Comparative genomic analyses of *Corynebacterium pseudotuberculosis*

Florence Elizabeth Pethick

B.Sc. (Hons)

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy at the Faculty of Veterinary Medicine, University of Glasgow

May 2013

To Mum and Dad with love and thanks...

...& to Grandpa,

whom I know has been beside me through every page.

Abstract

This study set out to sequence the genome of *Corynebacterium pseudotuberculosis* (*Cp*) 3/99-5, an ovine strain isolated from a naturally-occurring case of caseous lymphadenitis (CLA) in Scotland. The isolate was sequenced and assembled by 454 Life Sciences, and then gap closure performed by ‘PCR bridging’. The resulting sequence consisted of three contigs with a length of 2,319,079 bp and a G+C content of 52.18%. The genome was then annotated and predicted to contain 2,153 coding sequences. Analysis of the coding sequences revealed the presence of several putative virulence factors, including four sortases with multiple sortase target proteins containing LPXTG motifs.

A further two *Cp* strains, an Australian ovine and a North American equine isolate, as well as *C. ulcerans* NCTC 12077 were sequenced for comparison. Comparative genomics, both intra- and inter-species showed all the genomes to be highly homologous. However, the *C. ulcerans* genome is larger than the *Cp* genomes and is more distinct; it was found to be more similar to the equine *Cp* 1/06-A isolate which is the most diverged of the *Cp* isolates.

Phylogenetic analyses of the *Corynebacterium* genus were performed using house-keeping loci but also secreted protein loci from *Cp* 3/99-5. Bayesian analysis of house-keeping loci distinguished the bacteria to a species level. Inclusion of secreted protein loci did not distinguish the isolates any further.

The main objective of this work was to utilise the *Cp* genome sequence to identify potential diagnostic targets which could be used to augment the available ELITEST CLA or replace it. The ELITEST CLA is the only diagnostic test for CLA that exists on the commercial market in the UK. However, due to low specificity and sensitivity, it is only operated on a flock/group basis. Analyses of the *Cp* 3/99-5 genome identified several potential diagnostic candidates and seven protein targets were investigated further. Attempts were made to express these candidates as recombinant proteins, however, only two recombinants were successfully expressed and purified, Cp3995_0570 and CP40. The seroreactivity of these were then assessed by IgG ELISA using a panel of ten positive and ten negative CLA ovine sera. The sera were previously defined as positive or negative by PLD and whole cell ELISAs; both of which showed a significant difference between sera types. However, neither Cp3995_0570 nor CP40 distinguished between sera originating from *Cp*-infected and *Cp*-naïve animals.

Declaration

The work reported in this thesis was carried out under the supervision of Dr Michael C. Fontaine and Dr Alex F. Lainson at the Moredun Research Institute and Professor David G. E. Smith at the Moredun Research Institute and the Institute for Infection, Immunity & Inflammation, University of Glasgow. All results presented, unless otherwise stated, are the sole work of this author, as is the composition of this thesis.

Signed: Date:

Acknowledgments

First and foremost I would like to thank my supervisors, Prof David G.E Smith, Dr Alex Lainson and Dr Mike Fontaine for their guidance throughout this project. In particular thanks go to Alex but also Raja Yaga and Dushyanth Jyothi for introducing me to the ever-expanding world of bioinformatics. I would like to thank Dr Frank Wright for phylogenetics assistance and for helpful and informative discussions. Thanks must also go to Quality Meat Scotland, The English Beef and Lamb Executive and Hybu Cig Cymru (Meat Promotion Wales) who funded this work.

Importantly I would like to thank both past and present members of lab. 3.060, particularly Colin and Allen for their technical help, but also Sabrina, Anita, Tammy, Eleanor, Pam, Scott, Craig and Chris for making those long days in the lab much more enjoyable. Thanks also go to my fellow PhD students who were always a valuable source of encouragement and advice especially Sam, Heather and Cassie.

More recently I would like to thank my colleagues at the University of Strathclyde for their endless support. Huge thanks especially go to Gillian, Mark, Richard and Vartul for their encouragement at a time when completion of the thesis seemed unreachable.

I owe thanks to all the girls at Edinburgh Civil Service Ladies Hockey Club for their support, and helping me to vent when it was much needed.

Heartfelt thanks must go to my friends and family, particularly my Mum and Dad, who have always been incredibly supportive and understanding, not to mention patient. Their numerous phone calls have helped through the troublesome stages and they have constantly seen the potential in me to achieve something which at times, I did not think I could, but which I am now proud of.

Last but not least I would like to say a massive thank you to John for his support, understanding, love and kindness. But most of all, I thank him for keeping me calm amd strong and pushing me to stick with it when things got tough, coining the term ‘TUP’ in the process.

Contents

List of Figures

List of Tables

Abbreviations

°C	Degrees Celsius
μF	Microfarad
μg	Micrograms
μl	Microlitres
μM	Micromolar
μm	Micrometers
AHI	Anti haemolysin inhibition
APCs	Antigen-presenting cells
API	Analytical profile index
ATCC	American Type Culture Collection
ATP	Adenosine triphosphate
BHI	Brain Heart Infusion
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
BSA	Bovine serum albumin
C	Carbon atom
c.f.u.	Colony forming units
CAMP	Christie, Atkins and Munch-Petersen
CCDM	*Corynebacterium* Chemically Defined Medium
CD14	Cluster of differentiation 14
CDS	Coding sequence
CLA	Caseous lymphadenitis
cm	Centimetre
CMNR	*Corynebacterium, Mycobacterium, Nocardia* and *Rhodococcus*
Cp	*Corynebacterium pseudotuberculosis*
CP40	Corynebacterial Protease 40
CTLA-4	Cytotoxic T lymphocyte antigen-4
CTLA-4-HIg	CTLA-4 fused to human immunoglobulin G
dATP	Deoxyadenosine triphosphate
dCTP	Deoxycytidine triphosphate
ddH$_2$O	Double distilled water
dGTP	Deoxyguanosine triphosphate
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
dTTP	Deoxytyrosine triphosphate
EDTA	Ethylenediaminetetraacetic acid
ELISA	Enzyme-linked immunosorbent assay
g	Grams
G+C	Guanine and cytosine
gDNA	Genomic DNA
GI	Genomic island
h	Hours
HRP	Horseradish peroxidase
IFN-γ	Interferon-gamma
IgG	Immunoglobulin G

IgM	Immunoglobulin M
IHA	Indirect haemagglutination
KAAS	KEGG Automatic Annotation Server
kb	Kilobases
kDa	KiloDaltons
KEGG	Kyoto Encyclopedia of Genes and Genomes
kg	Kilograms
KO	KEGG ortholgy
kV	Kilovolts
l	Litres
LB	Luria-Bertani broth
LC-ESI MS/MS	Liquid chromatography electrospray ionization tandem mass spectrometry
M	Molar
Mb	Megabases
mg	Milligrams
min	Minutes
Ml	Millilitres
mM	Millimolar
M_r	Relative molecular weight
NCBI	National Center for Biotechnology Information
NCTC	National Collection of Type Cultures
ng	Nanograms
nm	Nanometres
OD	Optical density
ORF	Open reading frame
PAI	Pathogenicity island
PBP	Penicillin binding protein
PCR	Polymerase Chain Reaction
PLD	Phospholipase D
RNA	Ribonucleic acid
rpm	Revolutions per minute
rRNA	Ribosomal RNA
s	Seconds
SAP	Shrimp alkaline phosphatase
SDS-PAGE	Sodium dodecyl sulphate-polyacrylamide gel electrophoresis
SOC	Super Optimal broth with Catabolite repression
TAE	Tris-acetate EDTA
TBS	Tris-buffered saline
tRNA	Transfer RNA
U	Units
UV	Ultra violet
V	Volts
v/v	Volume for volume
w/v	Weight for volume
ΔPLD	Genetically inactivated PLD
Ω	Ohm

Chapter One: General Introduction

1.1 The genus *Corynebacterium*

The *Corynebacterium* genus consists largely of non-pathogenic species which commonly occur in soil and water, although some are part of the normal human skin flora. Non-pathogenic *Corynebacterium* species are important in industry with production of amino acids, such as that of glutamic acid by *Corynebacterium glutamicum*, being the most notable application. However, several species are pathogenic both to humans (such as *Corynebacterium diphtheriae*), and some to specific animal hosts (including *Corynebacterium pseudotuberculosis* and *Corynebacterium bovis*).

1.1.1 General characteristics of the genus *Corynebacterium*

The Corynebacteria are straight to slightly-curved Gram-positive rods that do not form spores and are non-motile. Bacteria belonging to the genus *Corynebacterium* are aerobic or facultatively anaerobic and have a fermentative metabolism, producing lactic acid (but not gas) from certain carbohydrate sources. These organisms are also catalase-positive. When stained, the cells in tissue sections or smears are observed in angular arrangements with "picket fence" or "Chinese letter" configurations which have become characteristic of this genus (Brown & Olander, 1987). Also, when stains such as methylene blue are used, the bacteria may be seen to contain metachromatic granules composed of inorganic polyphosphates which stain red compared to the blue colour of the remaining organism (Brown & Olander, 1987).

The *Corynebacterium* genus belongs to the family Actinomycetaceae which also includes the genera *Mycobacterium*, *Nocardia* and *Rhodococcus* (Songer *et al.*, 1988; Dorella *et al.*, 2006); collectively these are referred to as the 'CMNR' group (Dorella *et al.*, 2006). One feature of all four genera is a high Guanine + Cytosine (G+C) content of up to 74% (Dorella *et al.*, 2006). The CMNR group members also have a specific cell wall organisation characterised by a huge polymer complex consisting of complex lipid components including mycolic acids (Brown and Olander, 1987; Dorella *et al.*, 2006). The length of mycolic acid fatty acid chains varies between genera; hence analysis of mycolic acids has aided the taxonomic classification of members of the CMNR group. Mycobacterial mycolic acids generally consist of the longest chain lengths, ranging from 60 to 90 C atoms (Baird & Fontaine, 2007). Mycolates of organisms assigned to the genera

Rhodococcus and *Nocardia* are somewhat shorter, consisting of between 36 and 66 C atoms (Collins *et al.*, 1982). Mycolic acids of *Corynebacterium* spp., known as corynemycolic acids, are the shortest with chain lengths of 20 to 36 C atoms and are usually either saturated or contain one or two double bonds (Brown & Olander, 1987; Collins *et al.*, 1982). More specifically, it has been reported that *Corynebacterium pseudotuberculosis* (*Cp*) has chain lengths of C_{28} to C_{36} and has predominately saturated C_{14} side-chains (Collins *et al.*, 1982). In addition to their use in taxonomy, mycolic acids are of importance as they have been associated with the virulence of the bacterium. The roles played by these fatty acids in the virulence of *Corynebacterium pseudotuberculosis* are discussed in **section 1.2.2**.

1.1.2 Notable Corynebacteria

1.1.2.1 Non-pathogenic Corynebacteria

C. glutamicum is the most exploited of the genus *Corynebacterium*, and is hugely significant in industry. The reason for the importance of this organism is its ability to overproduce the amino acid L-glutamate (Mono-sodium-glutamate, MSG), which has an annual production of over two million tons (Burkovski, 2008; Kimura, 2003). Shiio *et al.* (1962) demonstrated that excess biotin in culture medium lowers the cellular permeability of amino acids, and hence, low levels of biotin results in overproduction of L-glutamate. It was then noted that *C. glutamicum* can be induced to overproduce L-glutamate even in excess biotin by the addition of certain detergents, such as polyoxyethylene sorbitane monopalmitate (Tween 40), or polyoxyethylene sorbitane monostearate (Tween 60) (Duperray *et al.* 1992). Importantly, the techniques constantly being developed to improve the productivity of *C. glutamicum* may well facilitate the understanding of other corynebacterial species including pathogens of humans and animals.

The closely related *Corynebacterium efficiens* is also used for the production of glutamate. The advantage of using *C. efficiens* is its ability to grow and produce glutamate at 40°C. This considerably reduces the cost otherwise associated with cooling *C. glutamicum* cultures to dissipate the heat which builds up during the glutamate fermentation process (Nishio, *et al.* 2003).

1.1.2.2 The pathogens

The most widely studied and well known species of the *Corynebacterium* genus is the important human pathogen *Corynebacterium diphtheriae*, the aetiological agent of diphtheria. Diphtheria is a disease of the upper respiratory tract and is transmitted by direct physical contact or inhalation of aerosolized secretions from an infected individual or asymptomatic carrier. The disease is characterised by a sore throat, low fever and loss of appetite, and as the infection develops, a greyish membrane forms in the throat usually affecting the tonsils, pharynx, and/or nasal cavity. The membrane, known as a pseudomembrane, consists of fibrin, bacteria and inflammatory cells and can reduce the air flow through the trachea and may eventually result in complete blockage, causing suffocation (Hadfield *et al.*, 2000). Once *C. diphtheriae* infection has become established, diphtheria toxin is produced which can cause damage to visceral organs. Holmes (2000) reported that when large doses of the exotoxin were injected parenterally into susceptible animals typical systemic lesions of diphtheria were observed, including myocarditis, polyneuritis, and focal necrosis in the liver, kidneys and adrenal glands. Fortunately, diphtheria has largely been eradicated in developed countries for many years through a successful vaccination programme. However, some re-emergence has since occurred and the epidemic experienced in the Newly Independent States (NIS) of the former Soviet Union in the 1990s (Golaz *et al.*, 2000) demonstrated the continued threat of this thought to be rare disease. Vaccination is achieved by administering toxoided diphtheria toxin, a formalin-inactivated derivative, as part of the diphtheria–pertussis–tetanus (DPT) vaccine.

Corynebacterium ulcerans is associated with disease in cows and domestic animals such as cats and dogs. More recently, *C. ulcerans* has been reported as a cause of diphtheria-like disease in humans, often associated with companion animals (Berger *et al.*, 2011). Indeed *C. ulcerans* infection was found to be responsible for more cases of toxigenic diphtheria in the UK between 2004 and 2008 than *C. diphtheriae* infection (Wagner *et al.*, 2010). *Corynebacterium pseudotuberculosis* can also potentially produce diphtheria toxin, although this is not common; Wagner *et al.* (2010) reported a single case (of a total of 102) of diphtheria-like infection caused by toxigenic *C. pseudotuberculosis. C. pseudotuberculosis* will be discussed in greater detail in **section 1.2**.

Another pathogenic *Corynebacterium* is the opportunistic pathogen *Corynebacterium jeikeium* which is part of the normal human skin flora. *C. jeikeium* has however been recognised as a nosocomial pathogen and can be responsible for disease such as septicaemia and endocarditis in immunocompromised hosts (Tauch *et al.*, 2005). Treatment of infected patients can be problematic largely due to the organism's resistance to a number of antibiotics (Soriano *et al.*, 1995).

1.2 *Corynebacterium pseudotuberculosis*

The first organism resembling *Corynebacterium pseudotuberculosis (Cp)* was described following its isolation from a case of lymphangitis in a cow by the French veterinarian Edmond Nocard in 1888 (Brown & Olander, 1987). A similar bacterium was identified three years later in cultures from a renal abscess in a ewe by Hugo von Preïsz, and the organism then became known as the *Preïsz-Nocard bacillus* (Baird & Fontaine, 2007). The bacterium was later known as *Bacillus pseudotuberculosis*, derived from 'pseudes tuberculosis' meaning false tuberculosis, referring to the apparent similarity of resulting lesions with those produced by *Mycobacterium tuberculosis* (Baird & Fontaine, 2007; Brown & Olander, 1987). The morphology of the organism distinctly resembled that of *C. diphtheriae* which led to a change of classification to *Corynebacterium,* as well as name to *Corynebacterium ovis*, referring to the predominant host species. However, it was further renamed by Eberson in 1918 to *Cp* following isolation of the bacterium from species other than sheep (Euzeby, 2005). The word 'corynebacterium' stems from the Greek words "coryne" meaning "club" and "bacteria" for "rod", hence *Cp* is a club-shaped rod which is capable of producing abscesses in several mammalian species resembling those of a *M. tuberculosis* infection .

1.2.1 Phenotypic characterisation

Typical of its genus, *Cp* is a Gram-positive, non-motile pleomorphic rod ranging in size from 0.5-0.6 µm by 1.0-3.0 µm, and forms characteristic palisade arrangements as mentioned above (Brown & Olander, 1987; Connor *et al.*, 2000; Selim, 2001). *Corynebacterium pseudotuberculosis* is catalase-positive but negative for oxidase production (Brown and Olander, 1987; Dorella *et al.*, 2006). Early reports showed that *Cp* isolates from different host species were identical biochemically, with the exception of their ability to reduce nitrate. Later Biberstein, Knight, & Jang (1971) suggested the existence of two biochemically diverse biovars of *Cp*: the biovars *ovis* and *equi*. Generally biovar *equi,* isolated from horses and cattle, is capable of reducing nitrates to nitrites, whereas the *ovis* biovar infects sheep and goats and is unable to reduce nitrate. Songer *et al.* (1988) also drew this conclusion following characterisation of sheep, goat, horse and cattle isolates from around the world by biochemical analysis and also by restriction endonuclease analysis of DNA. Interestingly, this group found that host preference marked by nitrate reductase production may not exist in cattle like the other host species studied. *Corynebacterium pseudotuberculosis* strains have been typed using pulsed-field gel electrophoresis (discussed further in **section 1.7**; Connor *et al.*, 2007; Connor *et al.*, 2000) and this has more recently confirmed the distinction of the two biovars. Indeed it would appear *Cp* isolates can vary in their biochemical properties, particularly in their ability to ferment carbohydrates; all strains produce acid but not gas from a range of carbon sources including glucose, maltose, fructose mannose and sucrose.

Corynebacterium pseudotuberculosis will grow either aerobically or anaerobically at 37°C. When grown on solid media, *Cp* colonies are dry, white (or pale) in colour and whole colonies can easily be pushed across the surface of the agar due to the waxy nature of the mycolic acid coat (Quinn *et al.*, 2009). *Corynebacterium pseudotuberculosis* growth benefits from addition of serum or whole blood to the growth media, and when whole blood is used as a supplement, β-haemolysis can be observed around bacterial colonies (although this may not be visible until up to 72 h) (Quinn *et al.*, 2009). Although *Cp* can be successfully cultured in liquid media, the organism tends to clump together; this property is thought to be related to the presence of fatty acids in the cell wall (Carne *et al.*, 1956). In the laboratory, an emulsifier such as Tween 80® can be added to liquid media to prevent this clumping.

Of the other Corynebacteria, it has been suggested that *Cp* is most closely-related to *Corynebacterium ulcerans*, predominantly because both species are unique from other coryneforms in their shared ability to produce phospholipase D (McNamara *et al.*, 1995). Also some strains of these two species can be lysogenised by corynephage β (which confers the ability to produce diphtheria toxin to *C. diphtheriae*) and *Cp* and *C. ulcerans* are the only species besides *C. diphtheriae* capable of producing that toxin (Groman *et al.*, 1984; Maximesc *et al.*, 1974a; Maximesc *et al.*, 1974b).

Cp is the causal agent of a number of diseases in several host species (**section 1.2.3**) including caseous lymphadenitis (CLA), a chronic disease primarily found in small farmed ruminants, which will be discussed in more detail in **section 1.3**.

1.2.2 Virulence factors of *C. pseudotuberculosis*

The virulence mechanisms of *Cp* are, as yet, poorly understood. Significantly, as no plasmids have ever been identified within *Cp* isolates (Baird and Fontaine, 2007), it is widely accepted that all virulence determinants are encoded within the chromosome. The vast majority of research over the years has been focussed on just two virulence factors, namely Phospholipase D and mycolic acid. The opportunity to identify further novel virulence determinants will be greatly enhanced by the sequencing and studying of the *Cp* genome.

Phospholipases are a group of enzymes capable of hydrolysing glycerophospholipids. There are four main classes of phospholipases, phospholipases A-D, and each group is classified depending on the specific ester bond that is cleaved within the substrate molecule. Eukaryotic cell membranes consist largely of phospholipids, and phospholipases are involved in membrane maintenance and the cellular inflammatory response (Schmiel & Miller, 1999). Phospholipases are secreted by several genera of bacteria and have been shown to be involved in virulence (Schmiel & Miller, 1999). The bacterial phospholipases target the phospholipids in eukaryotic cell membranes during invasion of host tissues, facilitating dissemination of infection (Ghannoum, 2000).

Phospholipase D (PLD) has been detected in every *Cp* isolate studied (Songer *et al.*, 1988) and the exotoxin has been implicated as the major virulence factor of *Cp* for some time (Batey, 1986b; Hodgson *et al.*, 1994). PLD inhibits the lysis of erythrocytes induced by staphylococcal lysin (Zaki, 1976) and what is more, when in the presence of an extracellular *Rhodococcus equi* factor, PLD has a synergistic effect on the lysis of erythrocytes (Fraser, 1961). The necessity of PLD for caseous lymphadenitis establishment has been credibly demonstrated in studies of *Cp* strains with inactivated PLD. *Corynebacterium pseudotuberculosis* isolates in which the *pld* gene has been deleted or inactivated by mutation are not capable of causing the abscessation classically observed in infected animals. One such study was performed by McNamara *et al.* (1994) who replaced the *pld* gene with an allele containing a nonsense mutation; the outcome was a reduced ability of this mutant to establish a primary infection compared to the wild-type strain used and it was also incapable of dissemination. The *pld* gene of *Cp* has been cloned and sequenced, and was found to encode a 31.4 kDa protein which is preceded by a probable secretory signal peptide sequence (Hodgson *et al.*, 1990). Egen *et al.* (1989) showed that large quantities of PLD can be collected for studies employing the enzyme using recycling isoelectric focusing. There are reports of several biological activities of PLD, and that with greatest significance to virulence is sphingomyelinase activity similar to that of sphingomyelinase prepared from *Staphylococcus aureus*; PLD catalyses the dissociation of sphingomyelin into ceramide phosphate and choline (Bernheimer *et al.*, 1980; Hodgson *et al.*, 1990). This hydrolysis of sphingomyelin causes increased vascular endothelial membrane permeability, which in turn leads to plasma leaking from blood vessels into surrounding tissues and subsequently into the lymphatic drainage (Jolly, 1965). This process may favour the lymphatic drainage of *Cp* in tissue fluid and hence aid pathogenesis (Batey, 1986b). PLD also activates complement via the alternative pathway, depleting complement components from the region surrounding the developing bacterial colony and limiting opsonisation (Yozwiak & Songer, 1993). Neutrophils treated with PLD failed to migrate to complement compounds indicating the enzyme may impair ovine neutrophil chemotaxis and consequently decrease the probability of phagocytosis in early infection (Yozwiak & Songer, 1993). Other activities of PLD include dermonecrosis and lethality (Brogden & Ester, 1990; Muckle & Gyles, 1986).

Corynebacterium pseudotuberculosis does not have a capsule and instead (as mentioned above) has a waxy coat comprised of mycolic acid, which reportedly has cytotoxic properties (Muckle & Gyles, 1983). It is thought that mycolic acid may act to increase pathogenicity in a number of ways. When mycolic acid extracted from *Cp* is subcutaneously injected into mice, a noticeable reaction is observed as swelling, congestion and a central region of haemorrhagic necrosis (Carne *et al.*, 1956). The same researchers also found when mycolic acid was phagocytosed by leucocytes, that degenerative changes were induced followed by the death of the cell (Carne *et al.*, 1956).

Corynebacterium pseudotuberculosis has relative resistance to environmental conditions and this survival of *Cp* in the environment may be assisted by the mycolic acid coat. Soil inoculated with pus and stored at a variety of ambient temperatures contained viable organisms for up to 8 months (Brown & Olander, 1987). Augustine and Renshaw (1986) inoculated several inanimate surfaces and particulate fomites (including hay, straw and faeces) with *Cp* in purulent exudate sourced from a naturally-occurring case of caprine

CLA. Fomites were then incubated at various temperatures and the *Cp* survival time was determined by isolation of viable bacteria from the fomite. The results of this study showed incubation at lower temperatures generally extended the bacterial survival. Also *Cp* organisms remained viable longer when pus was mixed on fomites than when spread on surfaces with a mean isolation period of 1-8 days on surfaces and 7-55 days from fomites (Augustine & Renshaw, 1986).

During the process of a natural infection, the mycolic acid coat provides a means of mechanical and possibly biochemical protection, helping to prevent enzymatic degradation within the lysosome. The ability to resist hydrolysis in lysosomes enables the bacterium to survive phagocytosis and remain intracellular (Baird, 2007; Baird & Fontaine, 2007). It is likely that the organism uses this mechanism to migrate to the site of lesion development. Also Muckle and Gyles (1983) found a direct relationship between the percentage of cell wall lipid produced by *Cp* isolates and the induction of chronic abscessation in artificially infected mice, demonstrating a role of mycolic acid in abscess formation and hence pathogenicity.

Other possible virulence factors may include a 40 kDa protein antigen, the identification of which was reported by Walker *et al.* (1994) using the application of a strategy that employed locally derived antibody-secreting cells. The protein was shown to be immunogenic and results of field trials indicated it to be highly protective (Walker *et al.*, 1994). Subsequently the protein was characterised as a serine protease, however database searches revealed a lack of homology with other known serine proteases suggesting this to be a novel protein and the authors proposed the antigen be termed corynebacterial protease 40 (CP40).

1.2.3 Diseases caused by *Corynebacterium pseudotuberculosis*

The most notable disease caused by *Cp* is caseous lymphadenitis (CLA) of small ruminants, particularly sheep and goats (this will be discussed in detail in **section 1.3**). However *Cp* is responsible for a number of infections in other animals including horses, cows and camelids.

Corynebacterium pseudotuberculosis infection in horses can cause ulcerative lymphangitis which affects the distal limbs and is generally thought to result from poor husbandry of animals (Brown & Olander, 1987). In this condition subcutaneous lesions rupture and subsequently form necrotic ulcers (Radostits *et al.*, 2000). Deep-seated abscessation (pigeon fever, Wyoming strangles, false strangles) is the most serious equine condition and is economically significant in much of western North America (Hall *et al.*, 2001; Hughes & Biberstein, 1959). It is typically characterised by lesions of the pectoral musculature although occasionally lesions are discovered in the ventral abdominal area (Hughes & Biberstein, 1959). The distribution of equine *Cp* disease is concentrated in western Canada and the US and the incidence of equine *Cp* infection in these areas is highest in the months of September, October and November (Aleman *et al.*, 1996). This reported seasonal variation may reflect the seasonal incidence of various arthropods implicated as disease vectors. Spier *et al.* (2004) conducted a real-time PCR-based analysis of flies to detect the presence of the *pld* gene which would indicate carriage of *Cp*. Three species of fly were identified as potential vectors for disease transmission, including up to 20% of examined houseflies (*Musca domestica*) from the vicinity of infected horses (Spier *et al.*, 2004).

Cp infections in Israeli dairy cattle have been reported, commonly characterised as deep subcutaneous abscesses that develop into ulcerative granulomatous lesions; normally occurrence of this condition is sporadic, although epidemic infections can also occur (Yeruham *et al.*, 2003). Indeed morbidity rates of up to 35% have been accounted (Yeruham *et al.*, 1997). *Corynebacterium pseudotuberculosis* has also been recognised as a cause of mastitis, which can have a range of outcomes from decreased milk production, to the more severe result of culling (Shpigel *et al.*, 1993).

In South America *Cp* commonly causes purulent lymphadenopathy which affects high numbers of farmed alpacas and llamas (Braga *et al.*, 2006), while in North America sporadic *Cp* infections have been encountered in companion camelids (Anderson *et al.*, 2004). The organism causes oedematous skin disease in buffaloes in the Middle East (Selim, 2001). Also, *Cp* has been isolated from sporadic naturally occurring infections in pigs (Zhao *et al.*, 1993) amongst other animal species (Baird & Fontaine, 2007).

Although *Cp* is predominantly an animal pathogen, it is potentially a zoonotic agent, and several cases of human infection have been reported (Mills *et al.*, 1997; Peel *et al.*, 1997). Mills *et al.* (1997) reported a case of suppurative lymphadenitis in an adolescent boy who had frequent contact with sheep on a family member's farm in Mallee, Victoria (Australia), though this could not be confirmed as the source of exposure. The patient's treatment consisted of excision of the right axillary lymph nodes and administration of various antibiotics (Mills *et al.*, 1997). A further case of necrotising lymphadenitis caused by *Cp* occurring in a twelve-year-old patient was described by Join-Lambert *et al.* (2006). The young patient had also been in contact with sheep and the causative agent was confirmed to be *Cp* using the 'API Coryne Kit' (**section 1.6.2**). This case is one of few reports of the zoonosis in Europe (Join-Lambert *et al.*, 2006). Also some years earlier, Peel *et al.* (1997) described 10 additional cases of the disease in humans to the 12 previously published, and the majority of all reported cases have been in people occupationally exposed to sheep, including farm and abattoir workers. Interestingly, this study represented a single Australian state (Victoria), yet almost doubled the number of previously reported cases worldwide which may indicate human lymphadenitis caused by *Cp* is more common than published data would suggest. Treatment of human cases has primarily been by antibiotic therapy and excision of the affected lymph nodes, an option unsuitable for most animals due to the labour and cost intensive nature of this procedure

1.3 Caseous lymphadenitis

1.3.1 Introduction

Caseous lymphadenitis (CLA) is a chronic disease prevalent in small ruminants such as goats and sheep. The disease is widespread across the world, although incidence is significantly higher in the majority of countries where intensive husbandry of small ruminants is practised. CLA is characterised by the development of pyogranulomatous lesions in the lymphatic system and visceral organs. Although the perceived importance of CLA varies worldwide, the disease can be responsible for considerable economic loss to the sheep and goat industries in affected areas.

1.3.2 Distribution and prevalence of CLA

CLA is of economic consequence in many countries, including North and South America, Australia, New Zealand, Europe and South Africa (Williamson, 2001). In 2002, an abattoir and questionnaire survey in Australia was conducted and it reported a prevalence of CLA in Western Australia of 20%, 23% in Victoria and 29% in New South Wales (Paton *et al.*, 2003). Although such a country-wide report of CLA incidence does not appear common, smaller scale studies have been reported for areas in other continents. A study conducted in the single state of Minas Gerais in Brazil estimated frequency of *Cp* infection to be 44% and farm frequency 100% (Guimaraes *et al.*, 2011). In Iran, the frequency of CLA in 468 sheep slaughtered in an abattoir in Tabriz was reported to be 13% based on bacteriological culture and 20% on histopathological study, 19% of animals had CLA confirmed by both bacteriological and histopathological studies (Zavoshti *et al.*, 2011).

It is thought that CLA was introduced into the UK in 1990 via an infected goat from Europe (Lloyd *et al.*, 1990). Since that time CLA has become increasingly common in terminal sire sheep breeds with at least 18% of flocks thought to be affected by the disease (Baird *et al.*, 2001). Disease incidences have been reported across the British Isles, Northern Ireland and the Republic of Ireland in the years following this initial prevalence assessment, and indicate the prevalence report of the UK in 2001 is likely an under-representation (Baird *et al.*, 2004). Indeed rather than being restricted to breeding animals CLA has become well established within the national flock (Baird *et al.*, 2004).

1.3.3 Economic significance of CLA in sheep

Partial or total condemnation of carcases is the most direct cost associated with CLA, in addition to other costs incurred at the abattoir by way of additional meat inspections and carcase trimming; in Australia the estimated losses at abattoirs has been estimated between $12 and $15 million per annum (Paton et al., 2003). CLA also appears to have a detrimental effect on the production of wool. In one study, unvaccinated sheep challenged with *Cp* produced 0.20 kg less clean wool than those unchallenged during the following 12 months (Paton *et al.*, 1988). Paton *et al.* (1994) assessed loss of wool production in a comparison study between *Cp*-infected and uninfected animals in three Australian sheep flocks. A 4.1 to 6.6% decrease in clean wool production and 3.8 to 4.8% decrease in greasy wool production was reported (Paton *et al.*, 1994). The authors also reported the approximate cost of CLA to the Australian sheep industry through lost wool production as $17 million annually (Paton et al., 1994). It was also found that vaccinated sheep had a higher bodyweight than unvaccinated animals 12 months post-challenge (Paton *et al.*, (1988).

In North America, CLA has been associated with a debilitating condition in mature sheep known as ‘thin-ewe syndrome’, which occurs despite the absence of clinical disease. It has been concluded that CLA in this form has a economically significant effect on both reproductive and culling rates in ewes (Renshaw *et al.*, 1979). However, although *Cp* is the principal causative agent, other bacterial species including *Moraxella*, *Staphylococcus* and *Pseudomonas* may be present in a mixed culture. Furthermore, occasionally *Moraxella* species can be isolated from infected tissue in the complete absence of *Cp* (Renshaw *et al.*, 1979). Other economic losses include reduced sales of breeding stock and increased rates of mortality (of particular impact within breeding stocks) (al-Rawashdeh & al-Qudah, 2000).

