shown to be essential and 1 is required for efficient stomach colonization. Hypothetical genes can represent up to 50% of the genes in a bacterial genome and those unique to an organism are arguably the most interesting, as they may be essential for an organism survival in general or its success as a pathogen. Those hypothetical proteins identified as being under positive selection here may thus represent a good starting-point for the functional characterization of these organism-specific groups of proteins, and the possible identification of novel virulence-associated genes and drug targets.

UNIVERSITY of the WESTERN CAPE

Table 4.1. *Helicobacter pylori* genes under positive selection. Genes containing regions where dN>dS were selected as candidates.

Gene name	Function/putative annotation
	A. Known/Putative Virulence factors
HP0547	CagA – Immunodominant antigen
HP0459	VirB4 – type IV secretion system component
HP0522	*cag3*, pathogenicity island protein
HP0731	Enterotoxin secretion protein
HP0069	Urease accessory protein
HP0916	Iron-regulated outer membrane protein
HP0227	Adhesin-Lewis antigen *babA*
HP1105	LPS biosynthesis protein
HP0279	LPS heptosyl transferase I
HP1191	ADP-heptose-LPS heptosyl transferase II
HP1274	Paralysed flagella protein
HP0818	Osmoprotection protein
HP0366	Polysaccharide biosynthesis protein
HP0006	Pantoate-beta-alanine ligase
	B. Other
HP0717	DNA polymerase III gamma and tau subunits
HP0648	UDP-N-acetylglucosamine 1-carboxyvinyltransferase
HP0329	NH(3)-dependent NAD+ synthetase
HP0832	Spermidine synthase
HP0640	Poly-A polymerase
HP1266	NADH-ubiquinone oxidoreductase
HP0290	Diaminopimelate decarboxylase
HP1237	Carbamoyl phosphate synthetase
HP1232	Dihdropterate sythetase
HP0293	Para-amino benzoate synthetase
HP0004	Carbonic anhydrase

UNIVERSITY of the WESTERN CAPE

Table 4.1. Continued…

Gene name	Function/putative annotation
	B. Hypothetical proteins
HP0060	Unique hypothetical protein
HP0063	“
HP0465	“
HP0469	“
HP0965	“
HP0969	“
HP0258	Hypothetical protein
HP0384	“
HP0668	“
HP0669	“
HP0781	“
HP0519	“
HP0554	“
HP0635	“
HP0800	“
HP0906	“

UNIVERSITY of the
WESTERN CAPE

Table 4.2. *N. meningitidis* genes on which positive selection may operate. Genes containing regions where dN>dS were selected as candidates and divided into functional classes.

Gene name	Function/annotation
	A. Known/Putative Virulence factors
NM0178	Outer membrane protein P1
NM0398	Class III outer membrane protein
NM0324	Adhesin mafB
NM0383	Porin
NM0264	Class I fimbrial protein
NM0368	Pilin glycosylation protein
NM0369	Pilin glycosylation protein
NM1700	Type IV pilus assembly protein *pilV*
NM0448	Iron(III)-ABC transporter
NM0457	Adhesion and penetration protein
NM0527	Lacto-N-neotetraose biosynthesis glycosyltransferase
NM0668	Hemolysin-related protein
NM0776	T- and B- cell stimulating antigen
NM1797	T- cell stimulating protein B
NM1057	LPS biosynthesis protein
NM1107	TonB dependent receptor
NM1739	Lactoferrin-binding protein A
NM1740	Lactoferrin-binding protein B
NM1926	Hemoglobin receptor
NM2024	Transferrin-binding protein A
NM2025	Transferrin-binding protein B
NM0905	Immunoglobulin A protease
NM1089	Pantoate beta-alanine ligase
NM1747	Glutathione synthase
NM1803	Sensor histidine kinase

UNIVERSITY of the
WESTERN CAPE

Table 4.2. Continued…

Gene name	Function/annotation
	B. DNA repair system components
NM0158	DNA processing chain A
NM0430	ATP dependent helicase
NM0995	Exodeoxyribonuclease V
NM1491	Transcription-repair coupling protein
NM1655	MutL DNA repair component
NM1656	DNA polymerase III subunit tau
	C. Other
NM1048	Peptide maturation factor
NM1179	Phosphoserine phosphatase
	D. Hypothetical proteins
NM0696	Hypothetical protein
NM0782	“
NM1050	“
NM1051	“
NM1090	“
NM1108	“
NM1561	“
NM1652	“
NM1841	“
NM1851	“