1.3.4 Pathogenesis and transmission

Corynebacterium pseudotuberculosis infection generally occurs via broken skin or mucus membrane wounds, particularly those introduced by shearing. The bacterium has also been reported to gain entry to the host through intact skin (Nairn & Robertson, 1974) but this requires further substantiation. Following entry into the host animal, *Cp* migrates to the local drainage lymph nodes where microscopic pyogranulomas develop which increase in size and ultimately combine to form larger abscesses. It has been postulated that the migration of the organism to the lymph node is achieved by carriage within phagocytic cells, such as macrophages (Baird & Fontaine, 2007). Indeed the organism is not only capable of surviving but also multiplying within macrophages (Williamson, 2001) which may aid the dissemination of infection via the lymphatic or blood system to other parts of the body, such as the viscera. The early phase of *Cp* infection in lambs was studied by Pepin *et al.* (1991) who described the infiltration of primary abscesses by polynuclear leucocytes (particularly neutrophils). From day three of infection the lesions then became enlarged and transformed to typical pyogranulomas with a central zone of necrosis and a peripheral mantle of mononuclear cells (neutrophils largely being replaced by an increase in macrophages) (Pepin *et al.*, 1991). This change is associated with a persistence of bacteria and their dissemination to other sites (Pepin *et al.*, 1991). When present in superficial lymph nodes, CLA abscesses progressively become swollen and encapsulated within fibrous tissue, and may finally result in the loss of overlying hair and lesion rupture (Radostits et al., 2000). A constant degradation and re-synthesis of the fibrous capsule surrounding the lesion produces distinct layers, resulting in an ‘onion ring’ appearance of cross-sectioned mature CLA abscesses (Batey, 1986b).

CLA lesions can be present in two forms: external and visceral, however, co-existence within a single host is possible. The cutaneous (external) form of CLA is characterised by abscesses in the superficial lymph nodes and subcutaneous tissues. In the visceral form of the disease lesions are present deep within the animal's tissues. It commonly affects both the internal lymph nodes and visceral organs, in particular the mediastinal lymph nodes and lungs but also the liver, kidneys, mammary glands and others less frequently (Fontaine & Baird, 2008; Papaioannou *et al.*, 2010; Stoops *et al.*, 1984). In a study by Stoops *et al.* (1984), the lung was found to be the most commonly affected visceral organ, with 13.08% of 4,089 animals having lung lesions compared to 11.79% having lesions of abdominal viscera. There are many similarities in the way CLA is manifested in sheep and goats; in both species this more severe visceral form is thought to be less frequent and primarily the superficial lymph nodes are affected, which is consistent with entry through the skin and drainage to the nearest local drainage lymph node. In the UK CLA lesions are commonly associated with the head and neck area (Baird & Malone, 2001), however, this is not the case in other parts of the world, in Australia the superficial lymph nodes of the shoulder and flank are commonly affected. Interestingly in other countries the lesion pattern observed in UK sheep is more-commonly associated with goats. The visceral form of the disease commonly observed in the USA, where it is regarded the most common cause of thin ewe syndrome, has been sparsely reported in the UK or elsewhere.

A possible respiratory route of infection for *Cp* has also been proposed (Stoops *et al.*, 1984), suggesting some sheep with affected lung tissue may have inhaled organisms in the form of an infectious aerosol, leading to colonisation of the lung parenchyma and the development of lesions. Severe lesions of the lungs not only diminishes the functional capacity of the respiratory system but can also increase the animals susceptibility to systemic infections (Stoops *et al.*, 1984). In addition, abscessation within the mediastinal lymph node has been postulated to exert pressure on the trachea, inhibiting normal rumination (Paton *et al.*, 2005).

In many parts of the world, shearing of sheep is thought to be the principal factor in CLA transmission. The shearing process can rupture superficial abscesses, releasing bacteria onto the skin and fleece and thus contaminating shearing equipment and other fomites such as fence posts. The concentrations of viable organisms contained within pus are estimated between 1×10^6 and 5×10^7 c.f.u/g (Brown & Olander, 1987), so the potential for environmental contamination from the purulent contents of even a single ruptured lesion is very high. Further animals may then become infected by direct physical contact with infected individuals, or via contaminated fomites (Fontaine & Baird, 2008). Also, wounds created by the shearing process provide an easy entry for *Cp* into the host as the skin barrier is broken. The effect of shearing on CLA incidence was studied in Awassi sheep in Jordan by al-Rawashdeh and al-Qudah (2000); they found that CLA prevalence increased with age and the incidence of the disease did indeed increase after shearing but only in sheep aged 1-2 years. This data provides evidence that there are multiple factors contributing to the occurrence of CLA. Importantly, shearing practices were also investigated within this study via farmers participating in a questionnaire, which revealed sheep were generally sheared under unhygienic conditions; none of the farmers disinfected shearing equipment and all farmers made wounds during the shearing process (al-Rawashdeh & al-Qudah, 2000). These practices and also the possible shared usage of equipment between farms will ultimately facilitate the transmission of CLA both within and between flocks.

Commercial dips used for lice control have been implicated as a means of CLA transmission. Nairn and Robertson (1974) researched the possibility of CLA establishment via unbroken skin and also the potential involvement of sheep dipping fluids. Two experiments were performed by this group; in the first *Cp* was placed directly onto a freshly shorn shoulder of sheep following pre-treatment with either water, a defatting agent or one of four different sheep dips used commercially; the second experiment attempted to increase infection rate using the same application technique. In the first experiment it was found that CLA could be reproduced in sheep by placing *Cp* on apparently-unbroken shorn skin, however low animal numbers meant the result was not statistically significant. The second experiment indicated no significant increase in disease when culture was applied to the skin on four consecutive days, however, it did reveal that incidence of induced CLA increased to 100% when *Cp* was applied to a larger surface area (Nairn & Robertson, 1974). Paton *et al.* (1996) found both dipping sheep and keeping the animals under cover for one h or more post-shearing increased the incidence of CLA. In a later study the same group investigated whether the interval between shearing and dipping of sheep had any effect on the spread of *Cp* infection (Paton *et al.*, 2002). In this study, 195 sheep were confirmed to be naive to *Cp* exposure by an assay of CLA toxin antibody; the sheep were then shorn and dipped at varying intervals in a solution containing *Cp*. It was found that the occurrence of CLA abscesses did not differ between groups shorn at 0, 2, 4 or 8 weeks prior to dipping, hence delaying dipping after shearing did not decrease *Cp* infection.

However, those sheep that were dipped immediately following shearing did develop higher concentrations of antibodies against *Cp* toxin and cell wall material than those sheep whose dipping was delayed (Paton *et al.*, 2002). Significantly, control of ectoparasites through the use of dips is no longer permitted in the UK, so this mode of *Cp* transmission has been (unwittingly) eradicated.

1.4 Treatment of CLA in sheep

The intracellular persistence of *Cp* makes CLA difficult to treat as the detection of infected animals is somewhat inadequate and drug therapy is of little effect; there is a general consensus that clinical CLA is refractory to antibiotics (Baird, 2006). However, Senturk and Temizel (2006) described a study in which the treatment of naturally occurring ovine CLA did appear successful using an antibiotic course of rifamycin given in combination with oxytetracycline. *Corynebacterium pseudotuberculosis* was confirmed to be the causative agent of disease in ten sheep admitted with clinical symptoms of CLA (enlarged superficial lymph nodes). These sheep were then administered Rifamycin Sv (10 mg/kg bodyweight) twice daily for ten days alongside oxytetracycline (20 mg/kg bodyweight) every three days for three doses. The authors classed this treatment as successful as the active lesions within the sheep resolved and no recurrence of CLA was observed in the follow-up period post treatment. However follow-up only lasted for one month and no indication was given of how long after this period animals remained free of CLA; what is more antibiotics used in this study are not licensed for veterinary use in the UK and resistance to rifamycins develops rapidly (Baird, 2006). It is thought antibiotics are incapable of penetrating the encapsulations around pyogranulomas and pus within lesions (Williamson, 2001). Olson *et al.* (2002) investigated antibiotic resistance of *Cp* when grown as a biofilm to more accurately reproduce the conditions of a natural infection. They reported that *Cp* in the planktonic state was susceptible to all antibiotics tested at minimal concentrations with the exception of streptomycin, which had a higher minimum inhibitory concentration of 256. Despite this, when grown as a biofilm the bacterium was highly resistant to all drugs tested (Olson *et al.*, 2002). It is as yet unknown whether *Cp* forms biofilms *in vivo,* but if this is the case, targeting the organism *in vivo* using antibiotic therapy is unlikely to be successful and hence a questionable course of action to take, certainly as a stand-alone measure, when attempting to eliminate disease.

1.5 Vaccines against CLA

As treatment of CLA with antibiotics is not practical, vaccination of animals at risk of contracting the disease may seem a more sensible approach, and is certainly one which is desirable to the sheep/goat producing industry. A significant amount of work on immunisation against *Cp* has been conducted over several decades, and several commercial CLA vaccines exist; however, there is currently no licensed vaccine available for routine use in the UK or indeed the whole of the European Union. Despite this fact, in cases where CLA is perceived to be sufficiently serious, permission may be granted by the Veterinary Medicines Directorate (VMD) for restricted use of unlicensed vaccines ([Anon], 2012). This permission allows the use of certain vaccines within a single affected flock or 'holding'.

1.5.1 Bacterin vaccines

Until recently, the CLA vaccines for which permission for use was most-frequently sought were made-to-order, so-called 'autogenous' vaccines. These are derived from *Cp* strains isolated from pus of affected animals, which are cultured in a laboratory and inactivated prior to addition of an approved adjuvant to create a bacterin vaccine. The (arguably flawed) theory behind the use of these vaccines is that all of the cases of CLA within a given flock will have derived from a single originator animal, and that a strain-specific (autogenous) vaccine will therefore be particularly efficacious at preventing further spread of that strain to naïve animals within the flock (Fontaine *et al.*, 2006). Evidence to support the efficacy of autogenous CLA vaccines is largely anecdotal. However, in one study designed to assess an autogenous vaccine, Fontaine *et al.* (2006) investigated the use of a formalin-killed *Cp* isolate adjuvanted with aluminium hydroxide. The vaccine was shown to confer significant protection against *Cp* infection and prevented dissemination of the homologous *Cp* challenge strain beyond the site of inoculation in an ovine experimental model of CLA (Fontaine *et al.*, 2006).

There has been a lack in studies of protection conferred by *Cp* bacterins against CLA in the field, however, Menzies *et al.* (1991) conducted a field trial of a whole-cell vaccine. The trial was performed over a period of three years in both a sheep flock and a goat herd.

Juvenile sheep and goats were vaccinated between 2.5 and 3.5 months old, and booster vaccines were administered at one and eleven months post initial immunisation. A persistently increased serum antibody concentration was observed for the full length of the trial in both sheep and goats. The vaccine also provided statistically significant protection against clinical CLA in sheep and although not significant, a similar effect was suggested in goats (Menzies *et al.*, 1991) with low animal numbers remaining in the trial following attrition losses given as the reason behind the insignificance.

Later Brogden *et al.* (1996) used a bacterin containing muramyl dipeptide in light mineral oil to vaccinate both lambs and kids. However the results were inconclusive and these authors also assigned the poor significance levels to low numbers of remaining animals (Brogden *et al.*, 1996).

1.5.2 Toxoid vaccines

Management strategies in Australia failed to prevent CLA spreading so in the 1970s efforts were concentrated on developing an effective vaccine and these studies eventually led to the development of the first commercial CLA vaccine, known as Glanvac™ (Eggleton *et al.*, 1991a; Eggleton *et al.*, 1991b). It had previously been reported that the *Cp* exotoxin, PLD, conferred protection against CLA. Subsequently technology became available to the Commonwealth Serum Laboratories (Parkville, Victoria, Australia) to produce an effective toxoid vaccine and details on antigen dose and combination with clostridial components were published in 1991 (Eggleton *et al.*, 1991a). Glanvac™ was released some years earlier in 1983 (Eggleton *et al.*, 1991b) and is a combined clostridial and corynebacterial vaccine which offers protection against CLA, tetanus, black disease, malignant oedema and blackleg in sheep and goats; currently, the vaccine is manufactured and marketed by Pfizer Animal Health.

1.5.3 Combined vaccines

During the development of Glanvac™, Eggleton *et al.* (1991c) reported no difference in protection when toxoid vaccine was supplemented with *Cp* cells. However, another commercially available vaccine against CLA, Caseous D-T™ (Colorado Serum Company, Denver, CO, USA), contains a combination of *Cp* bacterin and toxoid as well as *Clostridium tetani*, and *Clostridium perfringens* type D toxoids also offering protection against clostridial infections in addition to CLA. Piontkowski and Shivvers (1998) reported that this combined vaccine conferred some protection against experimental infection and reduced the incidence of CLA abscesses. However the vaccine remains unlicensed for use in the UK or anywhere outside of the US. In the UK vaccination of sheep with a combined *Cp* bacterin and recombinant PLD was reported to result in complete protection against *Cp* infection 3 weeks post experimental homologous challenge (Fontaine *et al.*, 2006).

More recently, İzgür *et al.* (2010) reported significant results from a combined bacterin and toxoid vaccine trialled in Turkey. The group found that development of CLA lesions and *Cp* isolation was significantly less ($P < 0.001$) in vaccinated lambs than in control animals.

1.5.4 Live vaccines

The potential value of an attenuated vaccine against CLA was indicated by Pepin *et al.* (1993) who reported an experimental infection in which *Cp* infected sheep were protected from further challenge, despite remaining carriers. Further support for the use of a live vaccine was demonstrated using PLD mutants: as PLD had been determined as an important virulence factor of *Cp*. Hodgson *et al.* (1992) hypothesised that inactivation of the *pld* gene may provide a basis for a live recombinant vaccine. Subsequently this group demonstrated that a single subcutaneous inoculation with up to 10^7 c.f.u of a PLD-deficient *Cp* strain designated Toxminus resulted in an absence of clinical disease associated with wild-type *Cp,* hence the Toxminus mutant strain was rendered incapable of inducing CLA (Hodgson *et al.*, 1992). However, at doses higher than 10^7 c.f.u transient abscesses did develop at the site of inoculation. Later the same group evaluated the use of Toxminus as a live oral vaccine in order to overcome such inoculation site reactions. They postulated that usefulness of the *pld*$^-$ mutant as a live oral vaccine was reduced as one of the major immunodominant antigens had been removed (Hodgson *et al.*, 1994). Therefore Hodgson et al. (1994) constructed a genetically inactivated PLD analogue by mutating the histidine$_{20}$ amino-acid residue to tyrosine to remove enzymatic activity of the protein. Two vaccines, Toxminus and Toxminus carrying this plasmid-borne inactivated *pld*, were administered to sheep orally and both vaccines induced significant specific antibody responses. Interestingly only animals vaccinated with the Toxminus carrying inactivated *pld* were protected against wild-type challenge and protection conferred by the Toxminus strain was negligible which contrasted with the researchers' previous work (Hodgson *et al.*, 1992).

This difference was attributed to the different routes of delivery and the inability of oral vaccination to stimulate a significant Th1 immune response (Hodgson *et al.*, 1994).

Further promise for a live CLA vaccine has been reported by others: attenuated mutants of two *Cp* strains were created by disrupting the *aroQ* gene which encodes 3-dehydroquinase, an enzyme essential in aromatic amino acid synthesis (Simmons *et al.*, 1997). One attenuated strain possessed and one lacked PLD and both strains were cleared from spleens and livers of challenge mice by eight days post-infection. The virulence of mutant strains was restored to some degree by the introduction of a plasmid encoding the wild-type *aroQ* gene. It was revealed that one of the attenuated mutants induced some protection against wild-type challenge and stimulated interferon-γ production by murine splenocytes when present at high doses (Simmons *et al*., 1997).

1.5.5 DNA vaccines

Another approach to producing an effective CLA vaccine, albeit less well studied, has been through the use of DNA antigens. Boyle, Brady and Lew (1998) developed a method of directing antigen to immune induction sites and particularly antigen-presenting cells (APCs) by fusing antigen-encoding DNA to cytotoxic T lymphocyte antigen-4 (CTLA-4). CTLA-4 is expressed on activated T cells and binds to CD80 and CD86 (B7 domains) on APCs (Boyle *et al.*, 1998). A derivative of CTLA-4 was created, designated CTLA-4-HIg, it was composed of human IgG fused to the extracellular domains of CTLA-4. This CTLA-4 derivative was found to be capable of dimerisation and proteins consisting of antigen-encoding sequences fused to CTLA-4-HIg were targeted to APCs specifically (Boyle *et al.*, 1998). Furthermore, genetically inactivated PLD (ΔPLD) antigen was shown to confer protection when delivered as a DNA vaccine intramuscularly, and when fused to CTLA-4-HIg the efficacy of the vaccine was increased (Chaplin *et al.*, 1999). Indeed the same research group later produced the only report of the development of a DNA vaccine protecting against *Cp* infection (De Rose *et al.*, 2002). As in the previous accounts, this report studied a vaccine consisting of plasmid DNA encoding ΔPLD antigen linked to CTLA-4-HIg and investigated protection gained from the vaccine using different immunisation routes. The authors found the route of immunisation to have a significant effect upon immune responses and protection of sheep to a live *Cp* challenge. Intramuscular injection of the vaccine induced a strong memory response and sterile immunity in 45% of challenged animals, a similar result to the protection observed with protein vaccination. However, animals which received subcutaneous injection or gene gun delivery of the vaccine had insignificant levels of protection (approximately 10%) against *Cp* challenge (De Rose *et al.*, 2002).

1.6 Diagnosis of CLA

The most important measures that can be taken to control CLA is preventing its introduction into flocks in the first place, and continually examining all animals within a flock to ensure they remain CLA-free. As discussed earlier, CLA can remain visceral and animals with this form of lesion development are an important source of infection within a flock; hence, the detection and removal of these individuals will therefore be crucial to the success of any CLA eradication scheme, and must be taken into great account when developing diagnostic tools for this disease.

1.6.1 Conventional identification

The presence of abscesses in superficial lymph nodes is strongly indicative of CLA, particularly if several animals from a single flock present with similar symptoms. Although there are other bacteria that can produce lymphadenopathy, such as *Actinobacillus licheniformis*, *Arcanobacterium pyogenes* and *Staphylococcus aureus* subsp. *anaerobius*, these infections are infrequent and very rarely affect a whole flock (Baird, 2007; Bek-Pederson, 1997). Positive diagnosis of the disease relies on the culture and identification of *Cp* from lesions within the affected animal. However, there are difficulties associated with culturing and identifying *Cp* which can result in a significant delay in CLA diagnosis, allowing time for further transmission within a flock prior to the initiation of any preventative measures being taken. Ruptured superficial lesions commonly become fibrosed and may contain very little pus and viable organisms, hence producing a negative culture as an outcome, such negative cultures have also been observed in humans (Baird, 2007; Peel et al., 1997). Indeed studies of human disease have illustrated a delay in diagnosis; Peel *et al*. (1997) reported cases of human disease where patients presented with a history of lymphadenopathy of many months' duration prior to diagnosis and experienced a protracted course of disease. However, the delay in humans is because *Cp* infection may be an unlikely cause of the observed symptoms compared to other possible causes (such as cancer), and hence the delay of months is for other possible causes to be discounted.

Once cultured, several biochemical tests can then be used to confirm the identification of *Cp*. Catalase, fermentative and oxidative metabolism reactions are vital for the differentiation of coryneforms however initial screening would also include a number of additional tests (Fontaine & Baird, 2008). Esculin hydrolysis, urea production are tested as well as testing the organisms ability to reduce nitrate to nitrites. Carbohydrate tests should result in acid but not gas production from glucose, maltose, sucrose, mannitol and xylose (Dorella *et al.,* 2006). A synergistic haemolytic effect can be observed when *Cp* and *Rhodococcus equi* are cultured together on media containing whole blood (Fraser, 1961); this is known as the CAMP-reaction and is another analysis available to aid identification of *Cp.* The lytic effect of the CAMP-reaction was found to be enhanced by the presence of 10% carbon dioxide (Fraser, 1961). Also a so-called reverse-CAMP reaction can be observed when haemolysis is prevented. When *Cp* is grown in close proximity with some staphylococcal species, staphylococcal β-lysin is inhibited and erythrocytes are protected from haemolysis (Barksdale *et al.*, 1981; Souckova & Soucek, 1972; Zaki, 1976). However as *C. ulcerans* can also produce PLD it will produce similar effects (Barksdale et al., 1981) and this should be taken into account when using the CAMP-reaction or reverse-CAMP reaction for identification purposes.

1.6.2 Analytical Profile Index (API) Identification

In the laboratory, *Cp* can be identified by an Analytical Profile Index identification system known as the 'API Coryne' kit. API kits are standardised kits which identify bacteria on their enzymatic profile and ability to use different carbohydrates. The now widely used API Coryne strip (API bioMérieux, La Balme-les-Grottes, France) became available in the early 1990s (Funke *et al.*, 1997). The API Coryne kit comprises 21 individual tests to be performed in either 24 or 48 h, depending upon whether bacterial growth is adequate or not at 24 h post inoculation. The kit consists of 20 tubes containing dehydrated substrates which allow for eleven enzyme tests and eight carbohydrate tests, a catalase and haemolysis test are also performed (Freney *et al.*, 1991). After inoculation with the test organism and incubation, results of both enzymatic and carbohydrate fermentation tests are observed as specific colour changes generated upon the production of metabolic end-products or a pH change. The negative and positive test results are collectively used to compile a numerical profile which is then compared to a computer software database to calculate the percentage fit of the test isolate to a *Cp* consensus profile (Baird & Fontaine, 2007).

In a study by Freney *et al*. (1991) identification of 240 isolates of coryneform and related bacteria by conventional biochemical methods were compared to those obtained using the API Coryne system. Identifications of the isolates with the API Coryne system showed a 97.6% agreement with conventional methods following a 24 or 48 h incubation. Although 31.8% isolates did require further testing and in three cases, the kit code book and the computer services of the manufacturer failed to identify the organisms (Freney et al., 1991). However, only three bacterial strains (1.2% of the total tested) were misidentified, and the authors of this comparison study concluded the API Coryne system to be rapid and reliable when compared to standard identification techniques.

1.6.3 Serological tests

There has been a focus of research on serological tests that may provide the ability to identify CLA even when the disease is not clinically evident without recourse to bacteriological tests.

The CAMP-inhibition test exploits the reverse-CAMP reaction (discussed in **section 1.6.1**) for the purposes of *Cp* identification and this has led to development of a new method; the haemolysis inhibition test. The haemolysis inhibition test has been used to diagnose disease in several animal species including sheep and goats (Hamid & Zaki, 1973; Burrell, 1980b). Also Shigidi (1978) reported an indirect haemagglutination (IHA) test which, following examination of sera from proven CLA cases, was shown to be more sensitive than the aforementioned haemolysis inhibition test. Additional tests that have been used to study CLA include tube agglutination assays (Cameron *et al.*, 1972), a PCR method (Cetinkaya *et al.*, 2002) and a double immunodiffusion test (Burrell, 1980a). The latter test relies upon obtaining high enough yields of *Cp* exotoxin (PLD) to allow the use of unconcentrated supernatant and serum as reactants.

Despite all of the tests discussed thus far, it is the studies using Enzyme-linked immunosorbent assays (ELISAs) as diagnostic tools that have shown particular promise. ELISAs using both cell wall and toxin antigens have been described with varying sensitivity and specificity. Initially, crude *Cp* cell wall preparations or supernate-derived exotoxin were used as antigens for CLA diagnostic ELISAs. However, despite having a good sensitivity these initial ELISAs usually had a relatively poor specificity (Baird & Fontaine, 2007; Sting *et al.*, 1998; Sutherland *et al.*, 1987). Some ELISAs have claimed sufficient sensitivity and specificity for field diagnosis of CLA including that described by Menzies *et al*. (1994) which employed an *Escherichia coli* recombinant PLD antigen and a test sensitivity of 86.3% and specificity of 82.1% were reported.

Perhaps the most promising CLA diagnostic tool being researched for the detection of *Cp* in sheep and goats is an interferon-gamma (IFN-γ) ELISA. The first test in which an assay of cellular immunity to *Cp* was used as a diagnostic tool was described by Prescott *et al.* (2002). This group developed a whole blood assay for the identification of *Cp* infected sheep based on the detection of an IFN-γ response to whole cell *Cp* antigens. The group employed a commercially available bovine IFN-γ ELISA (BOVIGAM™, CSL Veterinary Ltd., Parkville, Victoria, Australia) which has been used successfully in the eradication of bovine tuberculosis using a test-and-slaughter approach (Rothel *et al.,* 1990; Wood *et al.,* 1992). Cross-reactivity of the monoclonal antibody used in this test with both sheep and goat IFN-γ was demonstrated by Rothel *et al.* (1990). Prescott *et al.* (2002) used the IFN-γ ELISA to detect *Cp* experimentally infected sheep over 450 days with a reliability of 95.7%, and known non-infected sheep with a reliability of 95.5%. Later, Menzies *et al.* (2004) compared the bovine IFN-γ whole blood ELISA to a recombinant PLD antibody ELISA for the diagnosis of *Cp* infection in experimentally infected goats over a one year period. *Corynebacterium pseudotuberculosis* infection was accurately detected using the IFN-γ ELISA with a reliability of 89.2% and detection of non-infected animals with a reliability of 97.1%. Whereas reliability of the PLD ELISA to detect *Cp* was 81.0% for infected goats and 97.0% for those not infected. Regardless of these results, the PLD ELISA did prove to be more predictive of lesion observation at necropsy than the IFN-γ ELISA (Menzies *et al.*, 2004). That said the IFN-γ ELISA appears a reliable approach to CLA detection in goats and sheep at least experimentally (Menzies *et al.*, 2004; Prescott *et al.*, 2002).

Sunil *et al.* (2008) studied the performance of a whole blood IFN-γ enzyme immunoassay (EIA) for the detection of *Cp* in sheep. They found that an IFN-γ EIA response was not triggered by whole cell lysate antigens; however, formalin-inactivated whole cell antigen did evoke an IFN-γ response. The assay was validated in known-negative sheep and experimentally infected sheep as well as a naturally infected flock. The apparently reliable detection of a cell-mediated immune response to *Cp* using the IFN-γ EIA described in this study had a sensitivity of 91% and a specificity of 98% (Sunil *et al.*, 2008).

More recently Rebouças *et al.* (2011) described the use of an IFN-γ ELISA to quantify IFN-γ production in goats and sheep with CLA and also seronegative animals. The specificity of the assay was high at 100% and 93% for goats and sheep respectively; however, poor sensitivity was reported with levels at 55.8% and 56% respectively (Rebouças *et al.*, 2011). Although IFN-γ production appears a potential marker for determination of CLA infection status, more work is necessary to obtain an assay which is both highly sensitive and specific.

The use of a serological diagnostic test in a CLA eradication programme has been pioneered in the Netherlands. This CLA eradication and control scheme is in relation to dairy goats and is based on a PLD ELISA and Western blot analysis (Dercksen *et al.*, 1996). The PLD antibody ELISA employed in this programme is based on a double sandwich ELISA (Ter Laak *et al.*, 1992) which was later refined to improve specificity and sensitivity (Dercksen *et al.*, 2000). The modified version of this test was reported to have a specificity and sensitivity in goats of 98 ± 1% and 94 ± 3% respectively (Dercksen et al., 2000). A similar eradication scheme was proposed for sheep in the Netherlands (Schreuder *et al.*, 1994) however plans failed due to low sensitivity in certain flocks. In fact the overall sensitivity of the test in sheep was much poorer at 79 ± 5% (Dercksen et al., 2000).

This PLD ELISA was later used in a study investigating the control of CLA in sheep using clinical examination alongside ELISA testing (Baird & Malone, 2010). Of the six trial sheep flocks investigated, one was dispersed after two blood tests, and in one recommendations were not followed resulting in the retention of infected animals within the flock. However in the other flocks, the testing and advice given resulted in successful control of CLA to a level where two flocks had only a small number of seropositive animals in the final test and the other two flocks were seronegative in the final two tests (Baird & Malone, 2010).

Controls for CLA in the UK have been put in place and importantly, a UK pre-sale testing regime was launched in autumn 2005 ([Anon], 2005) with the aim of minimizing transmission of the disease between flocks. Also Binns *et al.* (2007) developed an economic ELISA for use in the UK; the test was optimised to detect either total antibody or IgG antibody in serum. One hundred and fifty positive and 103 negative reference sera were used to evaluate both versions of the ELISA at a specificity of 100%. The sensitivity was 71% and 83% to total antibody and IgG respectively; this difference could be explained by the greater affinity of IgG than IgM (the latter would also be detected in the total antibody ELISA). The ELISA test has been used in epidemiological studies (in the UK) and lack of sensitivity in individuals was overcome by using a greater sample size per flock (Binns *et al.*, 2007). Since this study, an ELISA for detection of *Cp*, 'ELITEST CLA', has become commercially available in the UK. The ELITEST CLA is available from Hyphen Biomed (France) and utilises a recombinant form of PLD to detect anti-PLD IgG antibodies in sera from sheep and goats with CLA. The test is currently being employed in Norway as part of a CLA eradication scheme in goats (Dr Michael Fontaine, personal communication). However, due to a sensitivity of 85% and specificity of 95%, the test is not ideal for individual animals and is operated on a flock/group basis.

1.7 Population biology

The study of bacterial populations and identification of bacteria at the strain level (bacterial strain typing) is important for diagnosis, treatment, and epidemiological surveillance of bacterial infections. This section will discuss some of the methods, including multilocus sequence typing, restriction fragment length polymorphism and pulsotyping, used to study population biology with regards to *Cp.*

1.7.1 DNA banding pattern-based methods

1.7.1.1 Restriction fragment length polymorphism

A restriction fragment length polymorphism (RFLP) is a difference in homologous DNA sequences which can be identified by generation of DNA fragments of different lengths following digestion of the homologous DNA samples with specific restriction endonucleases. An RFLP probe (a labelled DNA sequence that hybridises with one of more of the DNA fragments after they are separated by gel electrophoresis) can then be used to produce a unique Southern blot pattern characteristic to a specific genotype at a particular locus. RFLP analysis can be used in genome mapping and genotyping of isolates.

Bjorkroth *et al.* (1999) evaluated rRNA gene RFLP analysis (known as ribotyping) as a tool for identifying *Corynebacterium* and *Turicella* species. The authors analysed 74 strains, representing 26 species or subspecies, by BstEII, SmaI and SphI ribotyping; they found strains of the same species clustered together and that SphI resulted in the most heterogeneous patterns. However, the results from all three enzymes was considered essential for characterising an unknown strain, although the authors did conclude the ribotyping to be a useful tool for screening and characterising potentially novel *Corynebacterium* species. Recently Pavan *et al.* (2012) also used RFLP to successfully identify *Cp*. The group designed a PCR-RFLP based on a hypervariable region of the polymorphic RNA polymerase beta-subunit gene (*rpoB*). *Corynebacterium rpoB* sequences were analysed by computer-assisted restriction analysis, and the authors found that the pattern predicted by using restriction endonucleases MseI and StuI differentiated *Cp* from 61 other *Corynebacterium* species but also *Arcanobacterium pyogenes*, an ovine pathogen with similar clinical manifestations. The method successfully identified 12 wild *Cp* ovine isolates and a single caprine isolate; however, no *Cp* isolates belonging to biovar *equi* were included in the study (Pavan *et al.*, 2012).

More specifically, Abreu *et al.* (2008) used a PCR-RFLP method to compare *Cp* biovar *ovis* isolates obtained from abscesses in sheep and goats diagnosed with CLA in the region of Pernambuco (Brazil). Restriction endonucleases Hpy-Ch4 and MspI, and PstI and MspI were used to fingerprint the *rpoB* and *pld* genes of 35 *Cp* strains respectively (Abreu *et al.*, 2008). However, no difference in fragment banding pattern was observed, regardless of the host species or geographical origin, indicating a homogenous profile of *Cp* in the region.

1.7.1.2 Pulsotyping

Another method which has been described to genotype *Cp* is pulsotyping which differentiates isolates based on pulsed-field gel electrophoresis (PFGE). Connor *et al.* (2000) used PFGE (following restriction digestion with SfiI) to type 50 *Cp* isolates originating from the UK. Six pulsotypes were identified and these clearly differentiated ovine and caprine isolates from the equine isolates, with pulsotype P1 being a unique profile produced by the two equine isolates investigated. Pulsotype P2 comprised 43 of 46 ovine isolates and two caprine isolates, and Pulsotypes P3, P5, and P6 were each represented by a single ovine isolate, whereas P4 a single caprine isolate (Connor *et al.*, 2000). Interestingly, 80% of the strains tested were found to be epidemiologically related to the original goat outbreak strain for the UK.

The same group later applied this technique to 36 ovine and 6 caprine isolates from Australia, Canada, The Netherlands, Eire and Northern Ireland (Connor *et al.*, 2007). The *Cp* strains were found to be homogeneous and only four pulsotypes were differentiated, leading to the conclusion that *Cp* is highly conserved irrespective of country of origin (Connor *et al.,* 2007).

1.7.1.3 Random PCR

Random amplified polymorphic DNA (RAPD)-PCR is a PCR reaction that amplifies random segments of genomic DNA using a single primer of arbitrary sequence (Power, 1996). The technique has been used to effectively type many bacterial species including *Cp* (Dautle *et al.*, 2002; Foley *et al.*, 2004). Foley et al. (2004) characterised 54 *Cp* isolates from 49 horses, 4 cattle, and 1 sheep that were collected during perceived epidemics of infection in Colorado, Kentucky, Utah, and California (USA) using a RAPD-PCR. Ten genotypes were identified, seven of which were isolated from both horses and cattle. The distribution of the genotypes amongst the different states is consistent with a clonally expanding epidemic in Utah, but an increase in infections caused by multiple strains which have not derived from a single source in the other three states (Foley et al., 2004).

Dorneles *et al.* (2012) evaluated the use of the RAPD-PCR Enterobacterial Repetitive Intergenic Consensus (ERIC)-PCR as a tool for molecular typing of ovine *Cp* strains. ERIC-PCR was used to fingerprint 127 *Cp* isolates from sheep lesions were fingerprinted by the RAPD using the primers ERIC-1R, ERIC-2 and primer pair ERIC-1R + ERIC-2. Seventeen different genotypes were obtained for ERIC 1-PCR, whereas ERIC 2-PCR and ERIC 1 + 2-PCR both generated 21 genotypes (Dorneles *et al.*, 2012). For the majority of herds, at most three genotypes were observed amongst isolates from a single property; however, some flocks did have isolates from between four and nine genotypes. The ERIC-PCR described by this group is more discriminative of *Cp* isolates than many of the other DNA-based typing discussed here.

1.7.2 DNA sequencing-based methods

1.7.2.1 16S sequencing

The sequencing of the 16S rRNA gene has commonly been used to identify bacterial isolates and reconstruct bacterial phylogenies, and this method has been used to distinguish *Cp*. Pascual *et al.* (1995) found considerable phylogenetic depth in the *Corynebacterium* genus when they compared 16S rDNA sequences of 23 validated *Corynebacterium* species and seven non-valid *Corynebacterium* species. Furthermore, Cetinkaya *et al.* (2002) successfully used 16S rDNA PCR to identify *Cp* in 93 of 96 DNA samples isolated from abscessed lymph nodes; however, this group did observe cross-reactivity with *C. ulcerans.* The 16S rRNA gene was later used (along with *rpoB* and *pld*) in a multiplex PCR (mPCR) assay to identify *Cp* (Pacheco *et al.,* 2007). The mPCR assay allowed efficient identification of 40 *Cp* isolates previously confirmed to be *Cp* by biochemical testing, whereas related organisms (including *C. ulcerans*) did not produce the same mPCR profile.

1.7.2.2 Multi-locus sequence typing

Multi-locus sequence typing (MLST) is a typing procedure for characterising bacterial isolates based on the sequences of 450-500 bp internal fragments from (usually) seven housekeeping genes. The different sequences for each housekeeping gene are assigned as alleles, and the alleles at each of the seven loci define the sequence type (ST). Within the *Corynebacterium* genus MLST has focussed on *C. diphtheriae*, however, there has been at least one study using MLST of the wider diphtheria group to include *C. ulcerans* and *Cp.* MLST of *C. diphtheriae.* Bolt (2009) designed a MLST scheme for the diphtheria group comprising *C. diphtheriae, C. ulcerans* and *Cp* in which five MLST genes, *atpA, dnaE, fusA, odhA* and *rpoB*, were analysed to resolve the inter-species relationships of this group. Using this scheme Bolt *et al.* (2010) analysed one hundred and fifty *C. diphtheriae* isolates, representing 18 countries and spanning a period of 50 years, by MLST using these five genes and also *dnaK* and *leuA.* The MLST analysis clearly identified clonal complexes associated with disease outbreaks, and revealed two distinct lineages of the *C. diphtheriae* population, one exclusively composed of biotype belfanti isolates whilst the other contained multiple biotypes (Bolt *et al.,* 2010). The *pld* gene in addition to the five MLST genes were used to type 69 *C. ulcerans* isolates, which indicated that animal and human isolates were not distinct genetically (Bolt, 2009). Also, seventy three *Cp* isolates were analysed using the five MLST loci, *pld*, as well as *fagC* and *fagD* from an iron acquisition operon. The MLST distinguished between nitrate negative and positive strains but a shared ancestry was evident with the same alleles identified in two or three of the genes (Bolt, 2009). This MLST study also supports the studies previously discussed that indicate *Cp* biovar *ovis* is a genetically homogenous species.

Later, the same MLST scheme described above was used by Viguetti *et al.* (2012) to investigate the genetic relationship of six invasive *C. diphtheriae* strains from the urban area of Rio de Janeiro (Brazil). The six isolates were all non-toxigenic, and four presented atypical sucrose-fermenting ability. MLST showed five isolates represented new sequence types, and the invasive (sucrose-positive) *C. diphtheriae* isolates did not form a single clonal complex, but rather were found in clonal complexes containing strains responsible for non-invasive disease.

Farfour *et al.* (2012) also used the MLST as described by Bolt *et al.* (2010) and successfully characterised 42 *C. diphtheriae* isolates from Poland, New Caledonia (a French overseas territory) and mainland France. All of the 13 Polish isolates studied were biotype gravis and belong to the same sequence type ST8, a ST that belongs to a clonal complex associated with the former Soviet Union (FSU) epidemic. All of the isolates from New Caledonia were again biotype gravis and four (of five) isolates typed as ST82, and a single isolate as ST39. The 24 isolates from France (21 biotype mitis, two gravis and one belfanti) were distributed amongst eight STs, and the authors concluded that the high diversity resulted from a larger number of isolates from the region (Farfour *et al.,* 2012).

1.7.2.3 Genomic analysis

With the genomic era, the process of determining the complete DNA sequence of an organism's genome (whole genome sequencing; discussed in Chapter Three) has become readily available. As the cost of whole genome sequencing declines, the technique becomes increasingly available to scientists and routine diagnostic laboratories, and this brings with it the opportunity for new and/or improved typing methods to those already mentioned. Indeed the current whole genome sequencing costs have fallen below that of traditional MLST (Larsen *et al.,* 2012), offering scope for new approaches to this kind of typing. Larsen *et al.* (2012) developed a Web-based method for MLST of 66 bacterial species based on whole genome sequencing data (either short sequence reads or preassembled genomes). The publically available MLST Web server finds the best-matching MLST alleles of the specified MLST scheme using a BLAST-based ranking method, and the sequence type is determined by the combination of alleles (Larsen *et al.*, 2012).

There have also been reports of the use of genomic analysis in the facilitation of other typing methods. For example, Yang *et al.* (2013) developed a computational tool, Pan-PCR, which exploits genome sequence data to design discriminatory PCR assays. The authors applied the assay to the Gram negative bacterium *Acinetobacter baumannii* and found that it was capable of distinguishing strains with identical MLST profiles (Yang *et al.*, 2013).

Furthermore, Zankari *et al.* (2013) evaluated the use of whole genome sequencing as a routine tool for surveillance of antimicrobial resistance by comparing it to conventional phenotypic methods. Antimicrobial susceptibility tests were performed on 200 isolates representing four bacterial species. The same isolates were sequenced (using Illumina paired-end technology), their MLST types identified, and the antimicrobial resistance genes identified. A total of 3051 phenotypic tests were conducted and a high agreement between the phenotypic and predicted antimicrobial susceptibility was observed with only seven cases of disagreement reported (Zankari *et al.*, 2013). This study demonstrates the potential of whole genome sequencing as an alternative to current phenotypic methods.

In another study, the gene conservation amongst 73 *Salmonella enterica* genomes was determined and 2,882 core genes were identified, these genes were then evaluated as typing targets (Leekitcharoenphon *et al.*, 2012). Whole genome typing using these targets was compared to the traditional 16S and MLST methods, and a consensus phylogenetic tree based on variation of the core genes was found to be more resolved than that of 16S and MLST. The authors concluded that genomic variation within the core genome could be used to investigate molecular evolution and provide candidate genes for bacterial genome typing (Leekitcharoenphon *et al.*, 2012).

The increasing availability of whole genome sequencing means the approach to typing bacterial isolates and performing population studies is changing to adapt traditional methods by employing genome sequence data. Those studies already performed have indicated the high potential of genome sequencing as a tool in population biology. As the costs of whole genome sequencing decreases and the quality of sequencing increases, it is likely more studies of this nature will exploit whole genome sequences.

1.8 Aims

Prior to commencement of this work, the molecular characterisation of *Cp* has been rudimentary, and besides those discussed above, no other bacterial factor associated with virulence has been identified. However, since the genomic era there has become more scope for such studies in the field of bacteriology. Genomics will allow annotation and functional prediction of all of the predicted proteins encoded by the *Cp* genome. This ultimately will aid the identification of potential virulence factors in addition to those discussed above, and allow a new approach to identifying candidate vaccine and/or diagnostic targets. The genomic characterisation of *Cp* and associated studies will be covered in the following individual results chapters

The initial aim of this work was to obtain the complete genome sequence of the ovine *Cp* isolate 3/99-5, followed by predicting the function of, and manually annotating, all of the predicted coding sequences in the genome. An Australian ovine and North American equine *Cp* isolate (42/02-A and 1/06-A respectively), as well as *C. ulcerans* NCTC 12077, were also sequenced. However, the annotation of these three genomes was automated using web-based software.

The next objective of this thesis was to conduct comparative analyses of these genome sequences and also perform an updated phylogenetic analysis of the *Corynebacterium* genus using these and all other corynebacterial genomes available.

The final aim of this project was to identify protein targets which contain novel antigens either to augment the existing diagnostic ELISA already commercially available in the UK or replace it with a superior test.

Chapter Two: Materials and Methods

2.1 Chemicals

Restriction endonucleases and modifying enzymes were purchased from Promega (Promega UK Ltd.; Hampshire, UK) and all chemicals were purchased from Sigma-Aldrich (Sigma-Aldrich Company Ltd., Dorset, UK) unless otherwise stated.

Media and reagents were sterilised by autoclaving (121°C for 15 min) or by filtration. Large volumes were filtered by drawing through a 0.2 µm Stericup® filter unit (Millipore; Watford, UK) under vacuum and small volumes by passage through a syringe filter with 0.2 µm pore diameter (Millipore).

Oligonucleotide primers were synthesised to-order by Sigma-Genosys Ltd. (Haverhill, Suffolk, UK) or by Eurofins MWG Operon (Ebersberg, Germany), and are presented in **Tables 2.5-2.8**. Lyophilised primers were reconstituted to 100 µM in ddH$_2$O and were subsequently stored at -20°C until required.

2.2 Bacterial strains, plasmids and media

2.2.1 *Escherichia coli*

E. coli strains used for general cloning purposes (**Table 2.1**) were routinely cultured using Luria-Bertani (LB) medium. For maintenance of plasmids **(Table 2.2)**, media were supplemented with the appropriate antibiotics at the following concentrations: ampicillin (100 µg/ml), rifampicin (25 µg/ml), kanamycin and carbenicillin (50 µg/ml).

2.2.2 *Corynebacterium ulcerans*

A single *C. ulcerans* strain was used in this study, isolate NCTC 12077 which was purchased from the National Collection of Type Cultures (NCTC; PHLS, Central Public Health Laboratory, London, United Kingdom). The isolate originates from the UK and was found in a human throat swab taken at the Royal Hampshire County Hospital, Winchester. *Corynebacterium ulcerans* NCTC 12077 is the control strain used for biochemical tests.

Table 2.1: *E. coli* strains for cloning purposes

***E. coli* strain**	**Genotype**	**Source/Reference**
One shot® TOP 10 chemically competent cells	*(F- mcrA Δ(mrr-hsdRMS-mcrBC) φ80lacZ ΔM15 ΔlacX74 recA1 araD139 Δ(ara-leu) 7697 galU galK rpsL (StrR) endA1 nupG)*	Invitrogen, Paisley UK
ElectroMax DH5α-E	*(F- φ80lacZΔM15 Δ(lacZYA-argF) U169 recA1 endA1 hsdR17 (rk-, mk+) gal- phoA supE44 λ- thi-1 gyrA96 relA1)*	Invitrogen
One shot® BL21(DE3)	*F- ompT hsdSB(rB-mB-) gal dcm (DE3)*	Invitrogen
One shot® BL21(DE3)pLysS	*F- ompT hsdSB (rB-mB-) gal dcm (DE3) pLysS (CamR)*	Invitrogen

Table 2.2: Plasmid vectors

Plasmid	**Description**	**Selective antibiotic**	**Source/Reference**
pCR® Blunt II-TOPO®	Routine *E. coli* cloning vector	Kanamycin	Invitrogen
pET-15b	*E. coli* cloning/expression vector	Ampicillin	Novagen
Champion™ pET SUMO	*E. coli* cloning/expression vector	Kanamycin	Invitrogen

2.2.3 *Corynebacterium pseudotuberculosis*

A total of fifty *Cp* strains were utilised throughout this work (**Table 2.3**). The majority of studies were conducted on isolate 3/99-5, a virulent field isolate originating from a CLA outbreak in the Scottish Borders (Connor *et al.*, 2000). *Corynebacterium pseudotuberculosis* strain 3/99-5 is non-nitrate reducing which is typical of ovine *Cp* isolates. The isolate has been studied extensively in-house and has been used in the development of an ovine experimental model of *C. pseudotuberculosis* infection that closely mimics naturally occurring CLA.

The further *Cp* strains used in this study consisted of randomly picked field isolates previously described by Connor *et al.* (2007; 2000) and a further five isolates (two equine and three bovine) obtained from Dr Richard Walker (University of California, USA). Also one final *Cp* strain (NCTC 3450) purchased from the NCTC was included for comparison.

Corynebacterium pseudotuberculosis isolates were sustained on either standard sheep blood agar (E&O Laboratories Ltd; Bonnybridge, UK) or Brain Heart Infusion (BHI) agar (Oxoid; Oxoid Ltd., Basingstoke, Hampshire, UK). The bacterium was cultured in BHI broth (Oxoid) or *Corynebacterium* chemically-defined medium (CCDM; **Appendix one**) supplemented with 0.05% (v/v) polyoxyethylene sorbitane monooleate (Tween 80$^{®}$) from a filter-sterilised 20% (v/v) stock solution (Dubos & Middlebrook, 1948; McGinley *et al.*, 1985a; McGinley *et al.*, 1985b)

Table 2.3: *Corynebacterium pseudotuberculosis* isolates

The *Cp* strains originate from eight countries and include isolates representing both *equi* and *ovis* biovars. The pulsotypes of the isolates, determined by Connor *et al*. (2007; 2000 and unpublished data), also vary.

Isolate	Species of origin	Country of origin	Pulsotype	Source/Reference
3/99-5	Ovine	UK	P3	(Connor et al., 2000)
NCTC 3450	Ovine	South America	P4	(Connor et al., 2007)
40/01-1	Ovine	Northern Ireland	P4	(Connor et al., 2007)
40/01-2	Ovine	Northern Ireland	P4	(Connor et al., 2007)
40/01-3	Ovine	Northern Ireland	P4	(Connor et al., 2007)
40/01-4	Ovine	Northern Ireland	P4	(Connor et al., 2007)
40/01-5	Ovine	Northern Ireland	P4	(Connor et al., 2007)
40/01-6	Ovine	Northern Ireland	P4	(Connor et al., 2007)
40/01-7	Ovine	Northern Ireland	P4	(Connor et al., 2007)
40/01-8	Ovine	Northern Ireland	P4	(Connor et al., 2007)
40/01-9	Ovine	Eire	P4	(Connor et al., 2007)
41/01-1	Ovine	Northern Ireland	P4	(Connor et al., 2007)
41/01-2	Ovine	Northern Ireland	P4	(Connor et al., 2007)
13/02-A	Ovine	Northern Ireland	P2	(Connor et al., 2007)
13/02-B	Ovine	Northern Ireland	P2	(Connor et al., 2007)
28/02-A	Ovine	The Netherlands	P2	(Connor et al., 2007)
28/02-B	Ovine	The Netherlands	P2	(Connor et al., 2007)
28/02-C	Ovine	The Netherlands	P2	(Connor et al., 2007)
37/02-A	Caprine	Canada	P2	(Connor et al., 2007)
37/02-B	Ovine	Canada	P4	(Connor et al., 2007)
37/02-C	Ovine	Canada	P2	(Connor et al., 2007)
37/02-D	Ovine	Canada	P4	(Connor et al., 2007)
37/02-E	Ovine	Canada	P4	(Connor et al., 2007)
38/02-A	Ovine	Canada	P2	(Connor et al., 2007)
38/02-B	Ovine	Canada	P2	(Connor et al., 2007)
38/02-C	Ovine	Canada	P2	(Connor et al., 2007)
38/02-D	Caprine	Canada	P4	(Connor et al., 2007)
38/02-E	Ovine	Canada	P4	(Connor et al., 2007)
38/02-F	Ovine	Canada	P4	(Connor et al., 2007)
38/02-G	Ovine	Canada	P4	(Connor et al., 2007)
42/02-A	Ovine	Australia	P4	(Connor et al., 2007)
42/02-B	Ovine	Australia	P4	(Connor et al., 2007)
42/02-C	Ovine	Australia	P3	(Connor et al., 2007)
42/02-D	Ovine	Australia	P4	(Connor et al., 2007)
42/02-E	Ovine	Australia	P4	(Connor et al., 2007)
42/02-F	Ovine	Australia	P5	(Connor et al., 2007)
42/02-G	Ovine	Australia	P4	(Connor et al., 2007)
42/02-H	Ovine	Australia	P4	(Connor et al., 2007)
42/02-I	Ovine	Australia	P4	(Connor et al., 2007)
42/02-J	Caprine	Australia	P4	(Connor et al., 2007)
42/02-K	Caprine	Australia	P4	(Connor et al., 2007)
42/02-L	Caprine	Australia	P4	(Connor et al., 2007)
42/02-M	Caprine	Australia	P2	(Connor et al., 2007)
48/02	Ovine	The Netherlands	P2	(Connor et al., 2007)
1/06-A	Equine	California, USA	P1	this study
1/06-L	Equine	California, USA	P1	this study
3/99-4	Equine	UK	P1	(Connor et al., 2000)
2/06-C	Bovine	California, USA	P1	this study
2/06-G	Bovine	California, USA	P1	this study
2/06-K	Bovine	California, USA	P1	this study

2.3 Deoxyribonucleic acid extraction and purification

2.3.1 Small scale preparation of plasmid DNA

For rapid screening of plasmid constructs, small amounts of plasmid DNA were extracted from *E. coli* using the QIAprep Spin Miniprep Kit (Qiagen; Qiagen Ltd., West Sussex, UK), which employs a modified version of the classic alkaline lysis method (Birnboim & Doly, 1979). The kit was used according to the manufacturer's instructions. Briefly, a 5 ml culture of transformed *E. coli* was centrifuged at 3,893 × *g*, for 15 min at 4°C. The pelleted cells were resuspended in 250 µl of buffer P1 (containing RNase) and transferred to a microcentrifuge tube. Then, cells were lysed by adding 250 µl of buffer P2 to the cell suspension and gently mixing. Finally, the lysate was neutralised by addition of 350 µl of buffer N3 with thorough mixing. The lysate was then centrifuged for 10 min at 17,970 × *g* to pellet cell debris, following which the cleared lysate was applied to a QIAprep spin column and centrifuged for 1 min at 17,970 × *g*. The column was washed with 0.5 ml buffer PB by centrifugation for 1 min as above. The flow-through was then discarded and centrifugation was repeated to ensure complete removal of any residual wash buffer (which could inhibit downstream applications). The spin column was transferred to a clean microcentrifuge tube and the plasmid DNA was eluted by addition of 50 µl of ddH_2O followed by incubation for 1 min at room temperature and centrifugation at 17,970 × *g* for 1 min.

2.3.2 Extraction of high-purity plasmid DNA

To obtain plasmid DNA of sufficient quality for sequencing, the Qiagen Plasmid Mini Kit was employed. Like the Spin Miniprep Kit, this kit is based on a modification of the alkaline lysis procedure (Birnboim & Doly, 1979) and was also used according to the manufacturer's instructions. Briefly, a 5 ml culture of transformed *E. coli* was centrifuged at 3,893 × *g*, for 15 min at 4°C, following which the bacterial pellet was resuspended in 0.3 ml of buffer P1. To lyse cells, 0.3 ml of buffer P2 was added and the sample was thoroughly (but gently) mixed and incubated at room temperature for 5 min. Subsequently 0.3 ml of chilled buffer P3 was added, and after mixing by inverting the tube, the lysate was incubated for 5 min on ice, and the cell debris pelleted by centrifugation at 17,970 × *g* for 10 min. After equilibrating a Qiagen-tip 20 (by allowing 1 ml of buffer QBT to flow through the column by gravity pull), the sample was applied to the column and allowed to enter the resin. The Qiagen-tip 20 was then washed twice with 2 ml of buffer QC. DNA was eluted from the resin by addition of 0.8 ml of buffer QF (eluted DNA was collected into a clean microcentrifuge tube). The DNA was precipitated by adding 0.56 ml of isopropanol, mixing and centrifuging immediately at 17,970 × *g* for 30 min. The resulting DNA pellet was washed with 1 ml of 70% (v/v) ethanol and centrifuged at 17,970 × *g* for 10 min. Finally, the supernatant was discarded and the pellet was dried at room temperature for 10 min and then dissolved in 50 µl of TE buffer (**Appendix one**).

2.3.3 Extraction and purification of *Corynebacterium* genomic DNA

Extraction and purification of genomic DNA (gDNA) was performed by one of two methods depending on the bacterial strain.

2.3.3.1 Preparation of gDNA using the Wizard® Genomic DNA Purification Kit

Genomic DNA was purified from *Cp* 3/99-5 using the Wizard® Genomic DNA Purification Kit (Promega) according to the manufacturer's supplied protocol. Briefly, bacterial cells were harvested by centrifugation of 1 ml of a stationary-phase culture for 2 min at 17,970 × *g*. The cells were subsequently incubated for 1 h at 37°C in 480 µl of 50 mM EDTA and 120 µl of lysozyme (10 mg/ml in 50 mM EDTA). The suspension was centrifuged for 2 min at 17,970 × *g* and the pellet was resuspended in 600 µl of "Lysis Solution" and incubated for 5 min at 80°C. Three µl of RNase solution was then added to the sample, mixed, and incubated at 37°C for 1 h before allowing it to cool to room temperature. Subsequently, 200 µl of "Protein Precipitation Solution" was added to the sample, which was then vortexed and incubated on ice for 5 min before centrifugation at 17,970 × *g* for 3 min. The DNA was precipitated by addition of 600 µl of isopropanol prior to centrifugation for 2 min at 17,970 × *g*. The resulting DNA pellet was washed with 600 µl of 70% (v/v) ethanol, and centrifuged as above. The ethanol was then decanted, and the pellet was air dried for 10 min at 20°C prior to dissolution in 100 µl of "Rehydration Solution". Concentrations of gDNA (and other nucleic acids detailed elsewhere in this thesis) were determined using a NanoDrop® ND-1000 UV-Vis Spectrophotometer (Thermo Scientific; Wilmington, USA).

2.3.3.2 Preparation of gDNA using the phenol method

Corynebacterium pseudotuberculosis (or *C. ulcerans*) cells were harvested from 5 ml overnight cultures by centrifugation at 3,893 × *g* for 15 min at 4°C. Five milligrams of lysozyme was dissolved in 500 µl of buffer P1 (**Appendix one**) immediately before use, followed by the addition of 2 µl/ml RNase A (10 mg/ml). The resulting solution was subsequently used to resuspend the pellet. The sample was then incubated at 37°C for 1 h before being transferred to a 1.5 ml tube containing *ca*. 500 µl of 10 µm diameter zirconium beads. Homogenisation was performed at a speed of 5.5 for 3 × 20 s in a FastPrep instrument (Q. Biogene; Middlesex, UK). The tube was briefly centrifuged at 17,970 × *g* in a microcentrifuge and the supernatant transferred to a new 1.5 ml microcentrifuge tube. Then, 50 µl of a 10% (v/v) N-lauryl-sarcosine solution and 5 mg Proteinase K were added to the supernatant prior to an overnight incubation at 55°C. Subsequently the gDNA was extracted twice with 500 µl of phenol:chloroform:isoamyl alcohol (25:24:1); the solution was gently mixed by inverting the tube numerous times and centrifuging at 17,970 × *g* for 5 min. The upper aqueous layer was transferred to a new 1.5 ml microcentrifuge tube and a final DNA extraction performed using chloroform:isoamyl alcohol (24:1). The DNA was precipitated by addition of 1 ml of absolute ethanol and 50 µl of 3 M sodium acetate (pH 5.2). Precipitated DNA was harvested by centrifugation at 17,970 × *g* for 20 min at 4°C. The DNA pellet was washed with 500 µl of 70% (v/v) ethanol and the sample centrifuged for 10 min at 17,970 × g, again at 4°C. Finally the ethanol was decanted and the pellet air-dried before being resuspended in 50-100 µl of ddH_2O.

2.3.4 Isolation of genomic DNA for Colony PCR

Two-hundred µl of well mixed Insta-Gene™ matrix (Bio-Rad; Bio-Rad laboratories Ltd., Hertfordshire, UK.) was aliquoted into a 1.5 ml microcentrifuge tube. A single bacterial colony was transferred into the Insta-Gene™ matrix solution and the tube vortexed before being placed in a heat-block at 100°C for 8 min. The solution was again briefly vortexed and then centrifuged at 17,970 × *g* for 3 min. The supernatant (containing gDNA) was transferred to a clean 1.5 ml microcentrifuge tube, and 10-20 µl was used as template per polymerase chain reaction.

2.4 General DNA manipulations

2.4.1 Restriction endonuclease digests

Restriction endonuclease digestions were conducted essentially as detailed in the manufacturer's supplied instructions. Digests were performed in 1.5 ml microcentrifuge tubes, containing 1 × final concentration of the appropriate restriction endonuclease reaction buffer, 0.1 µg/µl of acetylated BSA, an appropriate volume of ddH$_2$O and *ca.* 5-10 U of restriction endonuclease per µg of DNA. Star activity was minimised by conducting digest reactions in sufficient volumes to ensure that the final concentration of glycerol (from the restriction enzyme storage buffer) was $\leq$10% (v/v). Subsequently, restriction digests were incubated at the recommended temperature for the enzyme for between 1-4 h.

2.4.2 Dephosphorylation of DNA

To prevent re-circularisation of vector during ligation reactions the phosphate groups of digested vector were removed using Shrimp alkaline phosphatase (SAP). Dephosphorylation reactions were performed in volumes of 30-50 µl containing 1 × SAP buffer and 1 U SAP per µg DNA. The reactions were incubated at 37°C for 15 min and then heat inactivated at 65°C.

2.4.3 Ligation of DNA fragments

Ligations were performed according to the traditional method employing T4 DNA ligase. Reactions consisted of 1 × Ligase Buffer (containing ATP), a 1: 9 ratio of vector: insert DNA and 2 Weiss Units of T4 DNA ligase and were made up to a final volume of 20 µl with ddH$_2$O. Reactions were incubated overnight in a water bath at 4°C, prior to inactivating the ligase by heating at 65°C for 10 min. Ligated DNA for subsequent bacterial transformations was de-salted by dialysing on Millipore VS membranes (**section 2.6.2**).

2.4.4 Polymerase Chain Reaction

Polymerase chain reactions (PCRs) were conducted using different DNA polymerase enzymes according to the downstream applications for which amplified DNA was to be used. PCR was routinely performed in 50 µl reaction volumes using 0.25 ml flat-topped

PCR tubes (Elkay Lab Products (UK) Ltd; Hampshire, UK) or 96-well ABgene® PCR plates (Thermo Scientific). In addition, template DNA varied in concentration (and hence volume) according to source. Thermal cycling for PCR was conducted using a GeneAmp® PCR system 9700 thermocycler (Applied Biosystems; Foster City California, USA). Primers for differing reactions are shown in **Tables 2.5-2.8**.

2.4.4.1 KOD DNA Polymerase

Amplification of DNA fragments for subsequent cloning was performed using KOD Hot Start DNA Polymerase (Novagen; Merck Chemicals Ltd., Nottingham, UK), an enzyme which generates PCR products without 3'-overhangs (Mizuguchi *et al.*, 1999). Standard reactions contained 1 × Polymerase Buffer, 1.5 mM $MgSO_4$, 0.2 mM each of dATP, dTTP, dCTP and dGTP, 0.3 µM of each sense and anti-sense oligonucleotide primer and 0.02 U/µl of KOD Polymerase. A standard thermal cycling protocol included an initial denaturation step at 95°C for 2 min, followed by 30-40 cycles of a denaturation step of 95°C for 20 s, an annealing step of 10 s (at a temperature dependent on primer melting temperature), and an extension step of 72°C (extension time being dependent on the length of target). On completion of thermal cycling, a final extension step of 72°C for 7 min was also included.

2.4.4.2 Expand Long-Template PCR System

Long-range PCR was conducted using the Expand Long-Template PCR System (Roche Diagnostics Ltd., Burgess Hill, UK), which comprises thermostable *Taq* DNA polymerase and *Tgo* DNA polymerase (a thermostable DNA polymerase with proofreading activity). Typical reactions were conducted in 50 µl volumes and contained 1 × buffer with $MgCl_2$, 500 µM each of dATP, dTTP, dCTP and dGTP, 0.3 µM of each sense and anti-sense oligonucleotide primer, 400 ng purified genomic DNA and 0.075 U/µl of Expand Long Template enzyme mix. Thermal cycling parameters used are shown in **Table 2.4**.

2.4.4.3 Taq master mix

Amplification of DNA fragments for screening purposes was performed using Taq PCR Master Mix (Qiagen). The Master Mix incorporates Taq DNA polymerase, PCR Buffer (with 3 mM $MgCl_2$) and 400 µM of each dNTP. PCR reactions were performed in 40 µl volumes containing 20 µl of the Taq PCR Master Mix, 0.1-0.5 µM of each sense and anti-sense oligonucleotide primers and 10-20 µl of gDNA template prepared using Insta-Gene™ Matrix (**section 2.3.4**). Thermal cycling conditions were as in **section 2.4.4.1**.

Table 2.4: Long-range PCR conditions

Stage	Temperature /°C	Time	No. cycles
Initial denaturation	94	2 min	1
Denaturation	94	10 s	10
Annealing	58	30 s	
Elongation	68	8 min	
Denaturation	94	15 s	25
Annealing	58	30 s	
Elongation	68	8 min +20 s elongation for each successive cycle	
Final elongation	68	7 min	1

2.5 Agarose gel electrophoresis

2.5.1 Analysis of DNA samples

Deoxyribonucleic acid samples, PCR products and restriction endonuclease digests were analysed by agarose gel electrophoresis using the Mini-Sub Cell GT Cell electrophoresis apparatus (Bio-Rad). Agarose gels were prepared containing agarose concentrations ranging from 0.8-2.0% (w/v) (depending on the size of DNA fragment) dissolved in 0.5 × Tris/acetate/EDTA (TAE) buffer (**Appendix one**). The TAE was heated (until the agarose had melted completely) in a microwave oven; the agar solution was then cooled to approximately 56°C before addition of GelRed (Cambridge Bioscience; Cambridge, UK) to a dilution of 1:10,000. The solution was poured into a mini-gel casting tray with comb, and allowed to set. Subsequently the gel was submersed into an electrophoresis tank containing 0.5 × TAE buffer, and loaded with DNA samples prepared in a 1 × final concentration of Blue/Orange loading dye (Promega). Electrophoresis was routinely conducted at 70-100 V/cm until the dye-front had reached the end of the gel, following which DNA was visualised over a UV light using an AlphaImager (GRI AlphaInnotech; UK) gel documentation system. Sizes of linear DNA fragments were determined by comparison with a DNA ladder molecular weight marker.