UNIVERSITY of the WESTERN CAPE

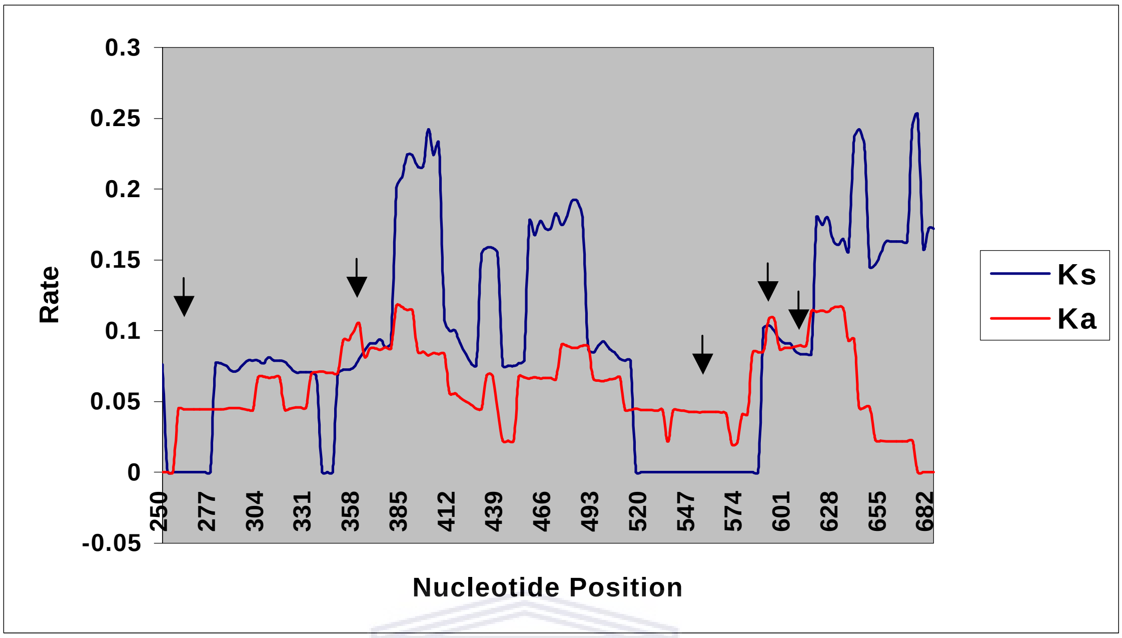


Figure 4.2. Plots of synonymous (Ks) and nonsynonymous (Ka) substitution rates calculated over 60 base pair windows of the *H. pylori* LPS-heptosyltransferase I gene (HP1191). Black arrows show regions undergoing adaptive evolution.