Table 2.5: Primers for closing the genome

Primer	Sequence (5'-3')	Target contig
Contig6F	TTATCGCCTGCTCGTTTTCCTCTC	Contig 6
Contig6R	GCCCGATTGCGCTGATGGTA	Contig 6
GC01	GTTGCAGAGGTCGAAGAAGG	Contig 7
GC02	TTTACCTCGATCGCCAAAAC	Contig 8
GC03	ATACACCCCTGTTGTGCCTC	Contig 8
GC04	AACAAGTGCTTGATGCCCTC	Contig 3
GC05	TAAGAACCCTGACAGCCCAC	Contig 3
GC06	TCGAGACGTGAATGATCAGC	Contig 4
GC07	TCAACTACTGCACGCGACTC	Contig 4
GC08	CCAATAAGCTCTGCCTTTGC	Contig 5
GC07B	GTTAGATCTGGGGGTCTTTCG	Contig 4
GC08B	GTTAGCGCGAGGATCCTTAAC	Contig 5
GC09	CTGGCCTCTTAGAGTCGGTG	Contig 5
GC10	TTTTACGCGGGGTAGAACAG	Contig 1
GC11	ACGCACAAGAAACAGCACAG	Contig 1
GC12	TTCCTGCTGCTGTCTTTGTG	Contig 2
GC13	AGTGCATTTTGGTCCGGTAG	Contig 2
GC14	AGCCGGATTCGATGTCTA	Contig 7

Table 2.6: Screening primers for amplification of target genes in isolate panel

Primer	**Sequence (5'-3')**	**Target gene**
CPTB0262_ScF	GAGGAAGTTGGCCCGGA	Cp3995_0167
CPTB0262_ScR	CTAGTTATTGCGACGCATCC	Cp3995_0167
CPTB0615_ScF	GCTACCTCCGACGCTGG	Cp3995_0518
CPTB0615_ScR	TTAAGGCTGCAGAACCACTTTTTC	Cp3995_0518
CPTB0666_ScF	GACGAATCTAATGGACCGATCA	Cp3995_0570
CPTB0666_ScR	TCATGCAAAGACTGAAGGGCA	Cp3995_0570
CPTB1273_ScF	GAGGATAAAGAACAAACCCCTAC	Cp3995_1191
CPTB1273_ScR	TTACTTCTTGGAGAAGAGGCGA	Cp3995_1191
Cptb_1605_ScF	GTGTTAAAAAAGTTCGCTGTGC	Cp3995_1510
Cptb_1605_ScR	TTAGCGATTGAAAAAAGGAAGATTC	Cp3995_1510
CPTB1788_ScF	CAAGAAAGCGACCCCATTAC	Cp3995_1712
CPTB1788_ScR	CTACCGGCTGAAAAATGAAGTA	Cp3995_1712
CP40_ScF	ATGCATAATTCTCCTCGATCAGT	Cp3995_1947
CP40_ScR	TTATCTAGAACCAGTTGGCTTTC	Cp3995_1947
Cp40_altF	GCATAACTTCCGCTCTCTTTTTGC	Cp3995_1947
Cp40_altR	CGCTTGGTAAAGGCAAAATCGGT	Cp3995_1947

Table 2.7: Cloning primers for amplification of target genes

Primer	Sequence (5’-3’)*	Target gene
CPTB0262_ClF	AAAAActcgagGAGGAAGTTGGCCCGGA[c]	Cp3995_0167
CPTB0262_ClR	TAATTggatccCTAGTTATTGCGACGCATCC[a]	Cp3995_0167
CPTB0615_ClF	TATTAcatatgGCTACCTCCGACGCTGG[b]	Cp3995_0518
CPTB0615_ClR	AAATTctcgagTTAAGGCTGCAGAACCACTTTTTC[c]	Cp3995_0518
CPTB0666_ClF	AATAAcatatgGACGAATCTAATGGACCGATCA[b]	Cp3995_0570
CPTB0666_ClR	ATATTctcgagTCATGCAAAGACTGAAGGGCA[c]	Cp3995_0570
CPTB1273_ClF	AAAAAcatatgGAGGATAAAGAACAAACCCCTAC[b]	Cp3995_1191
CPTB1273_ClR	TTATTggatccTTACTTCTTGGAGAAGAGGCGA[a]	Cp3995_1191
Cptb_1605_ClF	AATAAcatatgGATGAGGCTGTTGCTGCTA[b]	Cp3995_1510
Cptb_1605_ClR	AAATTctcgagTTAGCGATTGAAAAAAGGAAGATTC[c]	Cp3995_1510
CPTB1788_ClF	AAAATcatatgCAAGAAAGCGACCCCATTAC[b]	Cp3995_1712
CPTB1788_ClR	TTATTggatccCTACCGGCTGAAAAATGAAGTA[a]	Cp3995_1712
CP40_ClF	ATTATcatatgATGCATAATTCTCCTCGATCAGT[b]	Cp3995_1947
CP40_ClR	ATATTctcgagTTATCTAGAACCAGTTGGCTTTC[c]	Cp3995_1947

*oligonucleotide-incorporated restriction endonuclease recognition sites are indicated by lower-case letters and the suffixes a, b and c refer to *BamHI*, *NdeI* and *XhoI* respectively

Table 2.8: Miscellaneous oligonucleotide primers

Primer	**Sequence (5’-3’)**	**Reference**
M13 forward (-21)	TGTAAAACGACGGCCAGT	TOPO Blunt (Invitrogen)
M13 reverse (-29)	CAGGAAACAGCTATGACC	TOPO Blunt (Invitrogen)

2.6 DNA clean up

2.6.1 DNA extraction from agarose gels

Following agarose gel electrophoresis, DNA was extracted using the GENECLEAN® Turbo Kit (Q. Biogene). DNA bands were visualised over UV light and excised using a sterile scalpel blade, before being placed into a clean 1.5 ml microcentrifuge tube. The weight of the gel slice was determined and 100 µl of GENECLEAN® Turbo salt solution was added per 0.1 g of gel slice. The tube was incubated at 55°C with frequent vortexing for approximately 5 min or until the gel had melted. No more than 600 µl of DNA/salt solution was added to a GENECLEAN® Turbo cartridge and catch tube. The cartridge was centrifuged at 17,970 × *g* for *ca.* 5 s in a microcentrifuge, and the flow-through discarded. The DNA was then washed by adding 500 µl of GENECLEAN® Turbo wash buffer and centrifuging at 17,970 × *g* for 5 s. Finally, the catch tube was emptied, and the cartridge was centrifuged for a further 4 min at 17,970 × *g* to ensure complete removal of the wash buffer. The GENECLEAN® Turbo cartridge (containing bound DNA) was placed into a clean 1.5 ml microcentrifuge tube and 30 µl of ddH_2O pipetted directly onto the GLASSMILK®-embedded membrane. The tube was incubated at room temperature for 5 min before a final centrifugation at 17,970 × *g* for 1 min to elute the DNA.

2.6.2 De-salting of ligations prior to transformation

In order to reduce the risk of arcing during bacterial electroporation transformations, excess salts were removed by rapid dialysis. Dialysis of ligation reactions was performed using Millipore Type VS 0.025 µm filter discs (Millipore). Filter discs were placed (shiny-side–up) in Petri dishes containing ddH_2O. After a brief wetting period, ligation reactions were pipetted onto the centre of the discs and allowed to sit for approximately 10 min at room temperature. Ligation reactions were then transferred to clean microcentrifuge tubes and either used immediately or stored at -20°C.

2.7 Bacterial transformation

2.7.1 Transformation of *E. coli*

2.7.1.1 Chemical transformation

A 50-100 µl aliquot of chemically competent One shot® TOP 10 cells, One shot® BL21 (DE3) or One shot® BL21 (DE3)pLysS cells (all Invitrogen, Paisley, UK) were thawed on ice from -80°C. The cells were mixed with 0.5-1 µg plasmid DNA and then incubated on ice for 30 min. The cells were heat shocked at 42°C for 45 s prior to incubating on ice for 5 min. Subsequently 800 µl of SOC (**Appendix one**) was added to the cell suspension, which was then incubated at 37°C × 225 rpm for 1 h. The transformed cells were then plated on LB agar containing the appropriate antibiotics before a final incubation at 37°C until single colonies had developed.

2.7.1.2 Electroporation

An aliquot of electrocompetent *E. coli* ElectroMAX™ DH5α-E™ cells (Invitrogen) was thawed on ice, and 1 µg of plasmid DNA added. The cells and plasmid DNA were gently mixed and then incubated on ice for 10 min. The mixture was transferred into a pre-chilled, 0.1 cm electrode gap electroporation cuvette (Bio-Rad). The cells were electroporated in a Gene Pulser Xcell™ Electroporator (Bio-Rad) at 200 Ω, 25 µF and 2.0 kV, and immediately transferred to 500 µl of SOC medium in a sterile 1.5 ml microcentrifuge tube. Subsequently, the cell suspension was incubated at 37°C × 225 rpm for 1 h and then plated onto LB agar containing antibiotic prior to overnight incubation at 37°C.

2.8 General proteomic techniques

2.8.1 Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE)

Recombinant *Cp* target proteins were visualised by SDS-PAGE using the XCell *SureLock*™ Mini-Cell apparatus (Invitrogen) as per manufacturer's instructions. Briefly the comb was removed from a pre-cast NuPAGE® Novex 4-12% Bis-Tris Mini Gel and the gel was placed into XCell *SureLock*™ Mini-Cell electrophoresis tank. If only one gel was being run in the tank at a time, a buffer dam was used to seal the inner chamber of the tank. The inner chamber was filled with 200 ml of 1 × NuPAGE® MES SDS running buffer containing NuPAGE® Antioxidant, and the outer chamber was filled with 600 ml of ddH$_2$O. Five-10 µl of the protein samples were transferred to clean microcentrifuge tubes, and 2.5 µl of 4 × NuPAGE® LDS Sample buffer and 1 µl of 10 × NuPAGE® Reducing Agent was added to each. Samples were then incubated at 70°C in a hot block for 10 min, centrifuged in a microcentrifuge at 17,970 × *g* for 2 min, and loaded into the wells of the gel. SeeBlue® Plus2 pre-stained standards (Invitrogen) were used as a protein M_r marker and the gel was run at 200 V for 35-50 min.

2.8.2 Visualisation of proteins in SDS-PA gels

Gels were rinsed thrice with *ca*. 100 ml of ddH_2O for 5 min, discarding the ddH_2O after each rinse. The gels were then stained with enough SimplyBlue™ SafeStain (Invitrogen) to cover the gel (~20 ml) for 1 h at room temperature with gentle shaking. Subsequently gels were washed twice for at least 3 h with 100 ml of ddH_2O.

2.8.3 Drying of SDS-PA gels

The DryEase® Mini-Gel Drying System (Invitrogen) was employed to dry mini-gels as per manufacturer's instructions. Briefly, the gel was incubated in ~35 ml of Gel-Dry™ drying solution for 5 min. Two sheets of cellophane were immersed in Gel-Dry™ and one sheet then placed over one half of the drying frame. The mini-gel was then laid on the cellophane and the final sheet of pre-wet cellophane placed on top of the gel. After removing any bubbles and/or wrinkles from the gel/cellophane 'sandwich', the second half of the drying frame was aligned and the drying assembly clamped around the gel which was subsequently air-dried for *ca*. 12 h.

2.8.4 Western blotting

Following separation by SDS-PAGE, proteins were immediately transferred to nitrocellulose membranes using the XCell II™ Blot Module (Invitrogen). Briefly; the top left corner of the membrane was removed to aid orientation post-transfer. The nitrocellulose membrane and two pieces of Whatman™ 3MM filter paper were pre-soaked in Tris-glycine transfer buffer (**Appendix one**), then the membrane and gel were placed in-between the two pieces of filter paper and two sponges. Using the Invitrogen XCell II™ Mini-Cell apparatus, the gel was clamped into the assembly with the gel closest to the cathode core. The blot module was filled with 1 × Tris-glycine transfer buffer and the outer chamber with ddH_2O to serve as a coolant, prior to running at 30 V for 1 h.

Following transfer, the nitrocellulose membrane to which proteins were bound was blocked using 1% (w/v) casein (Thermo Scientific) in 1 × Tris-buffered saline for 1 h. The blocker was then poured off and replaced with 15 ml of 1 × Tris-buffered saline (**Appendix one**), supplemented with 0.1% (v/v) Tween 20® (TBST), containing either anti-His (C-Term)-HRP antibody diluted 1:5,000 or CLA sheep sera diluted 1:500. Following incubation with the antibody/sheep sera for 1 h at room temperature on a rotary shaker, the membrane was washed three times for 10 min in 15 ml of TBST on a rotary shaker. Membranes initially probed with sheep sera were then probed with a horseradish peroxidise-conjugated mouse monoclonal anti-goat/sheep IgG antibody (clone GT-34; Sigma) by incubating in 15 ml of TBST containing a 1:5,000 dilution of the secondary antibody for 1 h, with shaking. Membranes were washed thrice, as before, with TBST and subsequently the HRP-labelled proteins were detected by an ImageQuant LAS 4000 luminescent image analyser (GE Healthcare,) using Pierce® ECL Western Blotting Substrate (Thermo Scientific).

2.8.5 Determination of protein concentration

The concentration of purified recombinant protein was determined using Pierce® BCA Protein Assay (Thermo Scientific) as per manufacturers' instructions. A series of bovine serum albumin standards were prepared (concentrations used were 2.0, 1.5, 1.0, 0.75, 0.5, 0.25, 0.125 and 0.025 mg/ml). Twenty five µl of each standard and unknown sample were pipetted into a microplate well, in triplicate. Then 200 µl of Working Reagent (prepared by mixing 50 parts of BCA Reagent A with one part of BCA Reagent B) was added to each well, and the plate mixed on a plate shaker for 30 s. The plate was covered, incubated at 37°C for 30 min, and then the absorbance measured at 562 nm on an EL_X808_{IU} Ultra Microplate Reader (Bio-Tek Instruments Inc.).

2.8.6 Enzyme-linked immunosorbent assay (ELISA)

An ELISA plate was coated by transferring 100 µl of Coating Buffer (**Appendix one**) containing 1 µg/ml of antigen to each well of the plate. The plate was incubated overnight at 4°C. The antigen solution was then discarded and the plate washed four times with TBST. Complete removal of wash buffer was ensured by banging the plate upside down against a stack of paper towels. Non-specific binding was prevented by adding 200 µl of Blocking Buffer (**Appendix one**) to each well and incubating the plate at room temperature for 2 h. The Blocking Buffer was then discarded and the plate washed as previously. Each test serum was appropriately diluted in Dilution Buffer and 100 µl of each diluted serum added to the desired wells prior to incubation for 1 h at room temperature. The plate contents were discarded and the plate washed with TBST six times. Subsequently, 100 µl of diluted (1:10,000) monoclonal anti-goat/sheep IgG-peroxidase antibody (produced in mouse; Sigma) was added to each well and the plate incubated at room temperature for 1 h. The plate was washed as for the primary antibody. The plate was developed by the addition of 100 µl of HRP Substrate to each well, followed by incubation at room temperature for 6-20 min. The colorimetric reaction was stopped by the addition of 100 µl of 2 M sulphuric acid. Finally the absorbance of each well at 450 nm was determined using an EL_X808_{IU} Ultra Microplate Reader (Bio-Tek Instruments Inc.; Bedfordshire, UK).

2.8.7 Isolation of *C. pseudotuberculosis* secreted proteins

Fifty ml of *Corynebacterium* chemically-defined medium (CCDM) supplemented with 0.05% Tween 80® and 250 µM dipyridyl was inoculated with *Cp* 3/99-5, which was incubated overnight at 37°C × 225 rpm. This overnight culture was then used to inoculate 1 l of CCDM (containing 0.05% Tween 80® and 250 µM dipyridyl) in plastic disposable flasks (Nalgene, Rochester, NY, USA), incubated overnight as previously. The cultures were then centrifuged at 12,000 × *g* for 15 min at 4°C; the resulting supernatant was filtered through a 0.2 µm Stericup® filter unit (Millipore) under vacuum. Exported proteins were precipitated by addition of Trichloroacetic acid to a final concentration of 10% and incubation at 4°C overnight. Proteins were harvested by centrifugation at 10,000 × *g* for 45 min at 4°C, the supernatant discarded and the centrifuge tube drained upside down until the pellet was dry (approx. 30 min to 1 h). The protein pellet was washed three times with Tris-buffered ethanol before being resuspended in high-performance liquid chromatography grade H_2O.

2.8.8 Identification of secreted proteins by LC-ESI MS/MS

The *Cp* 3/99-5 exported protein fraction was separated by SDS-PAGE and the gel lane cut into 20 individual slices. The proteins within each of the gel slices were then digested with trypsin, and analysed by liquid chromatography-electrospray ionisation tandem mass spectrometry (LC-ESI-MS/MS) (Batycka *et al.*, 2006). The mass spectrometry data was submitted to an in-house Mascot server and searched against the manually annotated *Cp* 3/99-5 genome sequence using the Mascot search algorithm (Perkins *et al.*, 1999).

Analysis of the Mascot data was performed in ProteinScape™ version 2.1 (a platform developed by Bruker Daltonics). Molecular weight search (MOWSE) scores were only considered significant for individual protein identifications if two or more non-redundant peptides were matched for each protein, and each peptide contained a consecutive "*b*" or "*y*" ion series of at least four amino acid residues.

2.9 Expression and purification of recombinant proteins

2.9.1 Expression of recombinant proteins

Forty ml of LB broth supplemented with 10 mM glucose and 50 µg/ml carbenicillin was inoculated with transformed *E. coli* BL21(DE3), and incubated overnight at 26°C, 30°C or 37°C × 225 rpm. This overnight culture was then used to inoculate 1 l of LB broth (containing 10 mM glucose and 50 µg/ml carbenicillin) in a 5 l conical flask. The culture was incubated as above until it reached an OD_{600nm} of *ca*. 0.6. Expression of the cloned gene was induced by addition of either 1.0 mM or 0.5 mM Isopropyl β-D-1-thiogalactopyranoside and the culture incubated again as previously for 1 h. Cells were harvested by centrifugation at 12,000 × *g* for 15 min at 4°C and the supernatant was discarded.

2.9.2 Protein purification using native conditions

The cell pellet from a culture with induced protein expression (previous section) was resuspended in 20 ml of Lysis Buffer (**Appendix one**). Benzonase enzyme (Novagen) was added to 25 U/ml, and the cell suspension incubated at 37°C × 100 rpm for 20 min to allow cell lysis and degradation of nucleic acids. Subsequently cell debris was pelleted by centrifugation at 22,000 × *g* for 30 min at 4°C, and the supernatant decanted into a clean tube. Two 12 ml Econo-Columns (Bio-Rad) were loaded with 4 ml of Ni-CAM HC resin (Sigma) and allowed to drain by gravity flow to give a bed volume of 2 ml in each column. Each column was then washed with 10 ml of Binding Buffer (**Appendix one**). The end of the columns were sealed prior to loading them with 10 ml of cleared cell lysate, sealing the column tops and incubating overnight at 4°C on a tube rotator. The tops and bottoms were unsealed and the columns allowed to drain. The columns were then washed with 8 × 5 ml of Wash Buffer (**Appendix one**) before eluting column-bound proteins into clean collection tubes with 5 × 2 ml each of Elution Buffer (**Appendix one**). The eluted proteins were dialysed against 20 mM Na_2HPO_4 (pH 7.0) in a volume 100 × that of the dialysate. The dialysis was performed overnight at 4°C on a magnetic stirrer to allow rotation of the dialysis tubing and buffer. Once dialysis was complete, the dialysed protein was centrifuged at 3,080 × *g* for 30 min at 4°C. The cleared supernatant (containing purified protein) was decanted into a clean tube and the non-proteinaceous precipitate was discarded.

2.9.3 Protein purification using denaturing conditions

This was conducted as in section **2.9.2** however Guanidinium Lysis buffer and denaturing binding, wash and elution buffers were used (see **Appendix one** for buffer recipes).

2.10 DNA Sequencing

2.10.1 Genome sequencing

Automated sequencing of genomic sequences was carried out by either 454 Life Sciences (Branford, USA) or The GenePool (University of Edinburgh, UK) using a *de novo* whole genome shotgun sequencing technique. Subsequently sequence reads were assembled into contiguous sequences (contigs) using the Roche Newbler or Velvet (Zerbino & Birney, 2008) assemblers respectively and were provided as a multiple FASTA file. When used, hybrid assembly was performed by the GenePool using the Minimus2 assembler (Sommer *et al.*, 2007).

2.10.2 Sequencing of PCR products

Contract sequencing of PCR products was conducted by Eurofins MWG Operon (Ebersberg, Germany) using their Value Read Tube service. Various primers were used to sequence PCR products, however, PCR products cloned prior to sequencing were sequenced using universal primers shown in **Table 2.8**. Sequence data were provided as DNASTAR DNA files.

2.11 Basic DNA sequence manipulations and analyses

Basic DNA sequence analyses, including translation, mapping of restriction endonuclease recognition sites, and multiple DNA sequence alignments were carried out using the Clone manager Professional Suite Version 9 (Sci–Ed Scientific Ltd., Cary, NC, USA) or DNASTAR$^{®}$ Lasergene version 8 (DNAStar Inc., Madison, WI).

Throughout this project whole genome sequences were viewed and edited using Artemis software (Rutherford *et al.*, 2000), which allows visualisation of sequence features in the context of the sequence, and its six-frame translation.

2.12 Gene prediction and annotation

2.12.1 Open-reading frame (ORF) prediction

The *Cp* 3/99-5 genome sequence was searched using automated software Prodigal v.1.10 (Prokaryotic Dynamic Programming Genefinding Algorithm, University of Tennessee) (Hyatt *et al.*, 2010), a bacterial and archeal gene finding program, to predict open reading frames (ORFs) of genes. Prodigal was run with “-c” and “-m” flags. This commanded the program to avoid predicting genes running off the edges of the sequence and to treat runs of “n”s as masked sequence preventing genes being built across them. A second ORF prediction system, Glimmer v.3.02 (Delcher *et al.*, 2007), was also used to detect potential coding sequences in the genome sequence. This system creates a variable-length Markov model from a training set of genes and uses it to attempt to identify all genes in a given DNA sequence. Glimmer was run several times with varying minimum gene length and maximum overlap length settings as shown in **Table 2.9.**

Table 2.9: Settings used for running Glimmer3

Prediction	Min gene length* /nucleotides	Max. gene overlap /nucleotides
A	**75**	**25**
B	**100**	**50**
C	**100**	**25**

*Minimum gene length does not include the bases in the stop codon.

Glimmer3 was run three times with different settings corresponding to the minimum gene length accepted and the maximum overlap length between genes allowed.

2.12.2 ORF correction

A Smith Waterman alignment (Pearson, 1991; Smith & Waterman, 1981) was conducted for all predicted ORFs in the *Cp* 3/99-5 genome using default settings and the published *C. diphtheriae* genome (NCBI accession number NC_002935) as a reference sequence. ORF predictions were manually corrected in Artemis.

2.12.3 Analysis of subcellular location

The location of proteins within the bacterial cell was inferred by computational analysis of predicted ORFs. Five free, web-based software programs were employed to predict the subcellular location of the proteins: SignalP 3.0 server predicts the presence and location of signal peptide cleavage sites (Bendtsen *et al.*, 2004; Nielsen *et al.*, 1997); PSORT-B v. 2.0 (Gardy *et al.*, 2005) assigns a probable localisation site to a protein from an amino acid sequence alone; the SubLoc software (Chen *et al.*, 2006) also bases predictions on amino acid composition alone, but does not predict membrane proteins; CELLO v.2.5 uses a two-level Support Vector machine system to assign localisations to both prokaryotic and eukaryotic proteins (Yu *et al.*, 2006); and finally the program GPos-PLoc (Chou & Shen, 2008; Shen & Chou, 2007) which is exclusively for prediction of Gram-positive bacterial proteins and divides them into five subcellular compartments.

Discrepancies between the different programs used to predict subcellular location were generated from the predicted ORF sequences, so an overall average location was made. If the discrepancies were high, the GPos-PLoc location was taken.

2.12.4 Functional analysis

The function of predicted proteins was inferred using several web-based software programs described below.

Translated *Cp* ORFs were identified by comparison with sequences resident in the National Center for Biotechnology Information (NCBI) non-redundant, SWISS-PROT and Trembl databases, using the Basic Local Alignment Search Tool (BLAST).

InterProScan was used to predict the function of proteins encoded by the (predicted) ORFs of the genome. InterProScan (Zdobnov & Apweiler, 2001) is a computational tool that scans a given protein sequence against the protein signatures of the members of the InterPro database. The program uses a collaboration of various databases to give a unique, non-redundant characterisation of a given protein family, domain or functional site (Zdobnov & Apweiler, 2001). The program was installed locally and the genome sequence was submitted in FASTA format.

Ribosomal RNA genes in the *Cp* 3/99-5 sequence were identified using RNAmmer (Lagesen *et al.*, 2007). In addition tRNA scan-SE v.1.23 (Lowe & Eddy, 1997) was used to detect the presence of transfer RNAs in the genome sequence.

2.12.5 Gene annotation

Manual gene annotation of the *Cp* 3/99-5 genome was conducted by employing information from the functional analyses mentioned previously. Annotations were manually added to a compiled genbank file in Artemis.

2.12.6 Automated gene prediction and annotation

For *Cp* strains 1/06-A and 42/02-A and *C. ulcerans* NCTC 12077, ORF prediction was automated using x-BASE software (Altschul *et al.*, 1997; Chaudhuri *et al.*, 2008; Delcher *et al.*, 2007; Kurtz *et al.*, 2004; Lowe & Eddy, 1997). This software runs a number of prediction programmes to both assign ORFs to the sequence but also to allocate annotations to coding sequences (CDS). Gene prediction is performed using Glimmer run with a minimum gene length of 90 bp and a maximum gene overlap of 50 bp. The xBASE software searches for tRNA genes with tRNAScan-SE and ribosomal RNA genes with

RNAmmer. Also protein BLAST is run using the translated coding sequences as a query against a reference sequence selected (for this work this was always *C. diphtheriae*, NC_002935). The best result for each BLAST search is then imported as the gene annotation (if under the selected BLAST E-value cutoff of 1e^{-10}). The final output from xBASE is a GenBank file suitable for loading into a sequence viewer such as Artemis.

2.12.7 Analysis of codon usage

The codon usage of all predicted *Cp* 3/99-5 genes was determined using CodonW v.5a software (Peden, 1999). The software was run using the input option of bacterial genetic code (rather than universal). CodonW outputs the effective number of codons (ENc) and the G+C content at the third synonymous codon position (GC3s).

2.13 Genome comparison

2.13.1 ACT comparison

Comparisons of the genomes sequenced in this project were achieved using the Artemis Comparison Tool (ACT) Release 9 (Carver *et al.*, 2005) (Sanger Pathogen Sequencing Unit). Some publically available genome sequences were also used for comparison and these were all downloaded from the National Center for Biotechnology Information online database. In order to use ACT, a comparison file was generated using locally installed BLAST on a Linux operating system. A BLASTall search was performed using the "–p tblastx" option which searches a translated nucleotide database using a translated nucleotide query.

2.13.2 Genomic Island prediction

Genomic islands were computationally identified and visualised with IslandViewer (Langille & Brinkman, 2009). This web accessible application uses IslandPick (Waack *et al.*, 2006), IslandPath-DIMOB (Langille *et al.*, 2008) and SIGI-HMM (Langille *et al.*, 2008) methods for genomic island prediction. Genome data was input as a Genbank file.

2.13.3 Genome comparison using BLAST Ring Image Generator

Whole genomes were compared and visualised using the BLAST Ring Image Generator (BRIG; Alikhan *et al.*, 2011). BRIG generates a circular comparison image based on a BLAST comparison; similarity between a reference genome in the centre against other query sequences is displayed as a set of concentric rings coloured according to BLAST identity. Genome sequences were input as nucleotide FASTA sequence files and the program's default BLAST settings were used.

2.13.4 Genome comparison and alignment using BLASTx

An in-house script was generated by Raja Yaga, which aligned all genes from a given genome sequence against genes from all other genomes included in the study using local BLASTx. Output was in the form of a matrix file that indicated the most homologous match for each gene and whether this was a reciprocal-best hit.

2.14 Phylogenetic analyses

2.14.1 Genomes

Genome sequences employed in the current phylogenetic studies of the *Corynebacterium* genus consisted of 17 *Corynebacterial* genomes. Studies of housekeeping loci also used the genomes of 16 *Mycobacterium* species, *Rhodococcus jostii* RHA, and *Nocardia farcinica* IFM 10152 which served as an outgroup for analysis. NCBI genome accession numbers and corresponding species and strain information for all genomes is shown in **Table 2.10**.

2.14.2 Loci

A total of 65 gene loci were used in the phylogenetic studies of the *Corynebacterium* genus. These consisted of 33 genes encoding proteins concerned in general housekeeping roles of the organism, and a further 32 genes that encode secreted proteins whose function often remains as yet uncharacterized. Thirty-one of the 33 housekeeping loci were as described by Wu and Eisen (2008) and encode phylogenetic marker genes (*dnaG*, *frr*, *infC*, *nusA*, *pgk*, *pyrG*, *rplA*, *rplB*, *rplC*, *rplD*, *rplE*, *rplF*, *rplK*, *rplL*, *rplM*, *rplN*, *rplP*, *rplS*, *rplT*, *rpmA*, *rpoB*, *rpsB*, *rpsC*, *rpsE*, *rpsI*, *rpsJ*, *rpsK*, *rpsM*, *rpsS*, *smpB*, and *tsf*). The remaining two loci included were *hsp65* (the product of which is a heat shock protein) and *cpn60* encoding a chaperonin. The second group of 32 loci were proteins identified by the LC-ESI MS/MS analysis of a supernatant fraction of a *Cp* 3/99-5 preparation (**section 2.8.7**) that were also previously predicted to be extracellular by bioinformatic analyses (**section 2.12.3**).

2.14.3 Analysis of sequence alignment

Analyses of multiple sequence alignments were conducted in TOPALi v. 2.5 (Milne *et al.*, 2009) using either a maximum likelihood or Bayesian approach. Appropriate model selections were conducted for each analysis using the menu in the TOPALi software; the program selected the optimal model based on calculations involving hierarchical likelihood ratio tests or Bayesian information criterion (Milne *et al.,* 2009).

In preliminary studies the maximum likelihood algorithm PhyML-aLRT (v.2.4.5) (Anisimova & Gascuel, 2006) was employed. PhyML was performed using both protein and nucleotide alignments with 100 bootstrap runs.

Bayesian inference was produced using the program MrBayes (v.3.1.1) (Huelsenbeck & Ronquist, 2001). Preliminary studies were run using protein alignments and nucleotide alignments, assuming the sequence to be both simple DNA or protein-coding DNA. In these cases MrBayes analyses were run twice over 625,000 generations with a 20% burn-in and a sample frequency of 10. For larger studies containing multiple loci, a nucleotide alignment was used and analyses assumed the sequence to be protein-coding and were run twice over 1,250,000 generations with a 25% burn-in and a sample frequency of 10.

2.14.4 Basic phylogenetic tree analysis

To better visualise the resulting tree, the New Hampshire Tree file from TOPALi was loaded into Dendroscope version 3 (Huson & Scornavacca, 2011), a tool for visualizing phylogenetic trees and rooted networks, and the node labels edited appropriately.

Table 2.10: Genome sequences included in phylogenetic analyses

Accession #	**Species and strain**	**Genome size/ bp**
NC_012590	*C. aurimucosum* ATCC 700975	2,790,189
NC_002935	*C. diphtheriae* NCTC 13129	2,488,635
NC_004369	*C. efficiens* YS-314	3,147,090
NC_006958	*C. glutamicum* ATCC 13032	3,282,708
NC_003450	*C. glutamicum* ATCC 13032	3,309,401
NC_009342	*C. glutamicum* R	3,314,179
NC_007164	*C. jeikeium* K411	2,462,499
NC_012704	*C. kroppenstedtii* DSM 44385	2,446,804
NC014329	*C. pseudotuberculosis* FRC41	2,337,913
CP001809	*C. pseudotuberculosis* 1002	2,335,113
CP002251	*C. pseudotuberculosis* I19	2,337,730
CP001829	*C. pseudotuberculosis* C231	2,328,208
NC_016781*	*C. pseudotuberculosis* 3/99-5	2,337,938
CP003062*	*C. pseudotuberculosis* 42/02-A	2,337,606
CP003082*	*C. pseudotuberculosis* 1/06-A	2,279,118
(BioProject PRJNA179437)**	*C. ulcerans* NCTC 12077	2,616,855
NC_010545	*C. urealyticum* DSM 7109	3,369,219
NC_008595	*Mycobacterium avium* 104	5,475,491
NC_002944	*M. avium subsp. paratuberculosis* K-10	4,829,781
NC_002945	*M. bovis* AF2122/97	4,345,492
NC_008769	*M. bovis* BCG str. Pasteur 1173P2	4,374,522
NC_009338	*M. gilvum* PYR-GCK	5,619,607
NC_002677	*M. leprae* TN	3,268,203
NC_008596	*M. smegmatis* str. MC2 155	6,988,209
NC_009077	*M. sp.* JLS	6,048,425
NC_008705	*M. sp.* KMS	5,737,227
NC_008146	*M. sp*. MCS	5,705,448
NC_009565	*M. tuberculosis* F11	4,424,435
NC_000962	*M. tuberculosis* H37Rv	4,411,532
NC_002755	*M. tuberculosis* CDC1551	4,403,837
NC_009525	*M. tuberculosis* H37Ra	4,419,977
NC_008611	*M. ulcerans* Agy99	5,631,606
NC_008726	*M. vanbaalenii* PYR-1	6,491,865
NC_006361	*Nocardia farcinica* IFM 10152	6,021,225
NC_008268	*Rhodococcus jostii* RHA1	7,804,765

*Incomplete versions of these genomes were used for analyses as described in Chapters Three and Four.

**Genome sequence remains incomplete as a draft sequence.

Chapter Three: Characterisation of the *Corynebacterium pseudotuberculosis* 3/99-5 genome using basic bioinformatic analyses

3.1 Introduction

DNA sequencing uses several techniques or methods to establish the order of nucleotide bases in a given molecule of DNA. DNA sequences were first obtained in the early 1970s using labour intensive methods such as “wandering-spot analysis” described by Gilbert and Maxam (1973). However, it was the “plus and minus” method that was used to sequence the first complete DNA genome, that of bacteriophage ΦX174 which was reported by Sanger *et al.* (1977). A variation of the plus and minus method, the chain termination method (or simply Sanger method), then rapidly took over as method of choice due to its accuracy and speed. This method employs dideoxynucleotide triphosphates (ddNTPs) as chain-terminating molecules (Sanger *et al.*, 1977).

Currently most sequencing operations are performed using next-generation sequencing platforms, of which there are three that dominate the commercial market: the Roche 454 Genome Sequencer, the Illumina (Solexa) Genome Analyzer and the Life Technologies SOLiD System. The Roche system, employing pyrosequencing, produces the largest read length at 400 bases however the technology has high error rates in homopolymer repeats (Rothberg & Leamon, 2008). The Solexa platform uses a sequencing by synthesis principal resulting in considerably shorter reads of up to 100 bases. This technology is the most widely used platform and also has the capabilities to produce paired-end reads (Metzker, 2010). Life Technologies SOLiD is based on sequencing by ligation and only generates reads of 50 bases with longer run times than other platforms, however the technology used does produce a low error rate (Zhou *et al.*, 2010).

The production of low cost reads by NGS technologies makes them useful in a number of applications. These include *de novo* sequencing (the generation of primary genomic sequence of an organism) and variant discovery by either whole genome or targeted resequencing.

Transcriptome sequencing and analysis has also increased due to the next generation platforms available, and the deep sequencing approach RNA-Seq (Wang *et al.*, 2009) has enhanced understanding of the complexity of eukaryotic transcriptomes. RNA-Seq has allowed amongst others the accurate monitoring of gene expression during yeast vegetative growth (Nagalakshmi *et al.*, 2008), yeast meiosis (Wilhelm *et al.*, 2008) and mouse embryonic stem cell differentiation (Cloonan *et al.*, 2008). Additional applications of NGS include metagenomics, whereby DNA is extracted and analysed from uncultured microbial communities and epigenomic analysis, the study of heritable gene regulation using DNA sequence modifications and higher order structures (Zhou *et al.*, 2010).

Possessing genomic sequences of pathogens can allow the cataloguing of species diversity and improve taxonomy but may also provide greater understanding of pathogenicity and ways to control disease. Knowing the genome sequence of a pathogen does not necessarily mean information regarding gene function is held. However, it does allow for inference of gene function and from this, potential antimicrobial drug targets or vaccine candidates can be identified. For example, Wizemann *et al.* (2001) exploited the whole genome sequence of *Streptococcus pneumoniae*, and identified 130 CDSs encoding proteins with secretion motifs or similarity to predicted virulence factors. Mice were immunized with 108 of these proteins, and six were discovered to be protective against disseminated *S. pneumoniae* infection (Wizemann *et al.*, 2001).

Whole genome sequencing is the process of determining the complete DNA sequence of an organism's genome, in the case of bacteria this entails sequencing chromosomal DNA as well as any plasmids the organism may possess. The first bacterial genome to be completed was *Haemophilus influenzae* (Fleischmann *et al.*, 1995) and this marked the beginning of a revolution in bacterial genomics. According to the NCBI genome database, http://www.ncbi.nlm.nih.gov/, at the time of writing (December 2011) there are 1,843 bacterial genome projects complete and a further 5,230 ongoing. Of the complete genomes, 177 belong to the phylum Actinobacteria and more specifically 21 to the *Corynebacterium* genus. More specifically at the time that this work commenced (December 2008) there was no complete or partial *Corynebacterium pseudotuberculosis* genome sequence in the public domain. However, in the interim eight complete and five incomplete projects have been deposited in the NCBI genome database.

Here the objectives were to (i) sequence the genome of *Corynebacterium pseudotuberculosis* strain 3/99-5 (Genbank accession number CP003152, Pethick *et al.*, 2012b) using *de novo* sequencing, (ii) identify open-reading frames (ORFs) of genes and (iii) conduct bioinformatic analyses to assign an inferred cellular location and function to predicted protein coding sequences. The presence of structural RNA genes was also assessed.

3.2 Results

3.2.1 Genome sequencing of *Cp* 3/99-5

Roche 454 whole genome shotgun sequencing of *Cp* 3/99-5 generated a total of 367,232 reads with an average length of 248 base pairs (bp); this yield corresponded to 40 times coverage. Sequence reads were then assembled by 454 Life Sciences using the Newbler assembler into eight contigs with a total length of 2,319,057 bp. The largest contig, Contig 2, comprised ~722 kb, while the shortest, Contig 6, was ~6 kb in length, see **Table 3.1**.

An Artemis Comparison Tool (ACT; Sanger) comparison using the closest related, published genome *Corynebacterium diphtheriae* (accession NC_002935) showed that the *Cp* 3/99-5 contigs were co-linear with the *C. diphtheriae* reference sequence. The comparison provided a provisional framework for determining the order and orientation of *Cp* 3/99-5 contigs in the subsequent assembly and for design of primers for gap closing (**Fig. 3.1**). Interestingly it was determined that Contig 6 contained homologues of ribosomal ribonucleic acid (rRNA)-encoding genes from the *C. diphtheriae* genome. Within the *C. diphtheriae* genome there are five loci that contain such a region encoding 5S, 16S and 23S genes, known as the *rrn* operon. Analysis revealed that gaps between *Cp* contigs 7 and 8, 5 and 1, 1 and 2, and 2 and 7 corresponded to *rrn* operons in the *C. diphtheriae* genome, giving rise to the presence of four occurrences of Contig 6 in a predicted map of contig orientation (**Figs. 3.1** and **3.2**) The remaining *rrn* operon of *C. diphtheriae* mapped to within Contig 1 of the current genome.

Table 3.1: Contig statistics post assembly of read data

Assembly was performed by 454 Life Sciences using their Newbler assembler.

Contig	Size/ bp	G+C content/ %
1	367,241	63.45
2	721,944	52.98
3	244,724	50.09
4	150,189	50.77
5	276,346	53.07
6	5,912	52.26
7	405,086	51.83
8	147,615	49.23

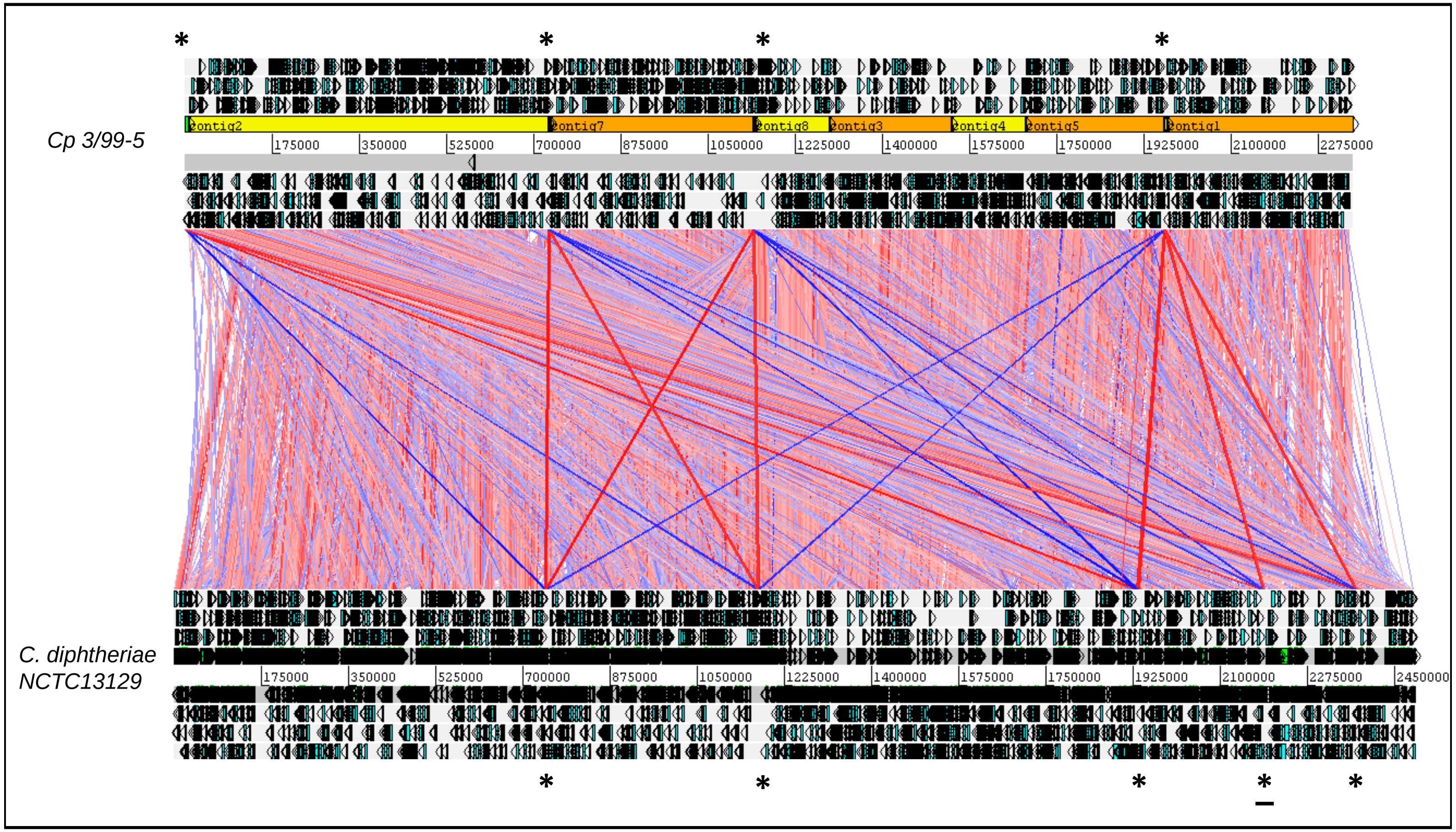


Figure 3.1: Map of *Cp* contigs and homology with *C. diphtheriae*

Following the sequence of the *Cp* 3/99-5 genome, the *Cp* contigs (top) were reordered based on homology with *C. diphtheriae* (bottom) and an ACT comparison of the two genomes is shown here. The *Cp* 3/99-5 genome was predicted to have four copies of the *rrn* operon indicated by the asterisks (containing 5S, 16S and 23S rRNA genes); hence Contig 6 (thought to be an amalgamation of reads from the four regions) is shown in four loci in the genome. *C. diphtheriae* contains five copies of the *rrn* operons, also indicated by asterisks, the fifth *C. diphtheriae rrn* operon mapped to within Contig 1 of the *Cp* genome.

3.2.2 Closing the *Cp* 3/99-5 genome

The tentative order and orientation of *Cp* 3/99-5 contigs was confirmed where possible by PCR bridging. Primer pairs were designed to anneal approximately 150 bp from the 3'-end of each contig to allow for sequence quality degeneration towards the ends of each contig (primer locations can be seen in **Fig. 3.2**).

PCR yielded a complex pattern of products for each reaction and the predominant product was analysed. In the case of amplification of the gaps between contigs 8 and 3, 3 and 4, and 4 and 5, PCR products were excised from an agarose gel and purified then cloned and sequenced. This approach successfully closed gaps between contigs 8, 3 and 4, where the overlaps between the PCR bridging products and adjacent contigs were 181, 312, 333 and 404 bp respectively. The PCR bridging attempt between contigs 4 and 5 produced sequence different to that targeted and the gap could not be closed.

Long-range PCR was used to analyse gaps between contigs 7 and 8, 5 and 1, 1 and 2, and 2 and 7, which potentially contained a Contig 6 homologue and added approximately 6 kb to the PCR product size. Each of the long range reactions produced a product larger than 6 kb, shown in **Fig. 3.3** (circled in red). Two unsuccessful attempts were made to sequence one of the products (GC01+GC02) and identify a Contig 6 homologue. Unfortunately the excessive sequencing cost quoted prohibited committing to sequencing all four products.

An alternative strategy of restriction analysis was adopted to determine the composition of the long-range PCR products. Digestion with SmaI, HaeII and KpnI was predicted to generate fragment sizes of 378 bp, ≥683 bp, 1,104 bp, 1,689 bp and ≥2,058 bp for SmaI, 52, ≥66, ≥845, 1431 and 3518 for HaeII and ≥688, 1142, ≥1158, and 2924 for KpnI. The observed banding pattern on agarose gels are shown in **Fig 3.4**.

This confirmed that the four larger genome gaps each contained an rRNA-encoding region, equivalent to that of Contig 6. Since this sequence is not of primary interest in this work it was decided to simply determine the sequences linking the Contig 6-like sequence and the adjacent contigs. PCR primers were therefore designed to extend out of either end of Contig 6 in order to bridge from either end of the rRNA region to the adjacent contig (**Fig. 3.5)**.

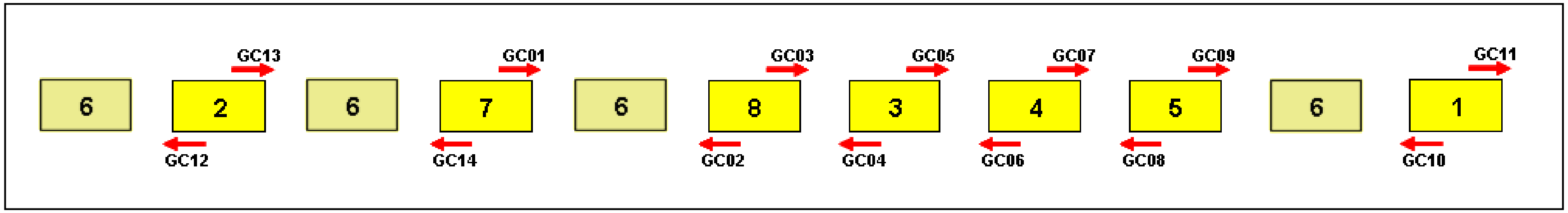


Figure 3.2: Map of *Cp* contigs and primers

The locations of primers designed for attempted genome closure in relation to the contigs are shown (red arrows). Contig 6 is shown in four copies as this represents the four copies of the *rrn* operon predicted to be present in the *Cp* genome.

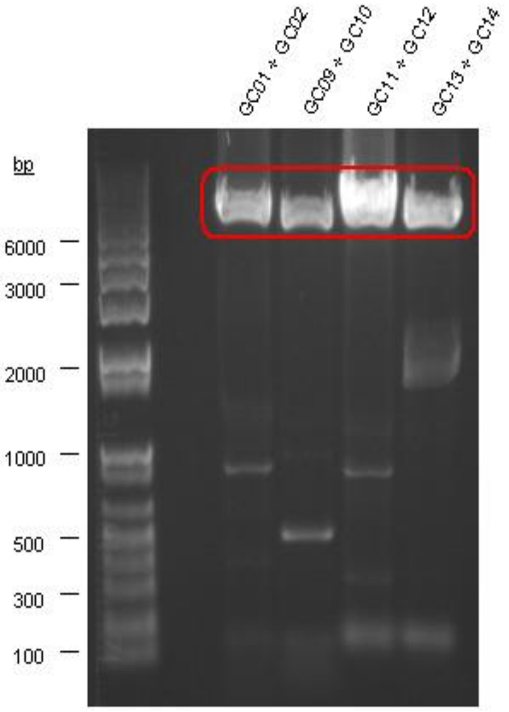


Figure 3.3: Long-range PCR bridging

Regions of unknown sequence expected to contain a copy of the *rrn* operon were amplified by long range PCR and are all greater than 6 kb in size. Primer pairs GC01+GC02, GC09+GC10, GC11+GC12 and GC13+GC14 correspond to primer locations as shown in **Fig. 3.2** and indicate an *rrn* operon to be present between contigs 7 and 8, 5 and 1, 1 and 2, and 2 and 7 respectively.

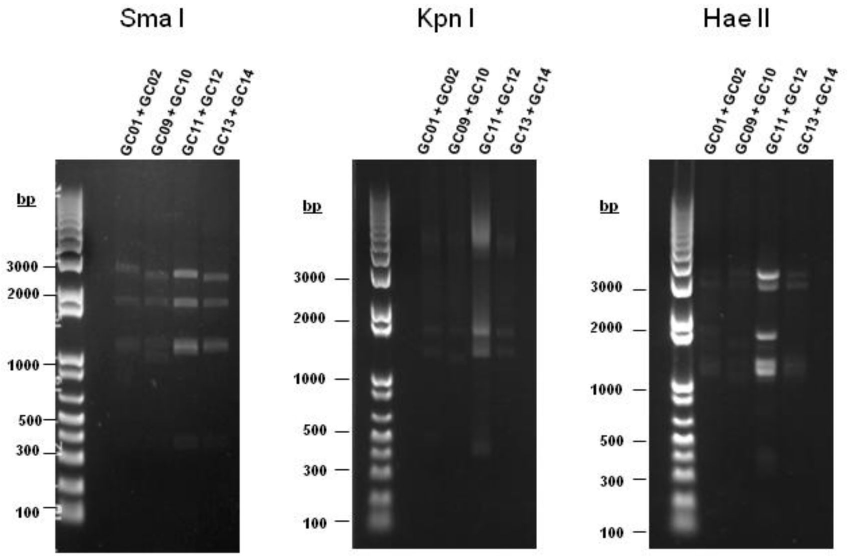


Figure 3.4: SmaI, HaeII and KpnI digests

The long range PCR products that were predicted to contain an *rrn* operon shown previously (**Fig. 3.3**) were digested with restriction endonucleases SmaI, HaeII and KpnI. Visualisation by agarose gel electrophoresis revealed restriction digests of the PCR products produced similar banding patterns for each of the endonucleases used.

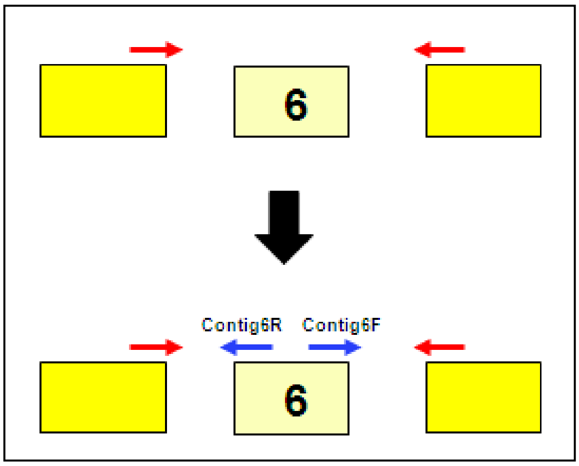


Figure 3.5: Adapted gap closure approach

Primers were designed to extend from Contig 6 (blue arrows) and were paired with primers previously designed to extend from neighbouring contigs (red arrows).

PCRs were performed using the new Contig 6 primers and the existing primers previously designed for the other contigs. The PCR reactions were carried out using each of the primers GC01, GC02, GC09, GC10, GC11, GC12, GC13 and GC14 paired with both Contig6F and Contig6R primers. Results of these PCR reactions are shown in **Fig. 3.6**. Multiple non-specific bands were evident in each lane, despite attempts to optimise PCR conditions (data not shown). However, dominant bands of ≤500 bp can be seen for each original primer paired with one or the other of the Contig 6 primers. These dominant bands were excised, purified, cloned and sequenced. This permitted two further gaps containing Contig 6 homologues to be closed i.e. contigs 2 and 7, and 7 and 8. At this stage, most of the seven original gaps in the genome had either been closed, or rationalized as shown in **Fig. 3.7**. Time limitations meant that this partially closed genome was used in further work described here. However in a following collaboration closing has been completed and the completed sequence submitted to the NCBI database, Genbank accession number CP003152 (**Appendix Four**, Pethick *et al.*, 2012b).

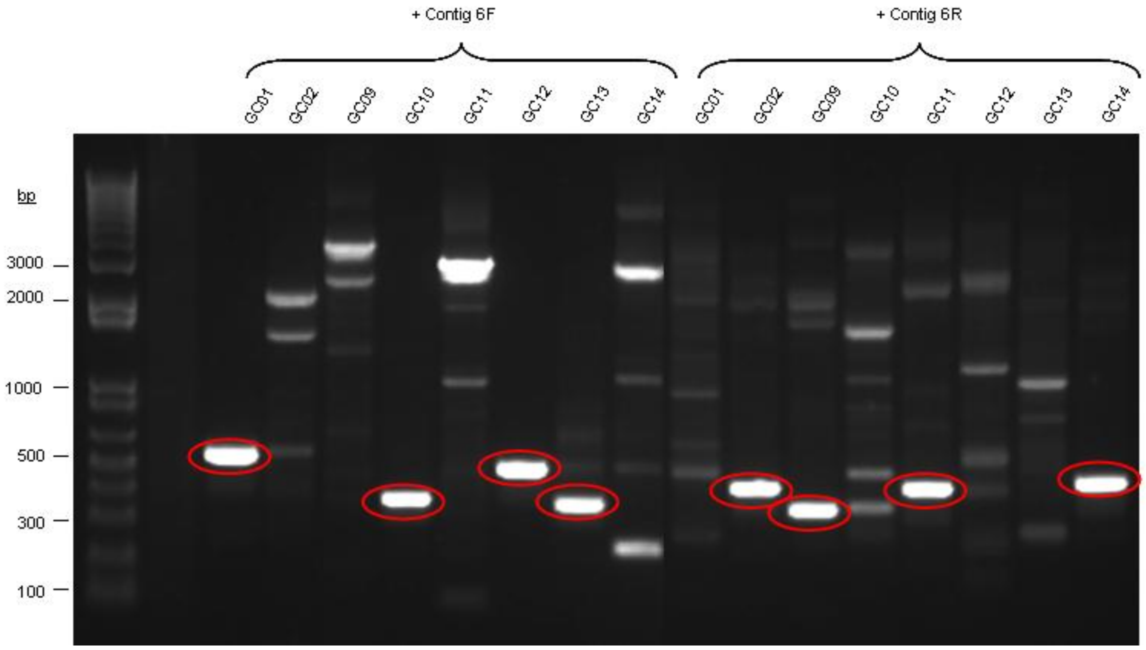


Figure 3.6: PCR bridging of gaps from ends of Contig 6

For each of the genome gaps containing an *rrn* operon, PCR bridging was continued with an adapted approach as depicted in **Fig. 3.5**. The primers GC01, GC02, GC09, GC10, GC11, GC12, GC13 and GC14 designed for original gap closure at these locations were each paired with both Contig6F and Contig6R primers. The primer location of the original primers in relation to the sequence contigs can be seen in **Fig. 3.2**, and that of primers Contig6F and Contig6R in **Fig. 3.5**. Following PCR reactions, dominant products of ≥500 bp were observed for reactions performed using original contig primers paired with either Contig6F or Contig6R primers (circled in red). These bands are the DNA fragments from the end of a Contig 6-like sequence (*rrn* operon) to the adjacent contig.

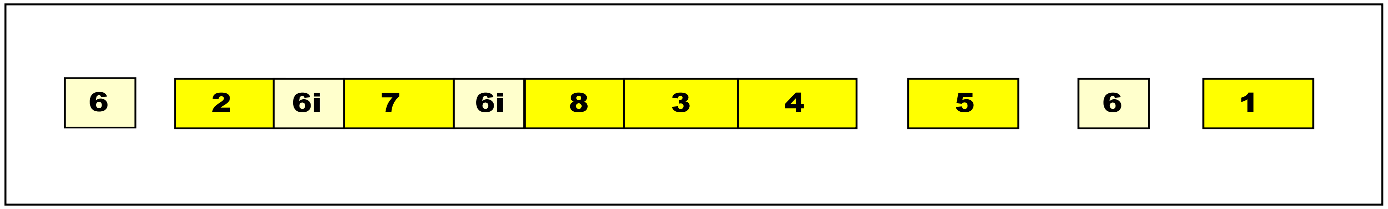


Figure 3.7: Linear map of contig arrangement post gap-closure

The locations of the four *rrn* operon-encoding regions similar to that of Contig 6 are shown amongst the other contigs. Two of the Contig 6-like regions are inverted (represented by 6i) as compared to the orientation of the *Cp* Contig 6, as provided by 454 Life Sciences.

3.2.3 Gene and protein prediction in *Cp* 3/99-5

Bioinformatic analysis of the *Cp* 3/99-5 genome commenced before the genome had been completely closed. This comprised open reading frame (ORF) prediction and assignment of function to these ORFS.

3.2.3.1 Open-reading frame (ORF) prediction

Two ORF prediction programs were used to identify potential coding sequences (CDSs); Glimmer v.3.02 and Prodigal v.1.10. Parameters were selected to suppress prediction of genes at junctions between contigs. A brief comparison of the output from these programmes is shown in **Table 3.2**.

Of the ORFs predicted by the different methods used, 923 were identical between all four predictions. Prodigal predicted the most ORFs but of a smaller average length. The Prodigal ORF prediction was used for all further work.

Table 3.2: Comparison of ORF predictions

Prediction	Prodigal	Glimmer A	Glimmer B	Glimmer C
Total no. genes	2153	1481	1376	1363
Average gene length/ bp	958	1139	1221	1231

To verify that these start and stop positions predicted by Prodigal had been correctly assigned, predicted ORFs from the *Cp* 3/99-5 genome were aligned to those from a published *C. diphtheriae* (NC_002935) genome using the Smith-Waterman global alignment algorithm. Where the predicted start and stop positions differed, the *Cp* 3/99-5 sequences were reviewed and edited where necessary.

3.2.3.2 Analysis of protein subcellular location

Predicted ORFs from *Cp* 3/99-5 were translated and the cellular location of these proteins was inferred using a set of predictive algorithms. The following programmes were used to predict subcellular location; PSORT-B, CELLO, SubLoc and GPos-PLoc; and SignalP predicted signal peptide cleavage sites. Discrepancies between the different programs used to predict subcellular location were observed, so an overall average location was made; the process used to assign a protein's location is shown in detail in **Fig. 3.8**. If the discrepancies were high, the GPos-PLoc location was taken.

To validate this methodology and the 'decision tree' (in **Fig. 3.8**), the assignment of subcellular location of protein examples of known locations were checked. The methods used here successfully assigned housekeeping proteins such as those encoded by *dnaA*, *pgk*, *rpoB* and *rplA* (as well as additional ribosomal proteins) to the cytoplasm. The integral membrane proteins SecF and SecD of the Sec pathway responsible for protein translocation were classified as membrane-bound, as was the cell division protein FtsW. Also the previously characterised serine protease CP40 was categorised as extracellular

Finally, of the complete predicted *Cp* 3/99-5 proteome, 122 (5.67%) proteins were predicted to be extracellular, 393 (18.25%) membrane-bound, 1599 (74.27%) cytoplasmic, eight (0.37%) were compartmentalised to the cell wall, and the subcellular location of a further 31 proteins (1.44%) remains "unknown".

3.2.3.3 Functional analysis of proteins

Predicted *Cp* 3/99-5 proteins were initially assigned a function by comparison to homologous sequences in the NCBI nr, Trembl and Swissprot databases using local BlastP. This information was used to build a GenBank formatted file that was subsequently edited and annotated using Artemis (Sanger). The *Cp* 3/99-5 genome annotation was added to as this work progressed using the output from multiple functional prediction programmes.

InterProScan, (locally installed) was used to predict protein functional domains using multiple analytical methods including the applications Gene3D, SignalP, HMMPanther and HMMPfam. The KEGG (Kyoto Encyclopaedia of Genes and Genomes) Automatic Annotation Server, KAAS (Moriya *et al.*, 2007), was also used to aid functional annotation of genes by BLAST comparisons against the manually curated KEGG GENES database. The program created KEGG Orthology assignments and automatically generated KEGG pathways (discussed in section **3.2.4.2**). Of the total 2,153 predicted proteins, 1,142 (53%) were assigned KEGG orthologues by KAAS. The large majority (approx. 70%) of proteins assigned KEGG orthologues have a metabolic function and more specifically this group is dominated by proteins involved in carbohydrate metabolism (16%) and amino acid metabolism (15%). However there are also a considerable number of proteins falling into both the genetic and environmental information processing categories (16% and 12% respectively). Functional groups of the KEGG orthologues can be seen in **Fig. 3.9**.

In addition to the 2153 predicted ORFS, 4 copies of the *rrn* operon (16S-23S-5S) were identified by the RNAmmer server and 49 tRNAs were also identified in the genome sequence using tRNA scan-SE. The location of these functional RNA genes can be seen in **Fig. 3.10**.

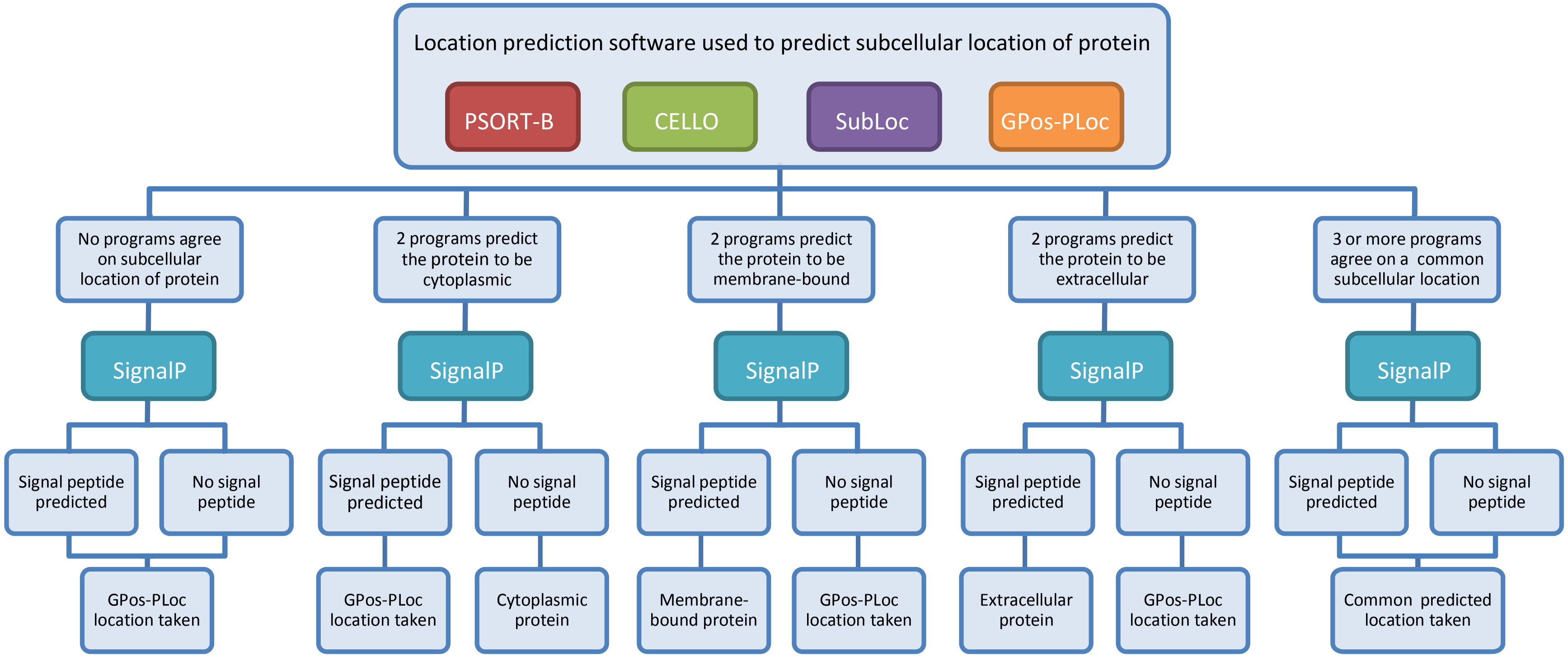


Figure 3.8: Flow chart indicating the determination of protein subcellular location

The flow chart is a “decision tree” of criteria used to determine the location of predicted *Cp* 3/99-5 proteins; this was worked through for each protein. The programs PSORT-B, Cello, SubLoc and GPos-PLoc to predict subcellular location are represented by a red, green, purple and orange box respectively. The program SignalP which predicts signal peptides is represented by the teal-coloured boxes. Proteins were input in to the programs as protein FASTA sequences.

3.2.4 Characterisation of the *Cp* 3/99-5 genome

General features deduced from the sequencing and bioinformatic analysis of the genome are shown in **Fig. 3.10**.

3.2.4.1 Chromosomal structure, G+C content variation and codon usage

The *Cp* 3/99-5 genome consists of a circular chromosome approximately 2.32 Mb in size. It has an overall G+C content of 52.18%; however, the G+C content is not constant across the genome. The GC skew ([GC]/ [G+C]) of the genome sequence (**Fig. 3.10**) splits the genome in two with the first half (approximately) of sequence showing a high GC skew compared to the average, and the other half has a much lower than average GC skew.

The codon usage variation amongst the different *Cp* 3/99-5 genes was determined using CodonW (version 5a) which calculated both the 'effective number of codons' (ENc) and the G+C content at the third synonymous codon position (GC3s). ENc values of the *Cp* 3/99-5 genes vary from 26.45 to 61.0, with a mean value of 53.4, and the GC3s range from 0.143 to 0.831 with a mean of 0.514. **Figure 3.11** shows the codon usage of *Cp* 3/99-5 genes compared to a reference curve representing the relationship between ENc and GC3s when codon usage is random. Although some of the *Cp* 3/99-5 genes lay on this curve, the majority do not.

3.2.4.2 Metabolic analysis

Rather than mention all metabolic pathways of the bacterium, focus is concentrated on a select few of importance, and those discussed can be seen in **Appendix Two.**

Analysis of carbohydrate metabolism revealed a complete set of genes encoding enzymes for glycolysis, gluconeogenesis and pentose-phosphate pathways. Furthermore, the tricarboxylic acid cycle is complete with the exception of the conversion between succinate and succinyl-CoA. All *de novo* amino acid biosynthesis pathways are present as is purine biosynthesis, although the pyrimidine pathway lacks cytidine triphosphate synthetase. The pathway for biotin production appears complete, however that for folic acid is not.