UNIVERSITY of the
WESTERN CAPE

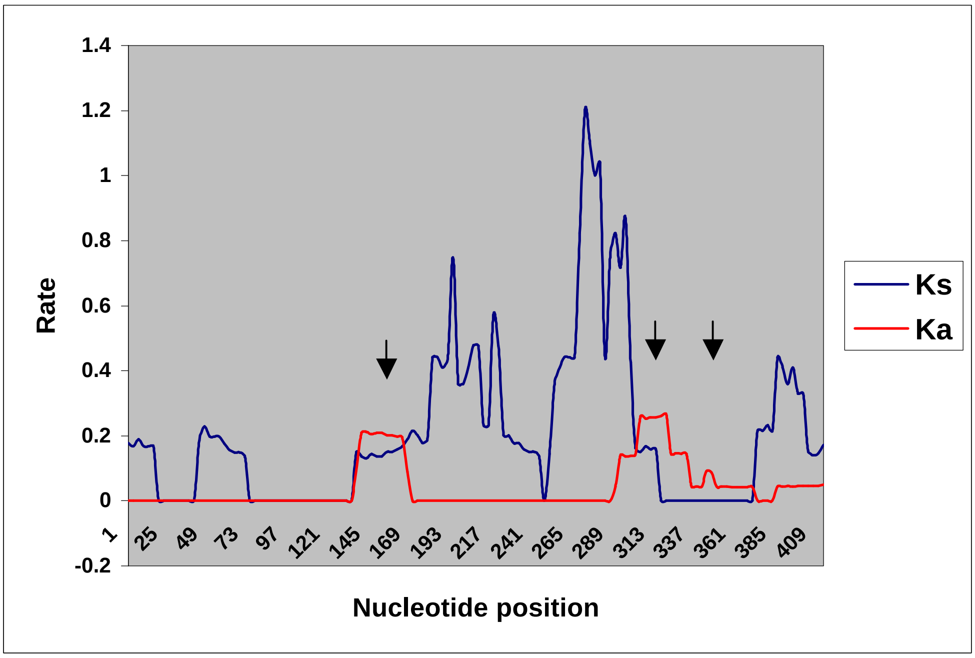
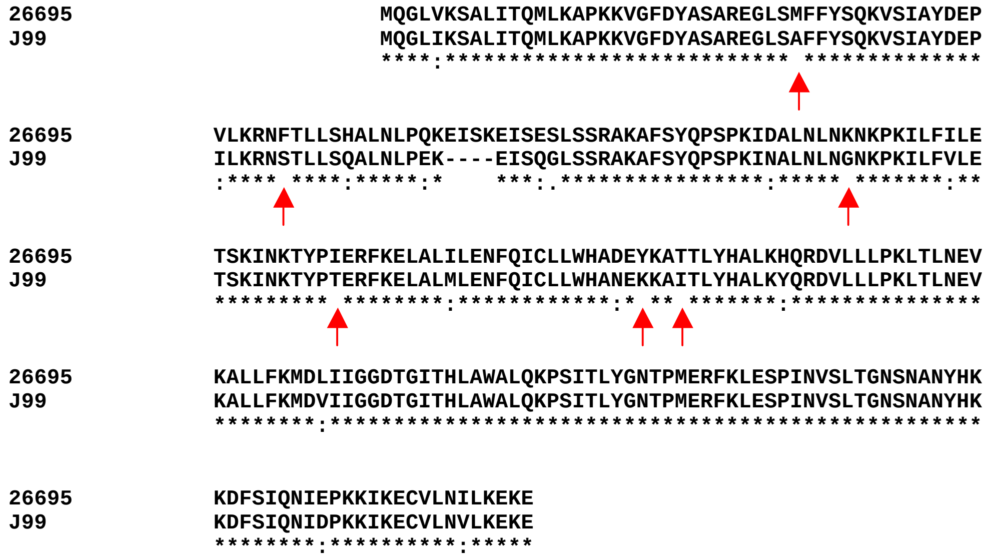


Figure 4.3: Plots of synonymous (Ks) and nonsynonymous (Ka) substitution rates calculated over 60 base pair windows of the *H. pylori* gene coding for NH_3-dependent NAD synthetase (HP0329). Black arrows show regions undergoing adaptive evolution.

UNIVERSITY of the
WESTERN CAPE



```
26695                        MQGLVKSALITQMLKAPKKVGFDYASAREGLSMFFYSQKVSIAYDEP
J99                          MQGLIKSALITQMLKAPKKVGFDYASAREGLSAFFYSQKVSIAYDEP
                             ****:*************************** **************

26695           VLKRNFTLLSHALNLPQKEISKEISESLSSRAKAFSYQPSPKIDALNLNKNKPKILFILE
J99             ILKRNSTLLSQALNLPEK----EISQGLSSRAKAFSYQPSPKINALNLNGNKPKILFVLE
                :**** ****:*****:*    ***:.****************:***** *******:**

26695           TSKINKTYPIERFKELALILENFQICLLWHADEYKATTLYHALKHQRDVLLLPKLTLNEV
J99             TSKINKTYPTERFKELALMLENFQICLLWHANEKKAITLYHALKYQRDVLLLPKLTLNEV
                ********* ********:************:* ** *******:***************

26695           KALLFKMDLIIGGDTGITHLAWALQKPSITLYGNTPMERFKLESPINVSLTGNSNANYHK
J99             KALLFKMDVIIGGDTGITHLAWALQKPSITLYGNTPMERFKLESPINVSLTGNSNANYHK
                ********:***************************************************

26695           KDFSIQNIEPKKIKECVLNILKEKE
J99             KDFSIQNIDPKKIKECVLNVLKEKE
                ********:**********:*****
```