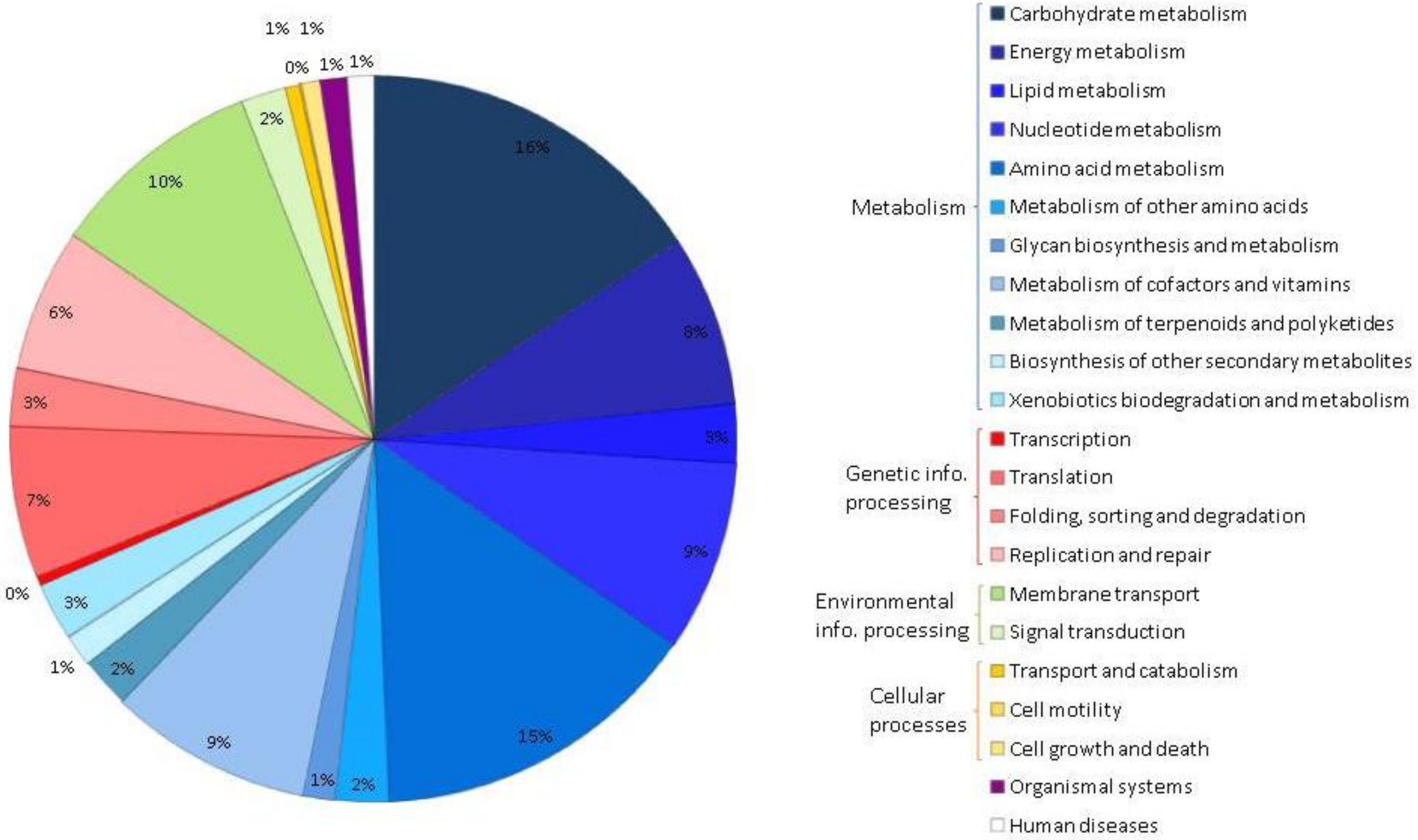


Figure 3.9: Functional classification of KEGG orthologues

Of the predicted proteome 53% has orthologues in the KEGG Orthology (KO) database, the majority of which have metabolic-functions. The largest functional category is carbohydrate metabolism, other large groups include amino acid metabolism and membrane transport, and the smallest groups include transcription and cellular processes.

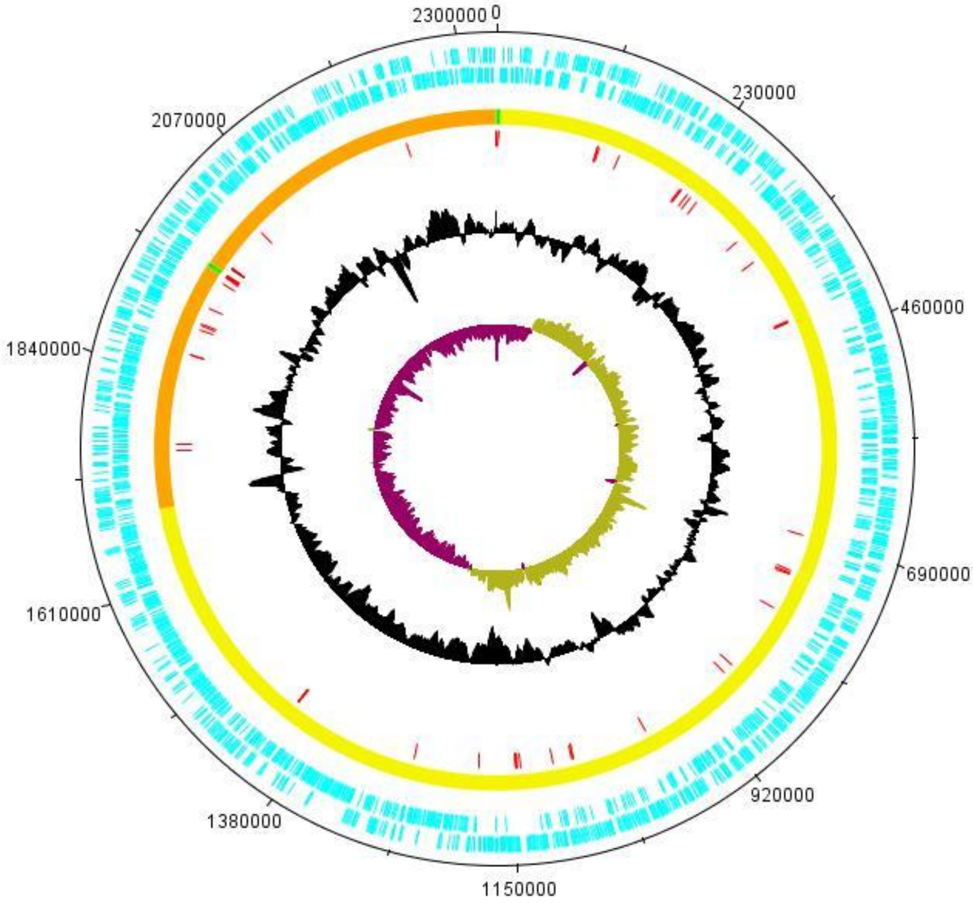


Figure 3.10: Circular DNA plot of the *Cp* 3/99-5 genome

(The plot shows the genome at an incomplete version of the genome prior to completion by collaborators.) The tracks from the outside represent: DNA bases (1);Forward CDS (2); Reverse CDS (3); Contigs (4); structural RNA molecules (5); %GC plot (6); GC skew ([GC]/ [G+C]) (7).

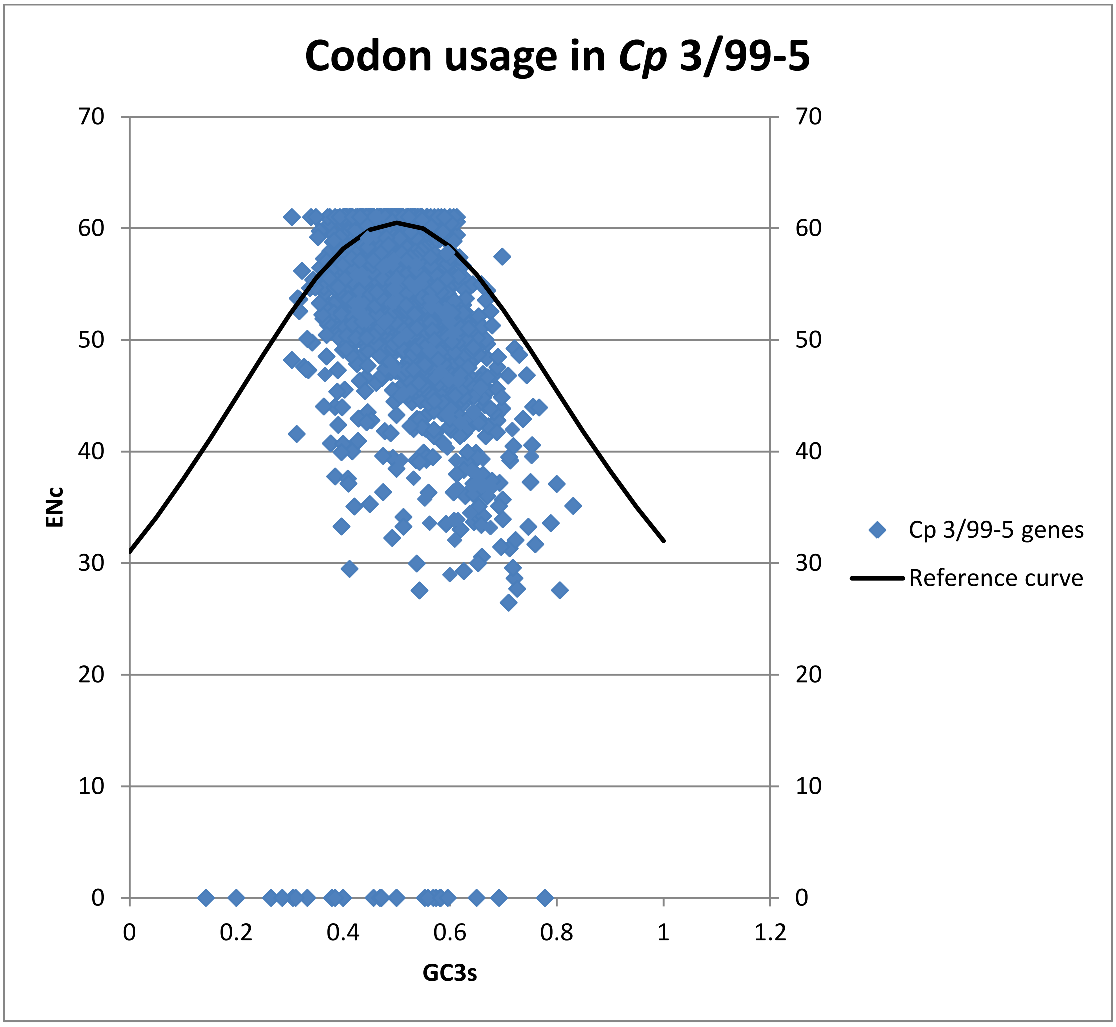


Figure 3.11: Graph of codon usage in the *Cp* 3/99-5 genome

The codon usage of all of the *Cp* 3/99-5 genes was determined using the effective number of codons (ENc), a measure of synonymous codon usage bias independent of gene length and amino acid composition. The ENc can range in value from 20, when one codon is exclusively used for each amino acid, to 61 when the use of alternative synonymous codons is of equal probability (Wright, 1990). CodonW software was used to calculate the ENc and the G+C content at the third synonymous codon position (GC3s), and the relationship between them is shown here in comparison to a reference curve which indicates the pattern generated by random codon usage.

3.2.4.4 Iron acquisition

During genome annotation several putative iron uptake systems were identified including several ABC transporters and putative siderophore systems (**Table 3.3)**.

Table 3.3: Putative iron uptake systems

In the host iron is not readily available and the *Cp* genome encodes several systems concerned with the acquisition of this essential element.

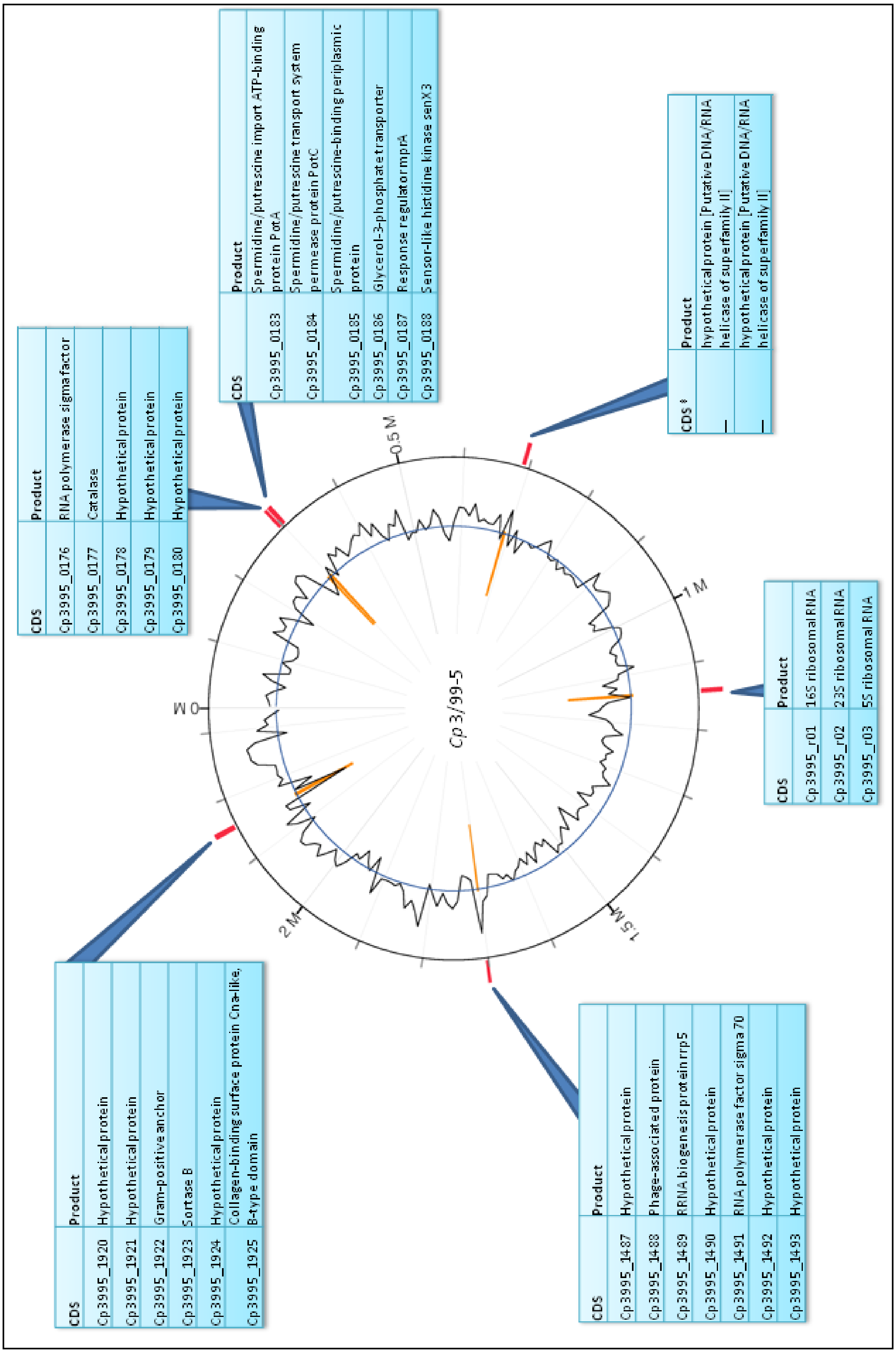
CDSs	**Genes**	**Characteristics**
Cp3995_0027-Cp3995_0030	*fagA*, *B*, *C* and *D*	Iron-siderophore uptake system. ABC transporter
Cp3995_0042-Cp3995_0045	*znuA, fecD,E, mgtE1*	Putative iron-siderophore uptake system
Cp3995_0457-Cp3995_0462	*htaA,B,C, hmuT,U,V,*	Haemin gene cluster. ABC transporter
Cp3995_0629-Cp3995_0637	*norM, cbs, lysA, arcB, tauD, oppA2,B2, C2,D1*	Iron-siderophore uptake system. ABC transporter
Cp3995_0890-Cp3995_0892	*fhuC,D,G*	ABC transporter
Cp3995_1002-Cp3995_1007	*ciuA,B,C,D,E,* and *F*	Iron-siderophore uptake system. ABC transporter
Cp3995_1510-Cp3995_1511		Cell-surface haemin receptor
Cp3995_1511-Cp3995_2043		Putative cell-surface haemin receptor

3.2.4.5 Genomic Islands

Genomic islands (GIs) were predicted using Island Viewer software which indicated six regions as GIs. The predicted GIs are within regions of anomalous G+C content and the CDSs in each of the GIs are shown in **Figure 3.12**.

3.2.4.6 Putative virulence factors

Annotation of the *Cp* genome has revealed a number of potential virulence factors which can be seen in **Table 3.4**. Potential virulence factors included in the table have homologues which have been linked to virulence in other bacteria; however, this is by no means an exhaustive list of *Cp* virulence factors.



CDS	Product
Cp3995_0176	RNA polymerase sigma factor
Cp3995_0177	Catalase
Cp3995_0178	Hypothetical protein
Cp3995_0179	Hypothetical protein
Cp3995_0180	Hypothetical protein

CDS	Product
Cp3995_0183	Spermidine/putrescine import ATP-binding protein PotA
Cp3995_0184	Spermidine/putrescine transport system permease protein PotC
Cp3995_0185	Spermidine/putrescine-binding periplasmic protein
Cp3995_0186	Glycerol-3-phosphate transporter
Cp3995_0187	Response regulator mprA
Cp3995_0188	Sensor-like histidine kinase senX3

CDS *	Product
–	hypothetical protein [Putative DNA/RNA helicase of superfamily II]
–	hypothetical protein [Putative DNA/RNA helicase of superfamily II]

CDS	Product
Cp3995_r01	16S ribosomal RNA
Cp3995_r02	23S ribosomal RNA
Cp3995_r03	5S ribosomal RNA

CDS	Product
Cp3995_1487	Hypothetical protein
Cp3995_1488	Phage-associated protein
Cp3995_1489	RRNA biogenesis protein rrp5
Cp3995_1490	Hypothetical protein
Cp3995_1491	RNA polymerase factor sigma 70
Cp3995_1492	Hypothetical protein
Cp3995_1493	Hypothetical protein

CDS	Product
Cp3995_1920	Hypothetical protein
Cp3995_1921	Hypothetical protein
Cp3995_1922	Gram-positive anchor
Cp3995_1923	Sortase B
Cp3995_1924	Hypothetical protein
Cp3995_1925	Collagen-binding surface protein Cna-like, B-type domain

Figure 3.12: Genomic islands predicted in the *Cp* 3/99-5 genome sequence

GIs (predicted by IslandViewer) and the genes within them are shown. The GIs are generally present at regions of abnormal G+C content. *Locus IDs are not given for these proteins as these CDSs were predicted using methodology in this thesis but are not present in the later version of the genome submitted to NCBI.

Table 3.4: Possible virulence factors encoded in the *Cp* 3/99-5 genome

CDS	Gene	Product
Cp3995_0026	*pld*	Phospholipase D
Cp3995_0128	*norB*	Nitric-oxide reductase, cytochrome b-containing subunit I
Cp3995_0391	*nanH*	Neuraminidase (sialidase)
Cp3995_0403	*mycP*	Subtilisin-like serine protease
Cp3995_0494	*pccB1*	Propionyl-CoA carboxylase beta chain 1
Cp3995_0495	*pccB2*	Propionyl-CoA carboxylase beta chain 2
Cp3995_0543		Hydrolase domain-containing protein
Cp3995_0570	*sprT*	Trypsin-like serine protease
Cp3995_0574		Non-ribosomal peptide synthetase
Cp3995_0603	*rpfA*	Resuscitation-promoting factor RpfA
Cp3995_0692	*rpfB*	Resuscitation-promoting factor RpfB
Cp3995_1099		Invasion-associated protein p60
Cp3995_1668		YcaO-like family protein
Cp3995_1673		Peptidase, S8A (subtilisin) family protein
Cp3995_1850	*mycB*	Mycosubtilin synthase subunit B
Cp3995_1927	*clpB*	ATP-dependent chaperone protein ClpB
Cp3995_1947		Serine protease CP40
Cp3995_2005	*accD*	Propionyl-CoA carboxylase beta chain 5
Cp3995_2164		Hypothetical protein (putative integral membrane protein *mviN*-like)

3.2.4.7 Fimbriae

Four sortase encoding genes are present in the genome (**Table 3.5**), one of which appears to be a 'housekeeping' sortase, Cp3995_2070. There are also several genes containing potential sortase recognition motifs (**Table 3.6**).

Interestingly all of the genes encoding sortase-like proteins except Cp3995_2070 are located in close proximity to at least one gene encoding a likely substrate protein. For example the Cp3995_1920 to Cp3995_1927 cluster of genes, a fimbrial associated region, includes two genes encoding sortases, three sortase substrate proteins as well as ClpB (**Fig. 3.13**).

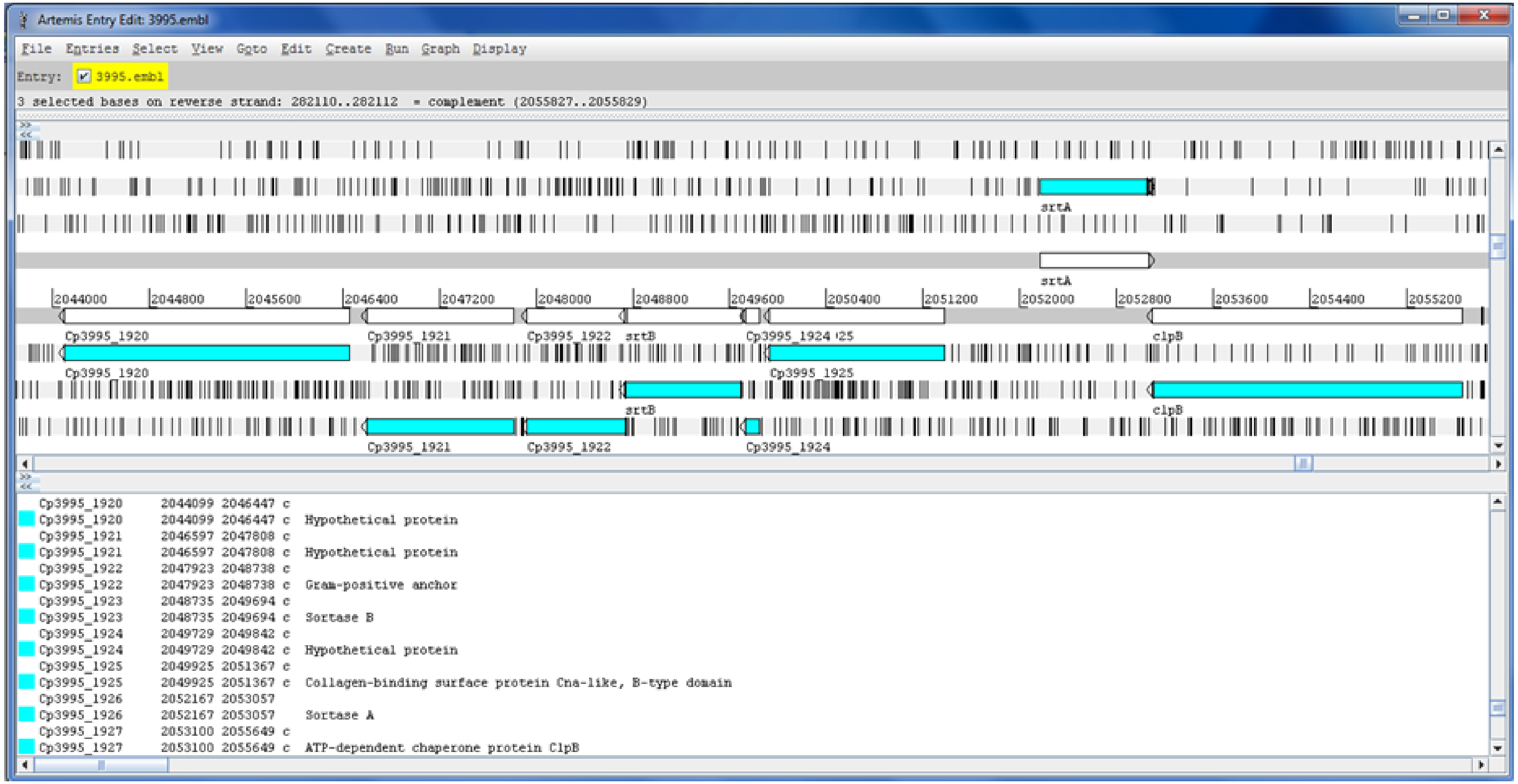
Table 3.5: Putative sortases present in the genome

CDS	Gene	Product
Cp3995_1926	*srtA*	Sortase A
Cp3995_1923	*srtB*	Sortase B
Cp3995_1954	*srtA*	Fimbrial associated sortase-like protein
Cp3995_2070	*srtB*	Sortase-like protein

Table 3.6: Putative sortase substrate proteins in the *Cp* 3/99-5 genome

Coding sequences containing potential sortase recognition motifs are shown.

CDS	Gene	Product	Potential sortase recognition motif
Cp3995_2123		Hypothetical protein	LANTG
Cp3995_0167		Hypothetical protein	LAATG
Cp3995_0169	*sdrD*	Serine-aspartate repeat-containing protein D	LAATG
Cp3995_0174		Surface antigen	LARTG
Cp3995_0316		Hypothetical protein	LPETG
Cp3995_1012		Hypothetical protein	LAKTG
Cp3995_1108		Hypothetical protein	LAKTG
Cp3995_1860		Hypothetical protein	LAKTG
Cp3995_1920		Hypothetical protein	LPKTG
Cp3995_1922		Gram-positive anchor	LPLTG
Cp3995_1925		Collagen-binding surface protein Cna-like, B-type domain	LPMTG
Cp3995_2020		LPxTG domain-containing protein	LAVTG



3 selected bases on reverse strand: 282110..282112 = complement (2055827..2055829)

```
  Cp3995_1920   2044099 2046447 c
■ Cp3995_1920   2044099 2046447 c  Hypothetical protein
  Cp3995_1921   2046597 2047808 c
■ Cp3995_1921   2046597 2047808 c  Hypothetical protein
  Cp3995_1922   2047923 2048738 c
■ Cp3995_1922   2047923 2048738 c  Gram-positive anchor
  Cp3995_1923   2048735 2049694 c
■ Cp3995_1923   2048735 2049694 c  Sortase B
  Cp3995_1924   2049729 2049842 c
■ Cp3995_1924   2049729 2049842 c  Hypothetical protein
  Cp3995_1925   2049925 2051367 c
■ Cp3995_1925   2049925 2051367 c  Collagen-binding surface protein Cna-like, B-type domain
  Cp3995_1926   2052167 2053057
■ Cp3995_1926   2052167 2053057    Sortase A
  Cp3995_1927   2053100 2055649 c
■ Cp3995_1927   2053100 2055649 c  ATP-dependent chaperone protein ClpB
```

Figure 3.13: The Cp3995_1920-Cp3995_1927 gene cluster

Both CDS (white) and gene (blue) features are shown; all code in the reverse direction except for *srtA*. Gene products are also indicated, and the cluster contains two sortases and three proteins, Cp3995_1920, Cp3995_1922 and Cp3995_1925, which contain an LPXTG sortase substrate motif.

The figure is taken from Artemis software.

3.2.5 *Cp* 3/99-5 genome comparison with *C. diphtheriae*

An almost complete *Cp* 3/99-5 genome was compared with the *C. diphtheriae* genome using Artemis Comparison Tool, (ACT). From which it would appear the *Cp* sequence is similar to the *C. diphtheriae* genome in size and G+C content as well as the number of genes contained within the sequence (**Table 3.7**). It became apparent that the genomes share an extremely high level of homology across the entirety of the sequence, so that the majority of genes were found in identical locations within the two genomes. Major differences appeared to be primarily associated with the acquisition of exogenous DNA through independent phage and transposon integrations. However comparison of the *Cp* 3/99-5 genome against the sequence of *C. diphtheriae* did reveal a number of predicted *Cp* proteins that are not present in the comparison genome.

Table 3.7: General features of the *C. pseudotuberculosis* and *C. diphtheriae* genomes

	***C. pseudotuberculosis* 3/99-5**	***C. diphtheriae*** NCTC13129*
Size (bp)	2,319,079	2,488,635
G+C content (%)	52.18	53.48
CDSs	2,153	2,320
Ribosomal RNAs	4 × (16S-23S-5S)	5 × (16S-23S-5S)
Transfer RNAs	49	54

***accession #: NC_002935**

3.3 Discussion

Genome sequencing of *Cp* 3/99-5 using 454 sequencing proved to be very efficient and allowed assembly of reads into eight contigs. This is in comparison to other projects that have been conducted in the same laboratory, whereby equivalent-sized genomes have assembled in >100 contigs (Colin Bayne; Sabrina Binkowski, both personal communication). Due to the low number of *Cp* contigs, it was hoped that the process of gap closing would be straight forward. In this respect, initial use of PCR to bridge the gaps in the genome successfully closed two gaps, joining contigs 8, 3 and 4. The large overlaps of sequence between the neighbouring contigs provided great assurance in the accuracy of these results.

Comparison of this preliminary version of the *Cp* genome with the available *C. diphtheriae* genome revealed high homology between the two species (along the entire length of both sequences). Interestingly the sequence of *Cp* Contig 6 shared significant homology with a region encoding 5S, 16S and 23S rRNA genes that is present in no less than five locations within the *C. diphtheriae* genome. By analogy, it was considered possible that the failure, during assembly, to associate Contig 6 with any of the other contigs could be a result of sequence data originating from multiple chromosomal locations being assembled into a single contig (Contig 6). Slight sequence differences in rRNA-gene regions across the chromosome would result in inevitable degeneration in the integrity of the sequence of Contig 6, so it would be impossible to assemble it within any of the other contigs. Further evidence for this hypothesis was obtained following a first-pass annotation of the available *Cp* genome contigs which revealed that four of the *C. diphtheriae* rRNA-operons mapped to gaps between contigs in the *Cp* genome, suggesting that *Cp* Contig 6 could be “slotted” in to fill the gaps. However, the final *C. diphtheriae* rRNA-encoding region mapped to within *Cp* Contig 1. There was no evidence that the assembly was incorrect (since it was co-linear with *C. diphtheriae* in every other way), and therefore it is assumed that the extra rRNA-encoding region in *C. diphtheriae* is due to divergent evolution of the two organisms, which likely arose from a common ancestor. Significantly, possession of multiple rRNA-encoding regions suggests an ability to generate greater numbers of ribosomes, which in turn will result in a higher level of protein synthesis. This could have a profound outcome on the ability of these two very successful pathogens to colonise and persist within their chosen host species.

Long-range PCR was used to confirm the presence of rRNA-encoding regions in the four predicted locations, and subsequent restriction analysis of PCR amplicons proved the predictions to be correct. Unfortunately, sufficient funding was not available to commit to sequencing all four of the rRNA regions, so as a preliminary step it was decided to sequence only one, and compare the sequence to "consensus" sequence of Contig 6, which must have derived from assembly of sequence reads originating from all four rRNA loci. However, repeated attempts at sequencing failed, despite confirmation that the DNA was of sufficient quality. Furthermore, the same primers used to amplify the PCR product were used for sequencing, and there should have been no doubt about primer specificity. Therefore, the reason(s) for the sequencing failure is unknown. As a result, restriction analyses were conducted instead, with a view to presenting the final genome assembly with Contig 6 placed in four locations and denoted as a consensus sequence. The results of these analyses revealed that sequences flanking the conserved core region varied in length between each product. It was therefore decided that, in order to allow Contig 6 to be accurately joined to its neighbouring contigs within four of the gaps, it would be necessary to determine the unknown sequences flanking the conserved core region. To achieve this, primers were designed to extend outwards from the ends of Contig 6 and PCR bridging reactions using these coupled with original bridging primers were performed. This technique further confirmed the accuracy of the predictions that Contig 6 is present in multiple copies within the genome, and allowed two of the Contig 6 loci to be joined to neighbouring contigs. Significantly, some of the *C. diphtheriae* rRNA-encoding regions are inverted as compared to the orientation of the *Cp* Contig 6. During the process of closing the gaps between Contig 6 and neighbouring contigs, two of the rRNA-encoding domains in *Cp* (between contigs 2 and 7, and 7 and 8) have been found to be inverted. At this point the gaps in the genome sequence had been significantly reduced from the original eleven contigs (assuming Contig 6 is present in four copies). Gaps still existed between contigs 4 and 5, 5 and 1, and 1 and 2; the latter two gaps were predicted to have rRNA regions within them. Also Artemis analysis of the *Cp* genome indicated the start of a gene just before the end of Contig 4, which appeared to continue into the start of Contig 5. *C. diphtheriae* comparison indicated the start of the same gene (*aceF*, encoding a dihydrolipoamide acetyltransferase) at a similar locus in this genome and suggested that the gene was present in *Cp* 3/99-5 and spanned the gap between contigs 4 and 5.

It was recognised that remaining gaps were very small; however, it was possible that sequences of particularly small genes had been missed. Also even very small gaps could contain sequence of importance, for example that of small RNAs (sRNAs). These are small usually non-coding RNA genes of 50-500 bp in length (Hershberg *et al.*, 2003). Although some sRNAs possess a protein-coding function, most do not instead acting as posttranscriptional regulators modulating target gene expression (Waters & Storz, 2009). Small RNAs have important regulatory roles involved in many crucial processes including stress response (Wassarman, 2002), quorum sensing (Lenz *et al.*, 2004) and biofilm formation (Mika & Hengge, 2013). Despite their importance, the small sequence length of many sRNAs may result in many annotations omitting them, due to conventional gene prediction and genome annotation software having defined gene sizes. Indeed in the current study, the minimum length specified for gene prediction was 90 bp which will have resulted in many of these sequences remaining unidentified. There are several sRNA databases now available, including sRNAdb (Pischimarov *et al.*, 2012) and the Bacterial Small Regulatory RNA Database (BSRD; Li *et al.*,2013), and these could be used to predict sRNAs in the *Cp* 3/99-5 genome

In the interests of time no further closure of the genome was performed in this work. That said the genome has since been completed by our Brazilian collaborators as part of a wider pan-genome project.

To ensure a timely continuation of the project, bioinformatic analyses, including gene annotation, were commenced using the near complete genome sequence (at the stage when only contigs 8, 3 and 4 had been joined); at this time it was not considered that a significant number of potential ORFs had been missed due to the gaps. Additional gap closure and indeed genome completion achieved since then has led to the identification of no new ORFs. Thus the “draft” genome produced in this work was of high quality, and the sequence needed to complete the genome was irrelevant in terms of using the genome to lead identification of candidate diagnostic targets.

To facilitate ORF prediction Prodigal software was used instead of the more commonly used Glimmer program. A preliminary comparison of the two programs revealed that while predicted ORFs were generally equivalent, under the parameters employed in this study Prodigal identified more ORFs than Glimmer. Glimmer was not as successful at predicting small genes, which is reflected in a smaller average gene length of the Prodigal prediction compared to that of the Glimmer predictions. However it is also likely that some of the very small ORFs predicted by Prodigal are predicted incorrectly. For the process of target identification in Chapter Six those genes that were small (less than 75 bp) and had no hits reported from BLAST searches and InterProScan were dismissed.

Although Corynebacteria typically have a high G+C content, the G+C content of species belonging to the *Corynebacterium* genus ranges greatly from 51 to 71%. Indeed the current sequencing project shows *C. pseudotuberculosis* is typical of its genus (with a G+C content of 52.18%). Analysis of the GC skew of the genome revealed a split of high GC skew for the first half of the sequence and a lower than average skew for the remaining half. This pattern is indicative of a bidirectional replication mechanism, and suggests the origin of replication to be around base 109,500 of the incomplete genome, further substantiated by the presence of *dnaA* and *dnaN* genes and similarity with the sequence of *C. diphtheriae* (Cerdeno-Tarraga *et al.*, 2003). The terminus of replication appears to be at a position 180° from the replication origin. These positions were confirmed following subsequent completion of the *Cp* 3/99-5 genome. Interestingly there is a region of lower G+C content approximately 750 kb in length which appears to cover the terminus of replication. This region of lower G+C content mirrors a similar region found in *C. diphtheriae*, however the non-pathogenic bacteria *C. glutamicum* and *C. efficiens* lack such a large genomic change in their G+C content.

It has been suggested that plotting the effective number of codons (ENc) against the G+C content at the third synonymous codon position (GC3s) could be used to effectively determine codon usage variation of genes (Wright, 1990). The codon usage of the majority of *Cp* 3/99-5 genes is biased, and has a similar bias to GC3s as previously observed for *C. glutamicum* (Liu *et al.* 2010). This variation of codon usage is likely influenced by amino acid conservation, translational selection and mutational bias.

There are 61 structural RNA genes in the current genome: 49 tRNA genes and as previously mentioned four sets of the 16S, 23S and 5S ribosomal RNA genes (*rrn* operon). Whilst other *C. pseudotuberculosis* isolates also contain four *rrn* operons (Cerdeira *et al.*, 2011a; Trost *et al.*, 2010; Silva *et al.*, 2011), the *C. diphtheriae* and *C. efficiens* genomes contain five *rrn* operons (Cerdeno-Tarraga *et al.*, 2003; Yukawa *et al.*, 2007), *C.*

glutamicum contains six (Kalinowski *et al.*, 2003) and *C. jeikeium*, three (Tauch *et al.*, 2005). In fact the ribosomal RNA operon copy number per bacterial genome can vary from one to fifteen (Andersson *et al.*, 1995; Bercovier *et al.*, 1986; Loughney *et al.*, 1983; Rainey *et al.*, 1996). There is some uncertainty over the reason for multiple copies of the *rrn* operon although the persistence of these in genomes of many species suggests a selective advantage to these organisms in retaining the extra copies. There is a general assumption that multiple *rrn* operons are required for high growth rates via increased production of ribosomes, however in some species not all *rrn* operons present are necessary for optimal growth rates (Ellwood & Nomura, 1980; Widom *et al.*, 1988) . Condon *et al.* (1995) reported that only five of the seven *rrn* operons in *E. coli* are required to reach near-optimal growth on complex media, however, all seven operons were required to rapidly adapt to temperature and nutrient changes. These findings are supported by similar studies including that of Stevenson and Schmidt (2004) who concluded that multiple *rrn* operons give a selective advantage allowing bacteria to respond quickly and grow rapidly in environments defined by fluctuations in resource availability. Klappenbach *et al*. (2000) also found that the number of rRNA genes correlated to the rate that bacteria responded to the availability of resources. Other explanations of the high copy number of *rrn* operons have also been made; Strätz *et al*. (1996) suggested the possibility of rRNA genes serving as natural vehicles for horizontal gene transfer.

The 49 tRNA genes present recognise 42 of the 61 possible sense codons. The number of tRNAs in the current genome is slightly less than those in other *Corynebacterium* species such as *C. diphtheriae* which has 54 (Cerdeno-Tarraga *et al.*, 2003), or *C. glutamicum* which has between 58 and 60 (Kalinowski *et al.*, 2003; Yukawa *et al.*, 2007). The present number does however appear typical of *Cp*; other *Cp* genomes published since this work was performed contain either 48 or 49 tRNAs (Trost *et al.*, 2010; Cerdeira *et al.*, 2011a; Silva *et al.*, 2011).

Thorough protein localisation analysis was conducted using four different software programs for predicting subcellular location; PSORT-B, CELLO, SubLoc and GPos-PLoc; as well as a program (SignalP) for predicting signal peptide cleavage sites. This latter program is important because signal peptides are short peptide chains that direct the transport of a protein to a specific subcellular location hence this may be indicative of extracellular proteins. With regards to this study, proteins located at the cell membrane and those transported out of the bacterial cell are crucial as these are the first proteins to which the host immune system is exposed. The result from GPos-PLoc was used if descrepencies were observed between location prediction software as this program has previously been described as the general localisation tool with the best predictive performance amongst those which do not base predictions on homology searches against SWISS-PROT in a validation study for the prediction of proteins of the related mycobacterial species (Restrepo-Montoya *et al.*, 2009). This methodology for assigning a location to each protein was validated by ensuring proteins of known locations were assigned correctly. The location of several proteins were identified correctly including ribosomal and housekeeping proteins which provided assurance in the methodology and the use of the 'decision tree' depicted in **Fig. 3.8**.

As would be expected, the large majority of proteins (~75%) are cytoplasmic, however, there are still a considerable number that are predicted as bound to the membrane or secreted out of the cell. There were 122 genes predicted to encode extracellular proteins, however, this number is considerably more than other studies have found. Pacheco *et al.* (2011), using a proteomic approach, found that for the two *Cp* strains investigated, 1002 and C231, 70 and 67 proteins of the exoproteome were identified respectively. Only eight proteins were predicted to be part of the cell wall; the reason for such a low number lies in the prediction software methodology as not all of the computational tools used assign proteins to this compartment. Hence proteins determined as belonging to the cell wall result from high discrepancies between different program outputs meaning a GPos-PLoc result was taken. Despite the methodology used the location of some proteins remains unknown, this is due to large discrepancies between location predictions and where the GPos-PLoc program has failed to assign a location-largely due to a short protein sequence.

KAAS was used to help determine function of the predicted proteins of the genome, and assign KEGG orthologues to the proteins, allowing characterisation of proteins into functional categories. The KEGG Automatic Annotation Server (KAAS) searches both the manually curated KEGG GENES database (which includes data from 1,358 bacteria, 151 eukaryotes and 114 archaea) and the KEGG Orthology (KO) database (containing 15,173 KO groups). The KO analysis indicated that approximately half of the predicted *Cp* 3/99-5 proteome have orthologues in the KO database. The large majority of these proteins are associated with metabolism, more specifically carbohydrate and amino acid metabolism. There are also a considerable proportion of proteins concerned with the processing of genetic and environmental information. The number of *Cp* 3/99-5 proteins with KEGG orthologues associated with cellular processes is minimal. Using the current methodology, the observed dominance of protein functional category by metabolic proteins is likely enhanced by the large focus of KEGG databases on this functional class. An alternative, widely used, database that could have been analysed to characterise proteins by function is 'The Cluster of Orthologous Groups' (COG) database. However, this database does not appear to be maintained and still only contains proteins from 66 genomes as described by Tatusov *et al.* (2003).

It is not the aim to mention all metabolic pathways of the bacterium in the current study and focus has been on the major carbohydrate, nucleotide and biotin metabolism, as well as biosynthesis of amino acids, fatty acids and folate. Analysis of metabolism indicated the presence of complete glycolysis, gluconeogenesis and pentose-phosphate pathways. The tricarboxylic acid cycle is almost complete, lacking only the conversion between succinate and succinyl-CoA. The enzyme succinyl-CoA synthetase (encoded by *sucC* and *sucD*) that usually catalyses this step is absent. However, *Cp* likely utilises the product of *cat1*, which has been shown to act as a succinyl-CoA: coenzyme A transferase in *Clostridium kluyveri* (Sohling & Gottschalk, 1996). Interestingly *C. diphtheriae* also lacks succinyl-CoA synthetase but has a homologue of *cat1*, *DIP1902*, (Cerdeno-Tarraga *et al.*, 2003). It stands to reason that these organisms compensate for the absence of succinyl-CoA synthetase by using an acyl-CoA: CoA transfer to form succinate and acetyl-CoA.

All *de novo* amino acid biosynthesis pathways are present, as is purine biosynthesis; however the pyrimidine pathway lacks cytidine triphosphate synthetase. The pathway for biotin production appears complete, however that for folic acid is not. The bulk of fatty acids appear to be synthesised as result of Fas, a fatty acid synthase encoded by *fas*. However, it would also appear that the product of Cp3995_0142 is responsible for some stages of the production of fatty acids; the encoded protein is similar to FabG, a 3-oxoacyl-[acyl-carrier protein] reductase.

Iron is essential for living organisms and contributes to many vital biological processes such as respiration, oxygen transport, gene regulation and DNA biosynthesis (Andrews *et al.*, 2003). In the host iron is not readily available to pathogens as the host actively limits iron availability to invading microorganisms as part of its innate immune system. By employing iron-binding proteins, such as transferrin and lactoferrin, mammals reduce the amount of free extracellular iron to levels insufficient for bacterial growth (Andrews *et al.*, 2003). In order to counter this iron restriction bacteria have evolved several mechanisms to obtain iron from their environment. Indeed *Cp* has several iron uptake systems and is able to sequester iron from both transferrin and lactoferrin in iron-chelated medium restoring growth to equivalent levels in un-chelated media (Walker, 2009). It is unclear by what mechanism the organism is able to achieve this, but it may involve the use of siderophores. Many pathogenic bacteria produce siderophores, which are low molecular mass proteins (<1000 Da) characterised by a high specificity and affinity for ferric iron, which scavenge iron competing against host iron-binding proteins. Generally siderophores are secreted by bacteria in response to iron restricted environments, however some remain associated with the cell envelope, such as the mycobactins of mycobacteria (Ratledge & Dover, 2000). *Cp* produces siderophores to acquire iron (Walker, 2009) and the *Cp* 3/99-5 genome encodes several putative iron-siderophore uptake systems including the *fagABC* operon previously linked to virulence (Billington *et al.*, 2002). Another *Cp* siderophore uptake system is the ABC transporter encoded by the *fhuCBG* operon which is concerned with ferric (Fe^{3+}) hydroxamate uptake. Cabrera *et al.* (2001) found strains of *Staphylococcus aureus* with mutations in this operon were unable to grow in iron-restricted media when supplemented with the ferric hydroxamate ferrichrome.

Corynebacterium pseudotuberculosis also has mechanisms in place to target haem, haemin and haemoglobin as sources of iron; the genome encodes numerous ABC transporters likely to be involved in iron uptake, including the *hmuTUV* operon. This operon is also found in *C. diphtheriae* and encodes products which have roles in the utilisation of haemin/haemoglobin (Brown & Holden, 2002). *C. diphtheriae* also scavenges iron directly from haem using haem oxygenase, HmuO encoded by *hmuO* (Schmitt, 1997), and *Cp* too possesses *hmuO* indicating that it may share a similar mechanism.

Genomic island (GI) is a general term for a large genomic region or cluster of genes that have been acquired by horizontal transfer (Langille *et al.*, 2010). These regions may be of particular interest as horizontal transfer can often lead to the acquisition of virulence factors and antibiotic resistance genes from one organism to another. Where the genes of a GI are involved in pathogenesis, the region is commonly referred to as a pathogenicity island (PAI) (Hacker *et al.*, 1990). In the current study IslandViewer software was employed to detect GIs. The resulting GIs predicted are all in regions of abnormal G+C content and include many proteins of unknown function. It is likely that these regions have occurred by horizontal acquisition from phage, indeed this statement is supported by the presence of phage-associated proteins in some of these regions.

IslandViewer has indicated a region containing a spermidine/putrescine transport system as a GI. The polyamines putrescine and spermidine are associated with a wide range of biological reactions necessary for cell growth including protein and nucleic acid synthesis (Tabor & Tabor, 1985). The region comprising Cp3995_1920 to Cp3995_1925, discussed later, was also flagged as a GI, and appears to be part of a gene cluster associated with fimbrae and is likely a PAI. Also one of the predicted GIs consists of an *rrn* operon; however, this prediction is questionable and may be due to the incompleteness of the genome sequence used for this analysis. This region spans a previous gap in the genome where Contig 6 has been ‘slotted in’

It was not the aim of this work to produce an exhaustive list of virulence factors, and indeed putative virulence candidates can not be confirmed as virulence factors without further work (including knock-out experiments) to substantiate homology searches. Furthermore, there is some debate over what constitutes a 'true' virulence factor as there are many different genes (such as housekeeping genes) that if absent would result in reduced cell growth or death, but do not as such add to pathogenicity other than allowing cell survival. Virulence gene discovery is also made difficult due to phenotypes often resulting from expression of multiple genes, so knocking out one or two genes may not render a mutant unable to survive or cause disease (McNeil & Aziz, 2009). It has been preposed that virulence determinants may be divided into four functional classes: (i) adhesion, (ii) toxin production, (iii) avoidance of host defences and (iv) somatic cell invasion and replication (Levin & Edén, 1990). However, some proteins are bifunctional having a biochemical role that may be conserved amongst many taxa but also a host-specific role with virulence potential. For example, streptococcal glyceraldehyde 3-phosphate dehydrogenase (GAPDH) is an essential enzyme involved in the glycolytic pathway, but it can also function in the adhesion to host components (Oliveira *et al.*, 2012). In the current work, should more time and resources been available, studies could have been performed to conclusively identify and investigate virulence determinants as well as satisfying molecular Koch’s postulates and classifying the proteins functionality as described above. However, instead here those proteins discovered during the annotation process to have a homologue linked to virulence in another organism were grouped together in a table of 'potential virulence factors'.

Virulence factors identified in the *Cp* genome include the previously described Phospholipase D and *Corynebacterial* serine protease CP40 enzymes (discussed in Chapter One). Analysis of the genome has also revealed two subtilisin-like serine proteases, a trypsin protease and ClpB, all possibly involved in virulence. Such serine proteases have roles in a wide range of biological processes but are also necessary for virulence in many pathogenic bacteria (Makinoshima & Glickman, 2006). This includes Gram-positive pathogens, for example Bonifait *et al.* (2010) found a subtilisin-like proteinase to be linked to virulence in *Streptococcus suis.* Also the trypsin-like protease *sprT* of the current genome contains a Hap active site, the Hap family, 'Haemophilus adhesion and penetration', contains proteins concerned with the interaction of epithelial cells. ClpB is an ATP-dependent chaperone belonging to the AAA family, a group of ATPases with varying cellular activities. The protein serves to disaggregate misfolded and aggregated proteins and is controlled by extracytoplasmic function sigma factor SigH. ClpB has been linked to virulence in other Gram-positive bacteria; Chastanet *et al.* (2004) revealed the protein to be involved in the pathogenicity of *Listeria monocytogenes* reporting affected virulence of a *clpB* mutant in a murine disease model. Also a *Mycobacterium tuberculosis* mutant lacking the *sigH* transcription factor was shown to be non-lethal despite achieving high bacterial counts in the lungs and spleen tissues in an identical manner to the wild-type (Kaushal *et al.*, 2002).

The *norB* gene encodes nitric oxide reductase which is important in the organisms defence against nitric oxide toxicity, which may be of particular importance in allowing survival within macrophages. Reactive nitrogen intermediates such as nitric oxide have antibacterial properties including direct bacterial killing, (Nathan & Shiloh, 2000) and macrophages produce nitric oxide through inducible nitric oxide synthase (Shiloh *et al.*, 1999).

There are three genes encoding β-chains of propionyl-CoA carboxylase in the *Cp* genome, these genes are necessary for fatty acid or mycolic acid synthesis (Gande *et al.*, 2007). Also resuscitation-promoting factors RpfA and RpfB, coded for by genes *rpfA* and *rpfB*, are responsible for resuscitation of bacterial dormancy and are required for virulence in *Mycobacterium tuberculosis* (Kana et al., 2008). Another candidate virulence factor, Cp3995_1668, encodes hypothetical protein similar to YcaO in *E. coli*; this protein is possibly involved in pathogenesis of *C. pseudotuberculosis*. Over-expression of YcaO is responsible for filamentous biofilm formation and reduced motility (Tenorio *et al.*, 2003).

In many Gram-positive pathogens, sortase-like proteins are responsible for anchoring surface proteins to the cell wall. The current genome has four genes encoding putative sortases (**Table 3.5**), one of which (Cp3995_2070) seems to be a 'housekeeping' sortase as it would appear to be part of the corynebacterial chromosomal backbone being present in *C. diphtheriae, C. jeikeium, C. glutamicum* and *C. efficiens.*

Also there are 16 genes encoding putative surface-anchored proteins with a cell wall sorting motif of LPXTG or LAXTG recognised by sortase-like proteins (**Table 3.6**). Although only LPXTG and LAXTG sorting motifs were discovered in *Cp*, SPXTG, PPXTG, LSXTG, and LGXTG are also common in Gram-positive bacteria (Pallen *et al.*, 2001). Interestingly one of the proteins containing such a motif (Cp3995_1012) was indicated as a cytoplasmic protein, which cannot be the case if it possesses this cell surface anchoring region. This possibly draws attention to the limitations of protein localisation software, which may have faults and cannot be relied upon fully.

The genes encoding sortase-like proteins with the exception of the housekeeping sortase are located close to at least one potential substrate protein. This close proximity of the sortase to substrate genes may indicate their acquisition at the same time by horizontal gene transfer. There appears to be a domain of potential virulence-related genes present in the current genome but mostly lacking from that of *C. diphtheriae* (Cp3995_1920-Cp3995_1927). This domain is associated with fimbriae and within this group there are two sortases (Cp3995_1923 and Cp3995_1926), three genes encoding potential sortase substrates (Cp3995_1920, Cp3995_1922 and Cp3995_1925) including one of which likely encodes a fimbrial subunit (Cp3995_1925), and Cp3995_1927 encodes ClpB.

At the time the present work commenced *C. diphtheriae* was the closest sequenced relative of *Cp*, and on comparison the genomes were found to be highly homologous to each other. Although the *Cp* genome is somewhat smaller than that of *C. diphtheriae*, homology extended across the entirety of the sequences and main differences appear to be associated with additional phage inserts present in the *C. diphtheriae* genome.

In the current study the *Cp* 3/99-5 genome was sequenced to almost entirety and analysed using suitable bioinformatic tools. Identification of the coding sequences within the sequence and the analysis of the subcellular location and function of corresponding proteins allowed characterisation of the genome. Insights into chromosomal arrangement, metabolic functions and iron acquisition have been gained as well as inferences made into the pathogenicity of *Cp,* with a number of candidate virulence factors identified. With the information this study brings to the field, there is a much greater hope of understanding the pathogenicity of *C. pseudotuberculosis* and finding novel proteins which may be exploited for vaccine and/or diagnostic purposes. The genome has since been fully completed as part of a subsequent collaboration and posted in the public domain via submission to the NCBI database. The current genome and indeed the additional *Cp* genomes that have recently been made publically available are invaluable tools in the study of this organism, its pathogenesis and the control of CLA.

Chapter Four: Sequencing of additional *Corynebacterium* genomes

4.1 Introduction

As mentioned in the previous chapter, there were no *Cp* genomes in the public domain when this project started, however several have since become available including one isolate, CIP 52.97 (Genbank CP003061), belonging to biovar *equi* that was isolated from a horse in Kenya (Cerdeira *et al.*, 2011b). Furthermore, there also lacked a *C. ulcerans* genome sequence in the public databases, however, since then two sequences have been published and described in a comparative paper (Trost *et al.*, 2011). (This genome information is correct at the time of writing, December 2011.)

Sequencing of additional *Cp* genomes and others in the *Corynebacterium* genus will allow comparative analyses to be performed both on a species but also genus level. Comparative genomics involves the direct comparison of the genetic material from one organism against that of another, and can determine the functions of genes and gain a better understanding of how species have evolved. Intra-species comparative analyses are becoming commonly used to indicate the ‘pan genome’ of bacteria, i.e. identifying those core genes common to all strains of a particular species. These studies are important as a single genome sequence does not reflect the variability driven by pathogenesis within a species, and hence limits screening for vaccine/diagnostic candidates. Tettelin *et al.* (2005) used comparative analyses to attempt to reveal the ‘pan genome’ of *Streptococcus agalactiae* and found that ~80% of any single genome belonged to the core genome. Another study which also shows that a significant proportion of the genes in a genome are specific to an individual strain was conducted by Hogg *et al.* (2007) who investigated *Haemophilus influenzae* and similarly found that ~80% of genes from a given genome belonged to the core genome.

Additional applications of comparative genomics have also been described such as its use in the clustering of regulatory sites (van Nimwegen *et al.*, 2007). Comparative analyses of fungal genomes have led to identification of putative targets for novel antifungal agents (Odds, 2005). Also regulatory motifs can be distinguished from other non-functional sequence using comparative analyses. T-box antitermination is a major mechanism of regulation of genes that are involved in amino acid metabolism in Gram-positive bacteria, and Vitreschak *et al.* (2008) used comparative analyses to identify T-box regulatory sites in various bacterial species.

Perhaps the most time consuming part of generating usable genome sequences is the annotation of coding sequences, and this is especially applicable when multiple genomes are to be annotated. However, numerous automated annotation tools have been developed to considerably reduce the time necessary for this step. These include The Joint Genome Institute's (JGI) Integrated Microbial Genome (IMG) system (Markowitz *et al.*, 2010), the National Microbial Pathogen Data Resource's (NMPDR) Rapid Annotation using Subsystems Technology (RAST) server (Aziz *et al.*, 2008), the KEGG Automatic Annotation Server (Moriya *et al.*, 2007) and the xBASE bacterial genome annotation service. The disadvantage to automatic annotation is a higher level of errors produced in the inference of coding sequence location and function. Most automatic annotation methods rely on homology to transfer annotations from the sequence of a reference to the new genome. However, if the sequence of a phylogenetically close reference is not available, annotation is insufficient resulting in many coding sequences remaining unannotated. Furthermore, even in homologous sequences, novel genes that do not have homology with the reference may not be annotated. This is problematic when considering the reason for sequencing a specific strain is often to identify how this strain is genetically different to its close relatives. Such problems in automated genome annotations can be improved by manual curation; including more accurately predicting the start of coding sequences and providing putative protein functions to coding sequences that are automatically annotated as hypothetical proteins. However, manual curation still requires considerable input from a genome analyst, the major limitation of which is time.

This chapter reports the genome sequencing of *Cp* 42/02-A, an Australian ovine isolate and *Cp* 1/06-A, a North American equine isolate, as well as the *C. ulcerans* control strain NCTC 12077. Annotation of these *Corynebacterium* genomes was automated using free web-based software xBASE (Chaudhuri *et al.*, 2008). A basic comparative analysis was also conducted, and this is followed up with more extensive comparative phylogenetic analyses in Chapter Five.

4.2 Results

4.2.1 Genome sequencing

Corynebacterium pseudotuberculosis strains 1/06-A and 42/02-A, and *C. ulcerans strain* NCTC 12077 were sequenced using both 454 and Illumina next generation sequencing (NGS) by The Genepool (The University of Edinburgh, UK). The 454 sequencing produced approximately 94,000, 113,000 and 138,000 reads for each sample respectively, while the Illumina technique yielded between 3.5 and 4.4 million reads per sample.

The assembly of sequence reads was conducted by The Genepool using the Newbler (Roche) assembler for 454 reads and Velvet (Zerbino & Birney, 2008) for Illumina reads, followed by a hybrid assembly of the 454 and Illumina data conducted for each *Cp* strain using the Minimus2 (Sommer *et al.,* 2007) assembly platform. Hybrid assemblies yielded between 8 and 18 contigs for each of the genomes corresponding to a sequence length of approximately 2.3 to 2.6 Mb (**Table 4.1**).

4.2.2 Gene prediction and genome annotation

Genome annotation was conducted on incomplete sequences, however in a following collaboration *Cp* 42/02-A and 1/06-A have been completed and the sequences submitted to the NCBI database under Genbank accession numbers CP003062 and CP003082 respectively (**Appendix Four**, Pethick *et al.,* 2012a; Pethick *et al.*, 2012b). The draft sequence of *C. ulcerans* NCTC 12077 remains incomplete, but has also been submitted to NCBI (BioProject PRJNA179437).

Automated gene prediction and annotation was performed using the web-based xBASE (Chaudhuri *et al.*, 2008) bacterial genome annotation service. The number of CDSs and tRNAs predicted for each of the genomes can be seen in **Table 4.1**. The number of rRNAs present in each of the sequences could not be confirmed as similarly to the *Cp* 3/99-5 genome discussed in Chapter Three, the *rrn* operon was only represented once in each of the assembled sequences.

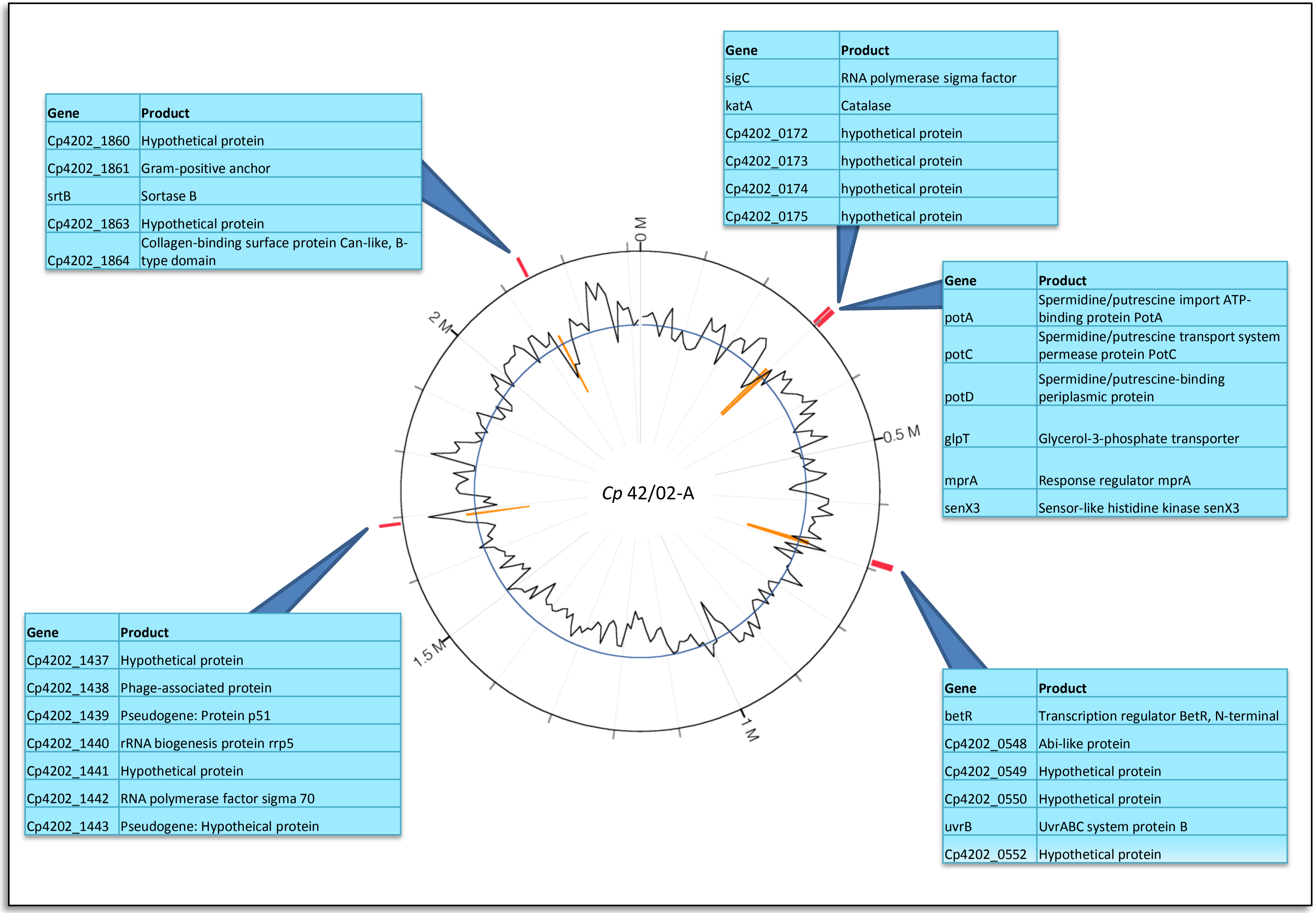
Table 4.1: Sequence data of the *Corynebacterium* genomes

The two *Cp* isolates have a similar sized genome with a similar number of predicted coding sequences, the *C. ulcerans* sequence however is somewhat larger and so contains more coding sequences. (Data is based on the incomplete sequences prior to any genome closure performed by our collaborators).

Bacterial isolate	***Cp* 42/02-A**	***Cp* 1/06-A**	***C. ulcerans* NCTC 12077**
Size of genome/ bp	2,319,594	2,365,069	2,616,855
No. contigs	8	13	18
G+C content/ %	52.18	52.08	53.38
CDSs	2115	2188	2434
Transfer RNAs	49	55	53

4.2.3 Genomic Island Prediction

IslandViewer software was used to determine the presence of genomic islands (GIs) in the genome sequences. The software runs several prediction programs to infer locations of GIs and the predictions made by the software are represented visually in **Figures 4.1-4.3**. The equine isolate, 1/06-A, was predicted to contain the least GIs with two (**Fig. 4.2**). In contrast five and four GIs were identified in the *Cp* 42/02-A (**Fig. 4.1**) and *C. ulcerans* NCTC 12077 (**Fig. 4.3**) respectively. All of the GIs flagged in *Cp* 42/02-A were also identified in *Cp* 3/99-5 (Chapter Three).