Figure 4.4. Protein alignment illustrating the degree of divergence of the *H. pylori* NH3-dependent NAD synthetase between the 29975 and J99 strains. Red arrows highlight radical (charge-changing) substitutions.

```
26695                                     VFSFHAIKPITTAEGGAVVTNDSELHEKMKLFRS
J99                                       VFSFHAIKPITTAEGGAVVTNNSELYEKMKLFRS
                                          *********************:***:********

26695           HGMLKKDFFEGEVKSIGHNFRLNEIQSALGLSQLKKAPFLMQKREEAALTYDRIFKDNPY
J99             HGMLKKDFFEGEVKSVGYNFRLNEIQSALGLSQLKKAPLLRQKREEVALIYDEIFKDNPY
                ***************:*:********************:* *****.** **.*******
                                                        ↑        ↑

26695           FTPLHPLLKDKSSNHLYPILMHQKFFTCKKLILESLHKRGILAQVHYKPIYQYQLYQQLF
J99             FTPLHPLLKHQSSNHLYPILMHQKFFTCKKLILESLHKLGILVQVHYKPIYQYQLYQQLF
                *********.:*************************** ***.*****************
                                                      ↑

26695           NTAPLKSAEDFYHAEISLPCHANLNLESVQNIAHSVLKTFESFKIE---
J99             NTAPLKSAEDFYSAEISLPCHANLDLESVQNIAHGVLKTFEGFNRMGFI
                ************ ***********:*********.******.*:
                            ↑
```

Figure 4.5. Protein alignment illustrating the degree of divergence of the Lewis antigens of the *H. pylori* 29975 and J99 strains. Red arrows highlight radical (charge-changing) substitutions.

UNIVERSITY of the
WESTERN CAPE

Table 4.3. A summary of genes selected for knockout analysis and colonization efficiency tests

Candidate	Motivation for choice as knockout candidate	Colonization defective (C) or putative essential (E)
HP0640	May play a key role in modulating RNA degradation [1]	C
HP0279	Involved in the production of core oligosaccharide region on LPS [2]	E
HP1191	“ [2]	C
HP1274	Knockout may produce a paralyzed phenotype [3]	C
HP1266	May be important for flagellar rotation[4]	E
HP1237	Key role in producing carbamoyl phosphate [5]	C
HP0329	May be involved in stress response [6]	E
HP0006	Pantothenate biosynthesis. Important for thiamine production [7]	E
HP0293	Key role in folic acid production [8]	E
HP0818	Role in cell wall biology [2]	
HP0366	“ [2]	
HP1232	Classic drug target [9]	E
HP0004	Inhibited by sulphanomides [10]	
HP0668	Homolog deleted in the virulence-attenuated *M. bovis* BCG strain	C
HP0669	Homolog deleted in the virulence-attenuated *M. bovis* BCG strain	
HP0060	Unique hypothetical gene that may play key organism-specific roles	
HP0063	“	
HP0909	“	E
HP0965	“	
HP0465	“	
HP0469	“	E

1. Yamanaka and Inouye, 2001
2. Tomb *et al.,* 1997
3. Yao *et al.,* 1994
4. Hase and Mekalanos, 1999
5. Holden *et al.,* 1999
6. Antelmann *et al.,* 1997
7. Enos-Berlage and Downs, 1997
8. Viswnanthan *et al.,* 1995
9. Maiden, 1998
10. Scozzafava and Supuran, 2000

UNIVERSITY of the
WESTERN CAPE

4.5. CONCLUSION

We present strong circumstantial and empirical evidence that identifying genes under positive selection in pathogenic bacteria is a powerful method of identifying novel virulence factors and potential drug targets. Of 21 candidates tested, 13 (61%) play a role in stomach colonization in mice or are essential. Furthermore, critical physiological processes may also be highlighted, providing important information that could be used in the design of novel therapeutic interventions. Those surface proteins that are under selective pressure from the host immune system could potentially provide candidates for multi-antigen recombinant vaccines, as they should be highly immunogenic. While we have presented evidence that our system effectively identifies important genes in pathogenic organisms, it is, however, dependent on the availability of at least two genomes from the same organism. It is likely that even when comparing two closely related species in the same genus, genes that have undergone adaptive evolution may be missed due to the elevated K_S due to different codon preferences. Re-engineering the technology to identify genes that have undergone divergent evolution in one of a pair closely related organisms e.g. virulent and avirulent species from the same genus, may permit the identification of pathogen-specific strategies for pathoadaptation, tissue tropism, immune evasion etc. In short, the system and any future derivatives bears the promise of maximizing the use of genome data by identifying ‘what is different between what is common’, rather than the traditional ‘what is unique?’.

UNIVERSITY of the WESTERN CAPE

4.6. REFERENCES

Akopyants NS, Eaton KA, Berg DE (1995) Adaptive mutation and co-colonization during *Helicobacter pylori* infection of gnotobiotic piglets. *Infect Immun*. 63(1), 116-121.

Alm RA *et al.* (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature.* 397, 176-180.

Altschul S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25, 3389-3402.

Antelmann H, Schmid R, Hecker M. The NAD synthetase NadE (OutB) of *Bacillus subtilis* is a sigma B-dependent general stress protein. *FEMS Microbiol Lett*. 153(2), 405-409.

Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. *Mol Biol Evol*. 13(5), 685-690.

Enos-Berlage JL, Downs DM (1997) Mutations in sdh (succinate dehydrogenase genes) alter the thiamine requirement of *Salmonella typhimurium*. *J Bacteriol*. 179(12), 3989-3996.

UNIVERSITY of the
WESTERN CAPE

Gerhard M, Lehn N, Neumayer N, Boren T, Rad R, Schepp W, Miehlke S, Classen M, Prinz C (1999) Clinical relevance of the *Helicobacter pylori* gene for blood-group antigen-binding adhesin. *PNAS (USA).* 96(22), 12778-12783.

Hase CC, Mekalanos JJ. (1999) Effects of changes in membrane sodium flux on virulence gene expression in *Vibrio cholerae*. *Proc Natl Acad Sci U S A*. 96(6), 3183-3187.

Holden HM, Thoden JB, Raushel FM. (1999) Carbamoyl phosphate synthetase: an amazing biochemical odyssey from substrate to product. *Cell Mol Life Sci*. 56(5-6), 507-522.

Iteman I, Nadjenski H, Carniel E (1995) High genomic polymorphism in *Yersinia pseudotuberculosis*. *Contrib Microbiol Immunol*. 13, 106-111.

Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*. 267, 275-276.

Maiden MC (1998) Horizontal genetic exchange, evolution, and spread of antibiotic resistance in bacteria. *Clin Infect Dis*. 27 Suppl 1, S12-20.

Moxon ER, Rainey PB, Nowak MA, Lenski RE (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr Biol*. 4(1), 24-33.

UNIVERSITY of the WESTERN CAPE

Nadjenski H, Iteman H, Carniel E (1995) The genome of *Yersinia enterocolytica* is the most stable of the three pathogenic species. *Contrib Microbiol Immunol*. 13, 281-284.

Parkhill J *et al.* (2000) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature.* 404, 502-506.

Scozzafava A, Supuran CT (2000) Carbonic anhydrase and matrix metalloproteinase inhibitors: sulfonylated amino acid hydroxamates with MMP inhibitory properties act as efficient inhibitors of CA isozymes I, II, and IV, and N-hydroxysulfonamides inhibit both these zinc enzymes. *J Med Chem*. 43(20), 3677-3687.

Sokurenko EV, Hasty DL, Dykhuizen DE (1999) Pathoadaptive mutations: gene loss and variation in bacterial pathogens. *Trends Microbiol*. 7(5),191-5.

Suerbaum S, Achtman M (1999) Evolution of *Helicobacter pylori*: the role of recombination. *Trends Microbiol*. 7(5), 182-184.

Tettelin H *et al.* (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science.* 287, 1809-1815.

UNIVERSITY of the
WESTERN CAPE

Thompson J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673-4680.

Tomb F *et al.* (1997) The complete genetic sequence of the gastric pathogen *Helicobacter pylori. Nature.* 388, 539-547.

Viswanathan VK, Green JM, Nichols BP (1995) Kinetic characterization of 4-amino 4-deoxychorismate synthase from *Escherichia coli*. *J Bacteriol*. 177(20), 5918-5923.

Wang G, Humayun MZ, Taylor DE (1999) Mutation as the origin of genetic variability in *Helicobacter pylori*. *Trends Microbiol.* 7,488-493.

UNIVERSITY of the
WESTERN CAPE

Yamanaka K, Inouye M (2001) Selective mRNA degradation by polynucleotide phosphorylase in cold shock adaptation in *Escherichia coli*. *J Bacteriol*. 183(9), 2808-2816.

Yao R, Burr DH, Doig P, Trust TJ, Niu H, Guerry P (1994) Isolation of motile and non-motile insertional mutants of *Campylobacter jejuni*: the role of motility in adherence and invasion of eukaryotic cells. *Mol Microbiol*. 14(5), 883-893.

CHAPTER 5

CONCLUSIONS, PITFALLS AND PROMISES

The overall objective of this project was assess the utility of systems that employ two important features of bacterial evolution, horizontal gene transfer and adaptive mutation, in the identification of potentially novel virulence-associated factors and possible drug targets.

Rather than searching for genes that were acquired by horizontal transfer from other bacteria, it was decided to determine whether acquisition of genetic material from eukaryotes played a role in the evolution of pathogenic bacteria. We developed a system that detects genes in a bacterial genome which have been acquired by interkingdom horizontal gene transfer, and initially identified 19 eukaryotic genes in the genome of *Mycobacterium tuberculosis*. Two of these, along with two novel eukaryotic genes were later found in the genome of *Pseudomonas aeruginosa*. Surprisingly, most of *M. tuberculosis* genes and all four eukaryotic genes in *P. aeruginosa* may be involved in modulating the host immune response through altering the steroid balance and the production of proinflammatory lipids.

UNIVERSITY of the
WESTERN CAPE

Although the system successfully identifies foreign genes that may be involved in the host-pathogen interaction, the similarity-search-based screen may miss eukaryotic-like genes that have been initially acquired by one bacterium and transferred to another. In this scenario, the protein of interest may score better against the bacterial protein than its eukaryotic homologs and would thus be eliminated from further analyses. One possible solution may be to use learning-based methods e.g. Hidden Markov Models that compare bacterial proteins to trained models of protein families found in both kingdoms, and selecting those proteins that match the eukaryotic model best.

We have also identified 41 and 44 genes that have accumulated adaptive mutations in the genomes of *Helicobacter pylori* and *Neisseria meningitidis* respectively. As approximately 50% of the genes encode known or potential virulence factors, the remaining genes may be implicated in virulence or pathoadaptation. A subset of 21 *H. pylori* genes, none of which are classic virulence factors, was tested for a role in colonization by gene knockout experiments. Of these, 61% were found to be either essential, or involved in effective stomach colonization in a mouse infection model. Searching for genes under positive selection is thus a reliable method of identifying novel virulence-associated genes and drug targets.

UNIVERSITY of the WESTERN CAPE

A major shortcoming of this system is its dependence on the availability of two completely sequenced genomes, from different strains of the same organism. Development of a method that compares orthologous gene sets of two organisms from the same genus, with the aim of identifying genes that have ‘specialized’ in either organism, would be a worthwhile endeavor as public data of this nature is readily available. One strategy may be to use methods that predict the last common ancestral sequence and then finding those proteins that have diverged from the ancestral form through the acquisition of adaptive mutations in one species, a strategy that may highlight genes involved in tissue tropism and adaptation to the host niche.

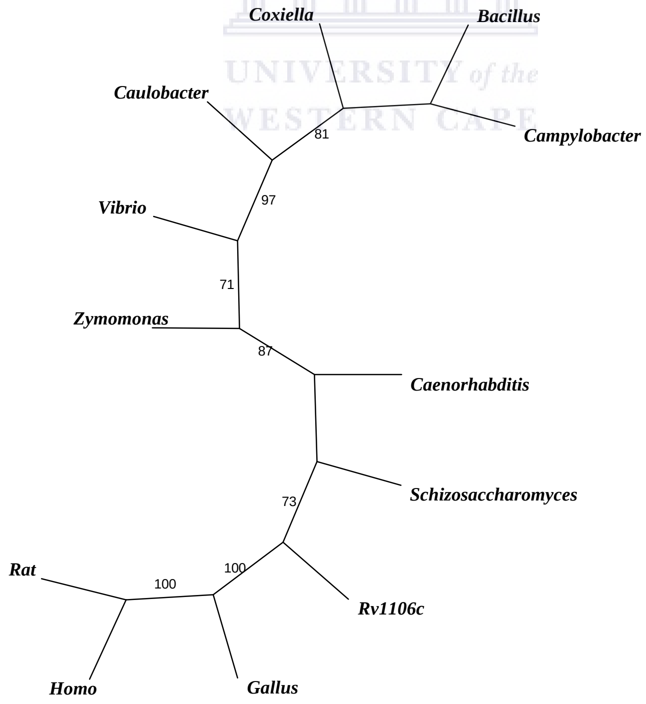
In conclusion, this study has successfully demonstrated that novel virulence-associated factors and mechanisms may be uncovered through the exploitation of phenomena known to be important in the adaptive evolution of bacterial pathogens. Furthermore, the circumstantial and empirical evidence presented here suggests that these technologies may present a reliable starting point for the development of screens for novel drug targets and vaccine candidates, drastically reducing the time for the development of novel therapeutic strategies.

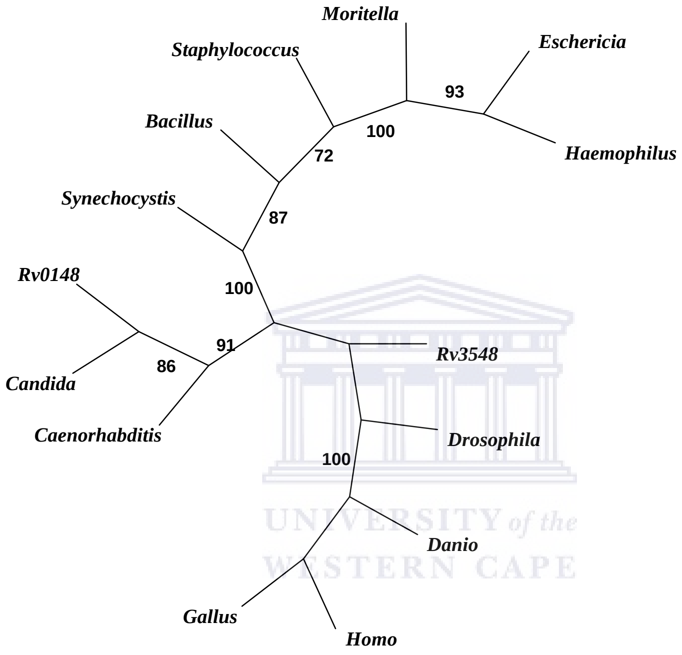
UNIVERSITY of the
WESTERN CAPE

APPENDIX II

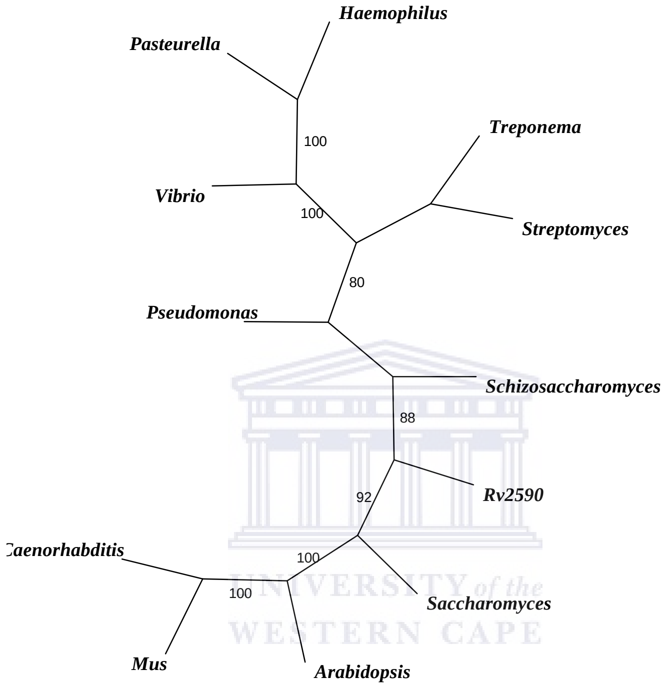
Phylogenies, illustrating the clustering of mycobacterial proteins with eukaryotic rather than prokaryotic proteins. Trees were constructed using the Neighbor Joining method. Numbers indicate the percentage of bootstrap replications above 70%, out of 1000 replications, in which the given node was observed. Very similar trees were obtained using maximum likelihood and parsimony-based methods. Mycobacterial genes are shown in the published Rv-nomenclature e.g. Rv0001.

UNIVERSITY of the
WESTERN CAPE

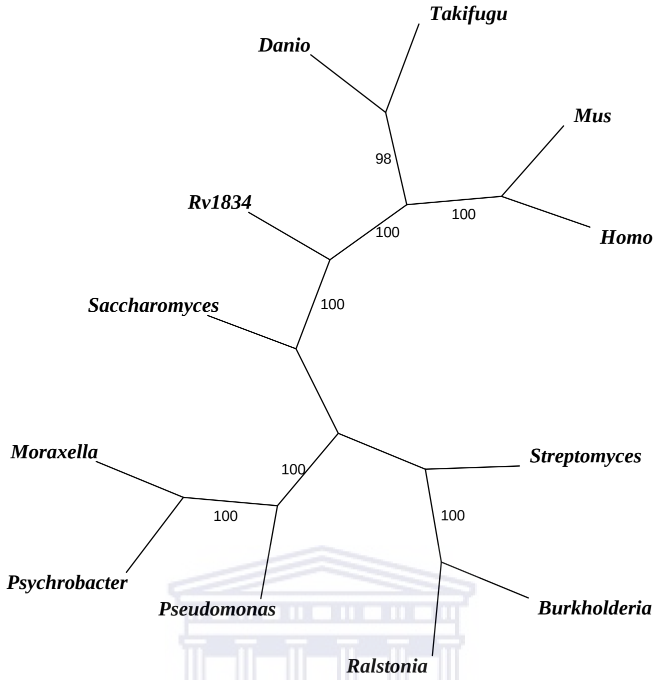




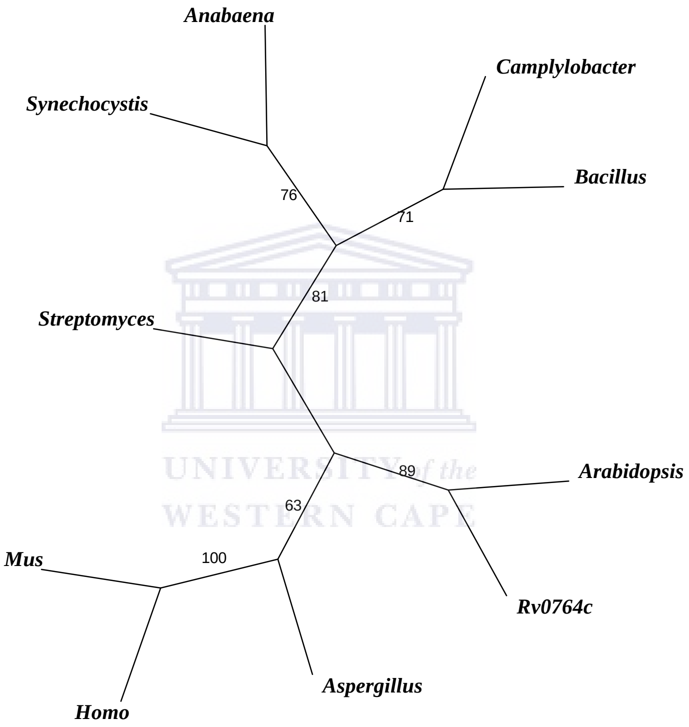
UNIVERSITY of the
WESTERN CAPE



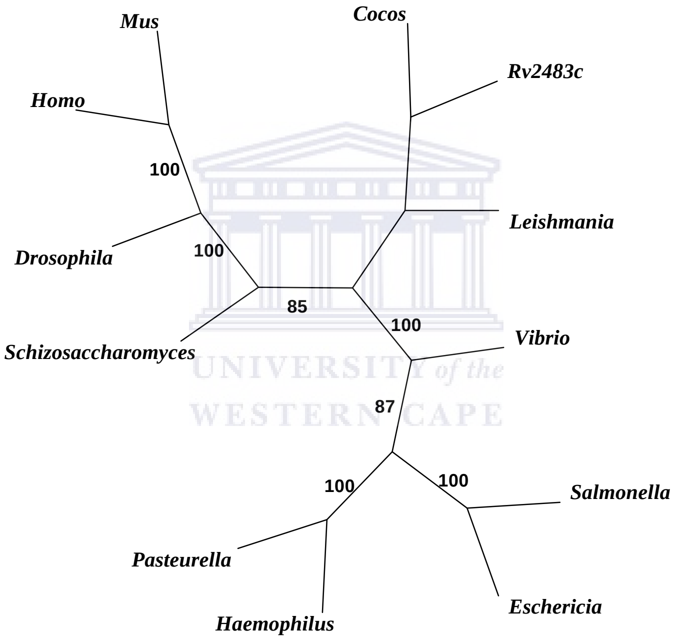
UNIVERSITY of the
WESTERN CAPE



UNIVERSITY of the
WESTERN CAPE



UNIVERSITY of the
WESTERN CAPE



UNIVERSITY of the
WESTERN CAPE