Gene	Product
sigC	RNA polymerase sigma factor
katA	Catalase
Cp4202_0172	hypothetical protein
Cp4202_0173	hypothetical protein
Cp4202_0174	hypothetical protein
Cp4202_0175	hypothetical protein

Gene	Product
potA	Spermidine/putrescine import ATP-binding protein PotA
potC	Spermidine/putrescine transport system permease protein PotC
potD	Spermidine/putrescine-binding periplasmic protein
glpT	Glycerol-3-phosphate transporter
mprA	Response regulator mprA
senX3	Sensor-like histidine kinase senX3

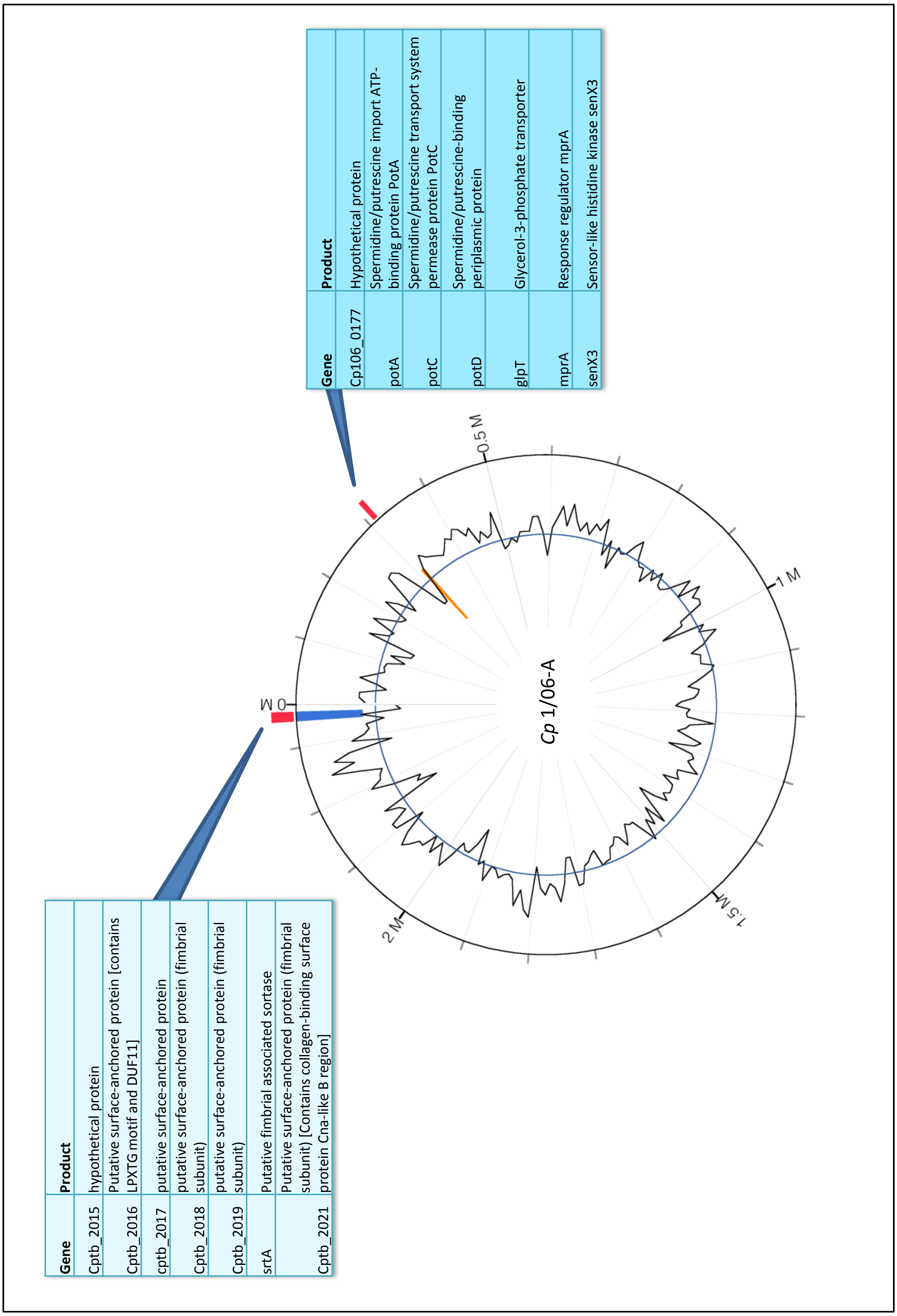
Gene	Product
betR	Transcription regulator BetR, N-terminal
Cp4202_0548	Abi-like protein
Cp4202_0549	Hypothetical protein
Cp4202_0550	Hypothetical protein
uvrB	UvrABC system protein B
Cp4202_0552	Hypothetical protein

Gene	Product
Cp4202_1437	Hypothetical protein
Cp4202_1438	Phage-associated protein
Cp4202_1439	Pseudogene: Protein p51
Cp4202_1440	rRNA biogenesis protein rrp5
Cp4202_1441	Hypothetical protein
Cp4202_1442	RNA polymerase factor sigma 70
Cp4202_1443	Pseudogene: Hypotheical protein

Gene	Product
Cp4202_1860	Hypothetical protein
Cp4202_1861	Gram-positive anchor
srtB	Sortase B
Cp4202_1863	Hypothetical protein
Cp4202_1864	Collagen-binding surface protein Can-like, B-type domain

Figure 4.1: Predicted GIs in *Cp* 42/02-A

GIs were predicted using IslandViewer software and the GI map produced is shown here with the genes in each GI added. The G+C content of the genome is represented by the graph on the first track.

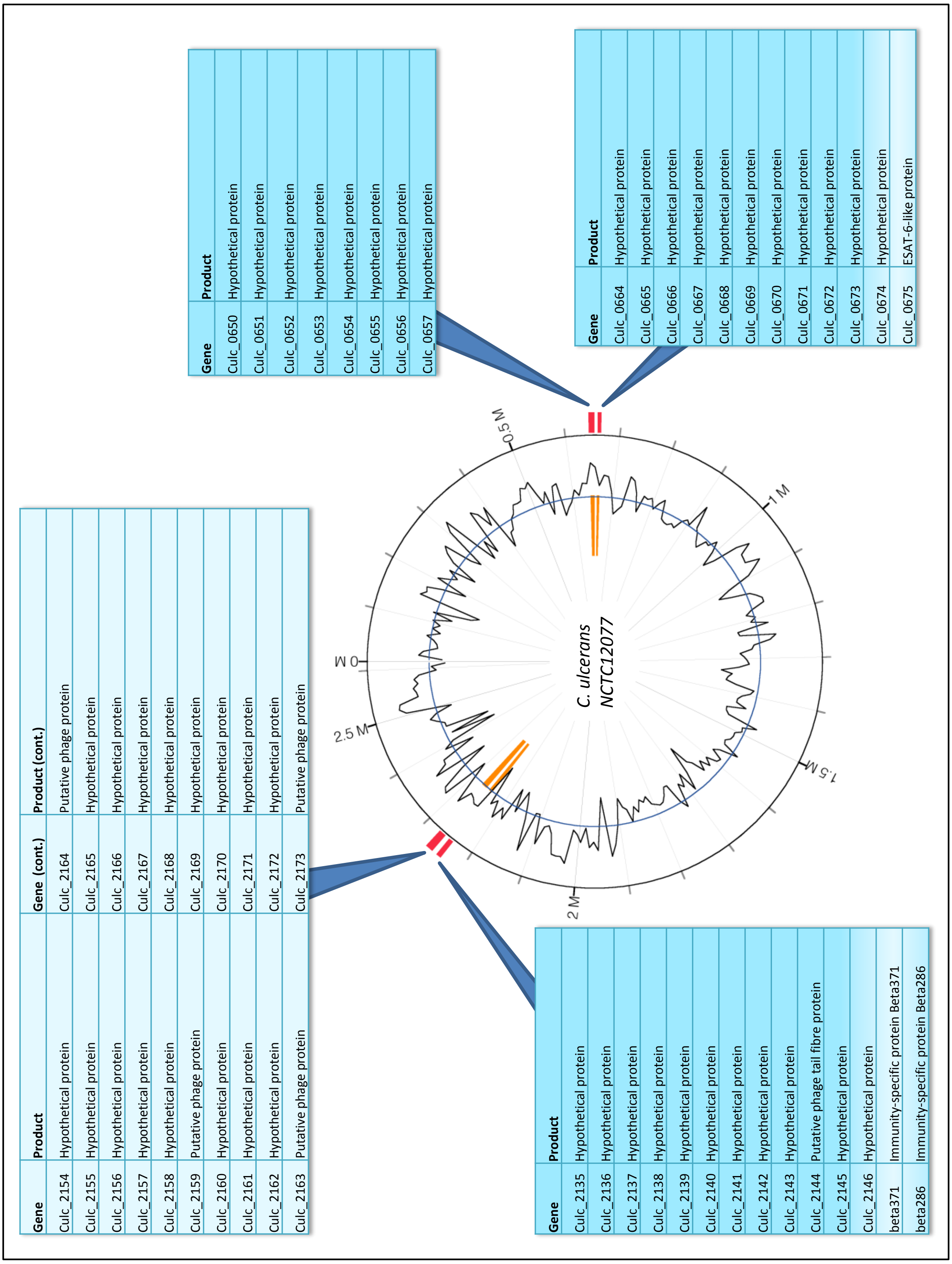


Gene	Product
Cp106_0177	Hypothetical protein
potA	Spermidine/putrescine import ATP-binding protein PotA
potC	Spermidine/putrescine transport system permease protein PotC
potD	Spermidine/putrescine-binding periplasmic protein
glpT	Glycerol-3-phosphate transporter
mprA	Response regulator mprA
senX3	Sensor-like histidine kinase senX3

Gene	Product
Cptb_2015	hypothetical protein
Cptb_2016	Putative surface-anchored protein [contains LPXTG motif and DUF11]
cptb_2017	putative surface-anchored protein
Cptb_2018	putative surface-anchored protein (fimbrial subunit)
Cptb_2019	putative surface-anchored protein (fimbrial subunit)
srtA	Putative fimbrial associated sortase
Cptb_2021	Putative surface-anchored protein (fimbrial subunit) [Contains collagen-binding surface protein Cna-like B region]

Figure 4.2: Predicted GIs in *Cp* 1/06-A

Locations of GIs in the *Cp* 1/06-A genome were predicted using IslandViewer; this isolate has considerably less GIs than the other *Corynebacterium* sequenced in this thesis. Again the G+C content of the genome is represented by the graph.



Gene	Product
Culc_0650	Hypothetical protein
Culc_0651	Hypothetical protein
Culc_0652	Hypothetical protein
Culc_0653	Hypothetical protein
Culc_0654	Hypothetical protein
Culc_0655	Hypothetical protein
Culc_0656	Hypothetical protein
Culc_0657	Hypothetical protein

Gene	Product
Culc_0664	Hypothetical protein
Culc_0665	Hypothetical protein
Culc_0666	Hypothetical protein
Culc_0667	Hypothetical protein
Culc_0668	Hypothetical protein
Culc_0669	Hypothetical protein
Culc_0670	Hypothetical protein
Culc_0671	Hypothetical protein
Culc_0672	Hypothetical protein
Culc_0673	Hypothetical protein
Culc_0674	Hypothetical protein
Culc_0675	ESAT-6-like protein

Gene	Product	Gene (cont.)	Product (cont.)
Culc_2154	Hypothetical protein	Culc_2164	Putative phage protein
Culc_2155	Hypothetical protein	Culc_2165	Hypothetical protein
Culc_2156	Hypothetical protein	Culc_2166	Hypothetical protein
Culc_2157	Hypothetical protein	Culc_2167	Hypothetical protein
Culc_2158	Hypothetical protein	Culc_2168	Hypothetical protein
Culc_2159	Putative phage protein	Culc_2169	Hypothetical protein
Culc_2160	Hypothetical protein	Culc_2170	Hypothetical protein
Culc_2161	Hypothetical protein	Culc_2171	Hypothetical protein
Culc_2162	Hypothetical protein	Culc_2172	Hypothetical protein
Culc_2163	Putative phage protein	Culc_2173	Putative phage protein

Gene	Product
Culc_2135	Hypothetical protein
Culc_2136	Hypothetical protein
Culc_2137	Hypothetical protein
Culc_2138	Hypothetical protein
Culc_2139	Hypothetical protein
Culc_2140	Hypothetical protein
Culc_2141	Hypothetical protein
Culc_2142	Hypothetical protein
Culc_2143	Hypothetical protein
Culc_2144	Putative phage tail fibre protein
Culc_2145	Hypothetical protein
Culc_2146	Hypothetical protein
beta371	Immunity-specific protein Beta371
beta286	Immunity-specific protein Beta286

Figure 4.3: Predicted GIs in *C. ulcerans* NCTC 12077

The presence and location of GIs in the *C. ulcerans* NCTC 12077 genome were predicted by IslandViewer software. The G+C content is shown on the first track of the map, and the predicted GIs are in areas of higher than average G+C content.

4.2.4 General genome comparisons

4.2.4.1 Genome comparison using BLAST Ring Image Generator (BRIG)

The *C. ulcerans* and *Cp* genomes sequenced during this project as well as the publically available *Cp* FRC41 and *C. diphtheriae* NCTC 13129 were compared using BLAST Ring Image Generator (BRIG) software. Two comparisons were made; using *Cp* FRC41 or *C. diphtheriae* NCTC 13129 as a reference, and the visual outputs of the program are shown in **Fig. 4.4** and **Fig. 4.5** respectively. The comparison using *Cp* FRC41 as a reference clearly shows the *Cp* sequences are extremely similar, especially *Cp* 3/99-5, 42/02-A and FRC41. However, there are three main regions of sequence which set these *Cp* isolates apart from the equine strain 1/06-A. The first of these is at approximately 600 kb of the *Cp* FRC41 genome, and contains genes encoding hypothetical proteins. The second region at ~1,050 kb also encodes hypothetical proteins but in addition contains an integrase with some homology to a *Mycobacterium* phage protein and a recombinase of the phage integrase family in *Enterococcus faecalis*. The final of these regions at ~1,600 kb contains many hypothetical proteins and a phage associated protein, as well as an rRNA biogenesis protein, an RNA polymerase factor and a DNA methylase. All three of these regions are also absent in the *C. ulcerans* and *C. diphtheriae* genomes.

The second comparison gives an indication of the extra sequence that *C. diphtheriae* possesses which *Cp* does not, of which there are many small regions. Interestingly there is a region at ~150 kb of the *C. diphtheriae* genome which is largely conserved in *C. ulcerans* but not in any of the *Cp* isolates (**Fig. 4.5**). This region in *C. diphtheriae* consists largely of genes encoding hypothetical proteins, but also genes encoding beta immunity-specific proteins and the *tox* gene encoding diphtheria toxin, as well as genes encoding phage prohead protease, integrase, capsid and tail fibre proteins. The *C. ulcerans* genome also contains the phage protein and beta immunity genes, however, it does not possess the *tox* gene and some genes encoding hypothetical proteins towards the end of this *C. diphtheriae* region.

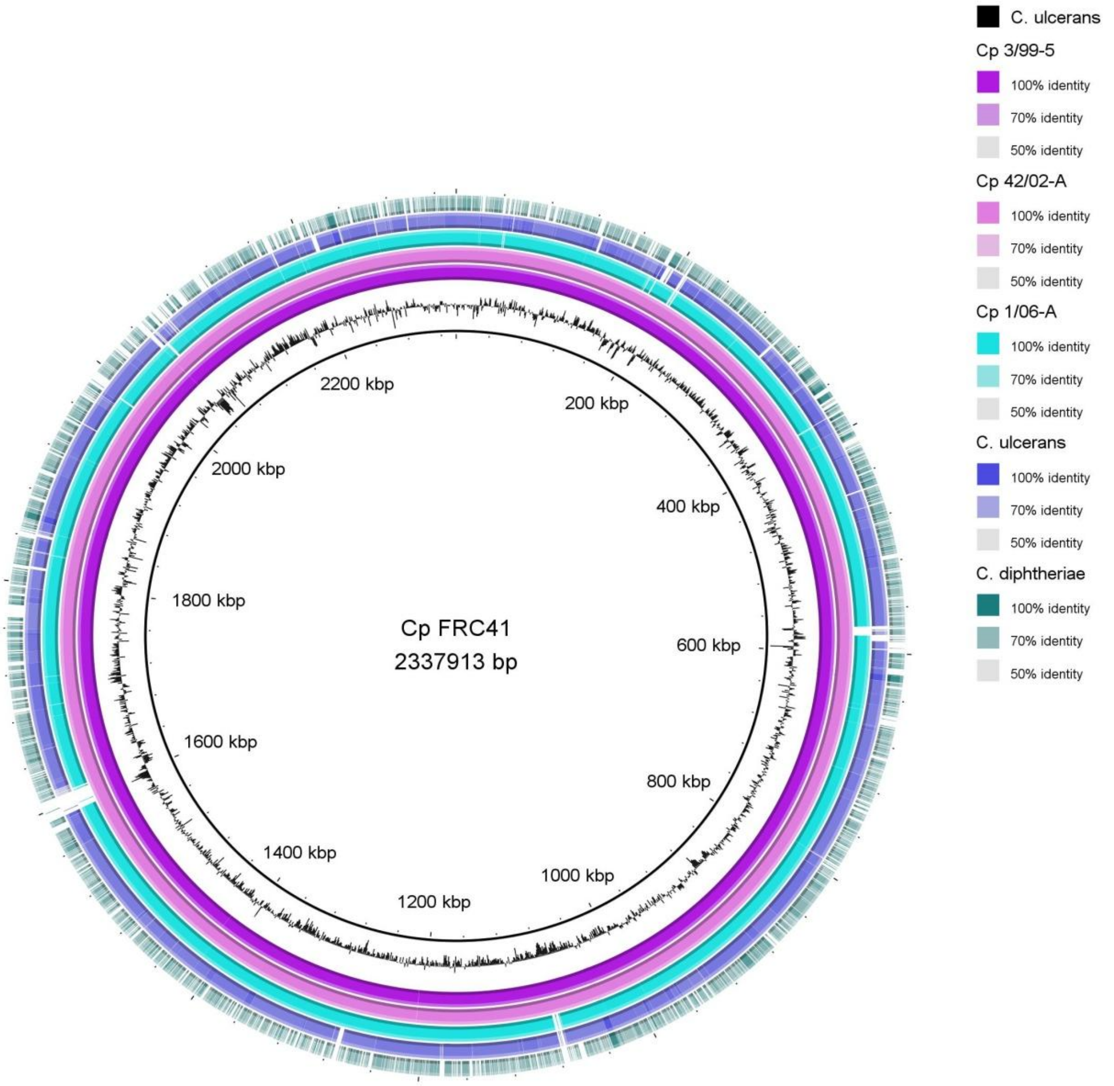


Figure 4.4: Genomic map of the *Corynebacterium* genome comparison

BLAST Ring Image Generator (BRIG) was used to compare *C. diphtheriae* and the *Corynebacterium* genomes sequenced in this thesis using *Cp* FRC41 as a reference. As shown in the key, darker shading represents more conserved sequence whereas lighter areas indicate more distinct sequence. Gaps in colour represent gaps in homology between genome sequences.

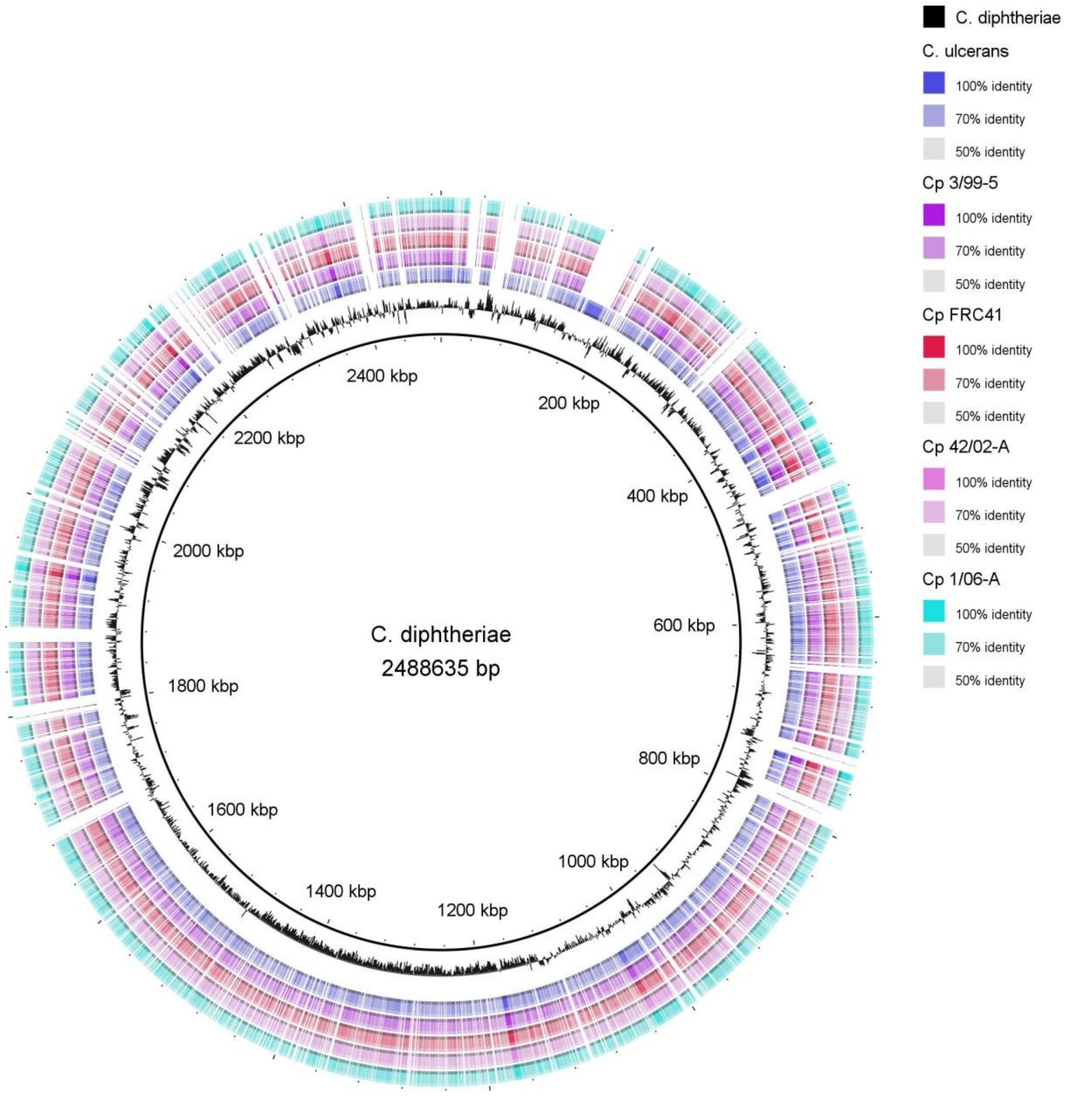


Figure 4.5: Genomic map of the *Corynebacterium* genomes with *C. diphtheriae* as a reference

BLAST Ring Image Generator (BRIG) was again used to compare the diphtheria group of sequences but using *C. diphtheriae* NCTC 13129 as a reference. As shown in the key, darker shading represents more conserved sequence whereas lighter areas indicate more distinct sequence. Gaps in colour represent gaps in homology between genome sequences.

4.2.4.2 Genome content comparison using BLAST

Corynebacterium genome sequences were compared using locally installed BLASTx to determine the amount of gene conservation between isolates. Genes were considered common between two genomes if they were a reciprocal best hit (otherwise known as a bidirectional best hit). The number of genes common or unique to genomes representing the diphtheria group (*Cp* 3/99-5, *C. ulcerans* NCTC 12077 and *C. diphtheriae* NCTC 13129) are shown graphically in a Venn diagram (**Fig. 4.6**). These genomes all share 1,660 genes, but also have 199-660 unique genes (i.e. approximately 10-20% of the genes in each genome are unique). Genome content of the *Cp* isolates was also investigated and genes common or unique to each of the *Cp* 3/99-5, 42/02-A, 1/06-A and FRC41 sequences are shown in **Fig. 4.7**. All of the *Cp* isolates share 1,860 genes, and *Cp* 1/06-A has the greatest number of unique genes with 217, whereas *Cp* 3/99-5, 42/02-A and FRC41 have 90, 84 and 40 respectively.

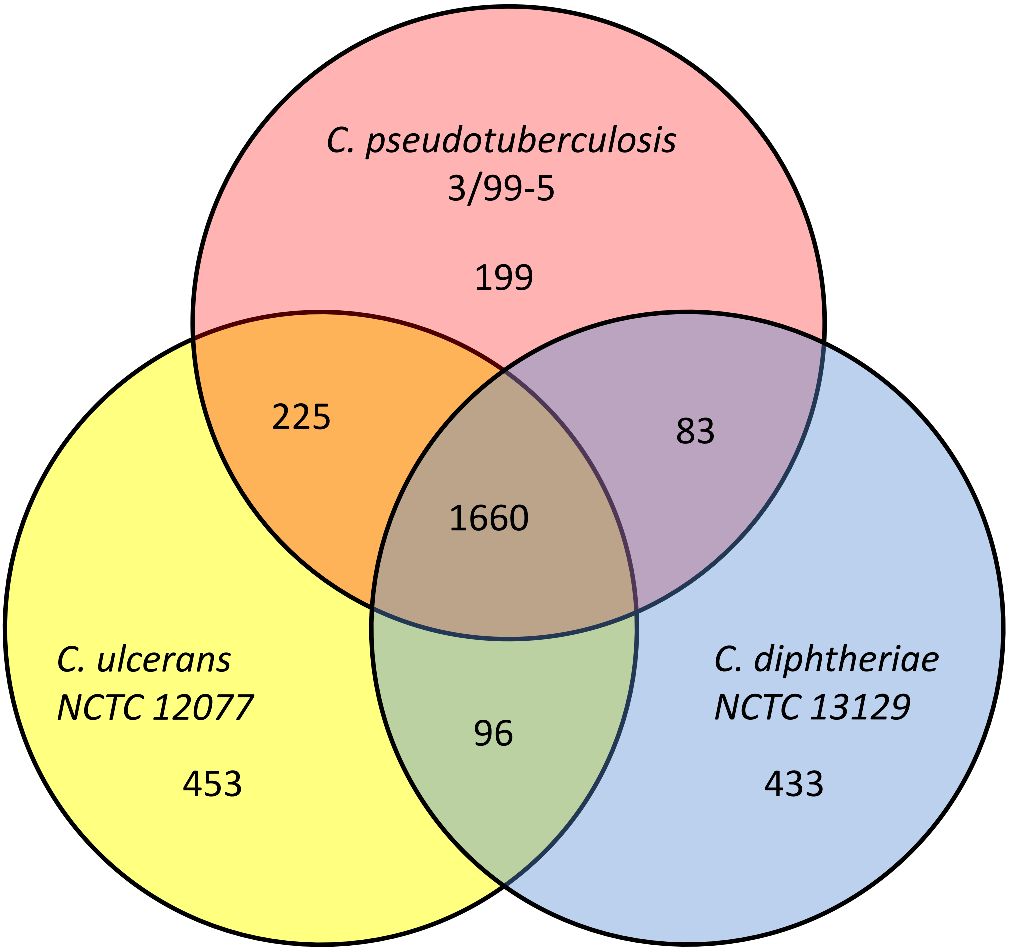


Figure 4.6: Genome content comparison of the diphtheria group

Genes common or unique based on reciprocal best hits, to *Cp* 3/99-5, *C. ulcerans* NCTC 12077 and the publically available *C. diphtheriae* NCTC 13129 were determined from BLASTx comparisons of the genome sequences.

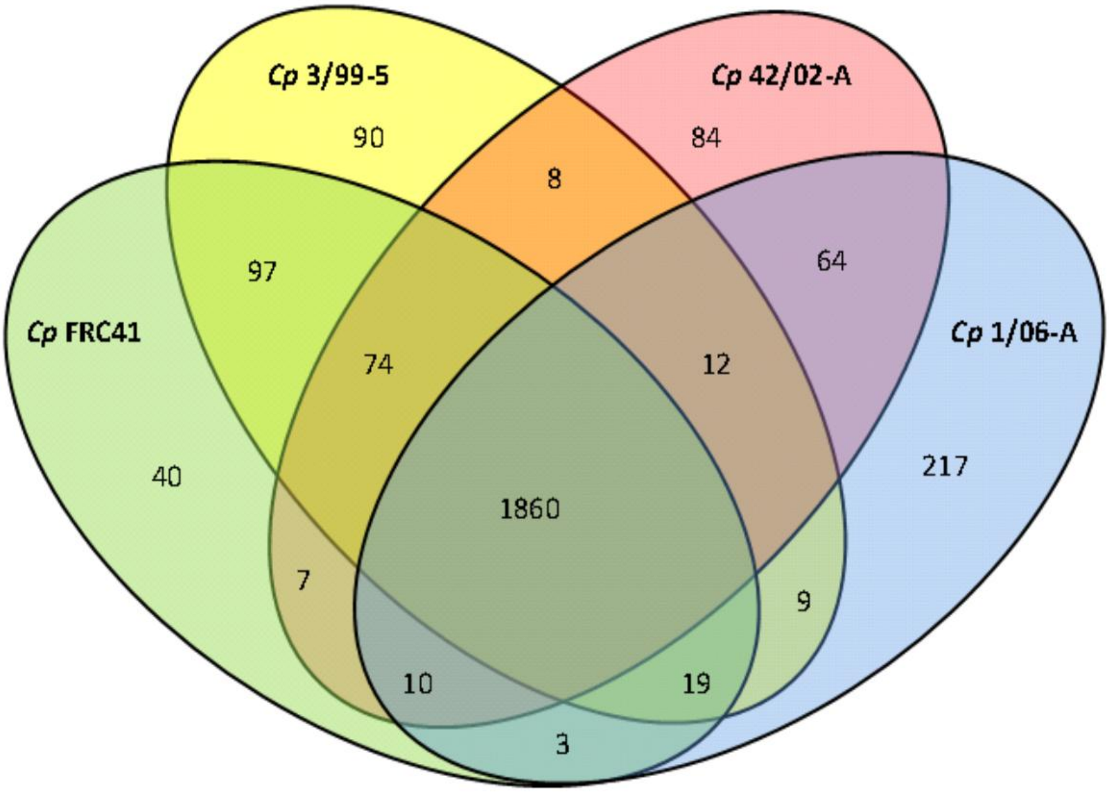


Figure 4.7: ***Corynebacterium pseudotuberculosis*** **genome content comparisons**

Venn diagram of genes unique or shared between *Cp* FRC41, 3/99-5, 42/02-A, 1/06-A. Genes were determined to be common or not based on reciprocal best hits from BLASTx comparison of genome sequences.

4.3 Discussion

Sequencing of the *C. ulcerans* NCTC 12077, and *Cp* 42/02-A and 1/06-A genomes was of a similarly high standard to that of *Cp* 3/99-5 in the previous chapter. In each case assembly resulted in a low number of contigs, the number of which increased with genome size. However, the additional Illumina sequencing of these genomes did not yield fewer contigs than the 454 sequencing alone for *Cp* 3/99-5, indicating that the greater coverage depth has not helped assemble what are clearly problematic regions such as the *rrn* operons. The number of *rrn* operons remained undetermined as in these incomplete sequences only one copy was present in each genome sequence. However this is almost certainly an under representation, and it is likely that the single operon is actually an amalgamation of reads from several loci as was discovered the case for *Cp* 3/99-5. Whilst this could not be proven without further work or genome closure, on comparison the *Cp* 42/02-A sequence had the same number of contigs as *Cp* 3/99-5 and the gaps were in the same locations. These gaps in *Cp* 3/99-5 include four *rrn* operon loci and it is highly likely that the equivalent gaps in *Cp* 42/02-A are also loci of *rrn* operons. Furthermore, these predicted *rrn* operon loci were confirmed during subsequent genome closure by our collaborators.

General characteristics of the genomes are similar, especially the *Cp* isolates but also, although to a slightly lesser extent, *C. ulcerans*. The number of tRNAs ranged from 49 to 55, which is equivalent to other sequenced *Corynebacterium* (Cerdano-Tarraga *et al.*, 2003; Silva *et al.*, 2011; Yukawa *et al.*, 2007). The G+C content of *Cp* 42/02-A is the same as *Cp* 3/99-5, and that of the equine isolate is also approximately 52%. However, the *C. ulcerans* genome has a marginally higher G+C content of 53.38%. The number of coding sequences for all of the *Cp* isolates is approximately 2,150, although the *C. ulcerans* genome contains more (2,434) reflecting the genome size of this organism being approximately 0.3 Mb larger.

Genomic islands were predicted in the genomes using IslandViewer software as in the previous chapter. The amount of GIs identified ranged from two in the equine *Cp* 1/06-A isolate to five in the ovine *Cp* 42/02-A isolate. The *Cp* 42/02-A genome has extremely high homology with the previously sequenced *Cp* 3/99-5 and the GIs predicted in this sequence were also identified in 3/99-5 and discussed in Chapter Three. It is likely that in the lineage of this species these regions of sequence were at some point acquired by horizontal transfer and that these two strains share a common ancestor which had acquired these regions. Often GIs are associated with virulence and hence known as pathogenicity islands, and if this was the case for all of the GIs predicted, one could be forgiven for thinking the ovine *Cp* strains are more virulent. However, further investigation, such as the identification of genes in the GIs as virulence factors and confirming the isolates pathogenicity, is necessary to draw conclusions regarding this. Of the two GIs identified in *Cp* 1/06-A, at least one may be linked to virulence as it contains genes encoding a sortase and sortase substrate proteins which are associated with fimbriae. Sortases and their substrates have important roles, including adherence, in the virulence of many Gram-positive bacteria (Popp and Ploegh, 2011). Furthermore, Ton-That and Schneewind (2003) showed that Sortase and surface proteins were necessary for pilus formation in *C. diphtheriae.* The other GI identified in *Cp* 1/06-A contains genes encoding proteins involved in the transport of spermidine and putrescine. This region was also flagged as a GI in the other *Cp* strains and is necessary for many biological processes.

C. ulcerans has four GIs, however extremely little is known about what these regions encode. The genes in these GIs predominantly encode hypothetical proteins; however, two of the GIs also have putative phage proteins encoded in them. It is likely that these regions have been acquired by horizontal transfer from phage. Another of the GIs contains an ESAT-6-like protein; ESAT-6 (Early Secreted Antigen Target 6 kDa) is a secreted antigen of *Mycobacterium tuberculosis*. It has been described that ESAT-6 and related proteins require a dedicated secretory apparatus which is encoded by a cluster of genes (Brodin *et al.*, 2004). Perhaps the remaining proteins in this GI (which are all annotated as hypothetical proteins) are responsible for the secretion of this ESAT-6-like protein. Importantly, ESAT-6 is a potent T-cell antigen and has been suggested as a virulence factor (Brodin *et al.*, 2004), indicating this GI as a possible pathogenicity island.

The comparisons of *Corynebacterium* genomes using BRIG indicates that *C. ulcerans* shares much more homology with *Cp* genomes than *C. diphtheriae* does. This is further supported by the amount of genes shared between isolates as depicted in the Venn diagram for the diphtheria group. In addition the *C. ulcerans* genome appears more similar to the equine *Cp* isolate than the other *Cp* strains. The BRIG analysis using *Cp* FRC41 as a reference revealed three notable regions absent from *C. ulcerans, C. diphtheriae* and the equine *Cp* strain but present in all other *Cp* isolates. Two of these regions contain only hypothetical proteins with the exception of one gene in the second of these regions which has been annotated with a predicted function, it encodes an integrase. This protein has homology with a *Mycobacterium* phage protein and also a recombinase of the phage integrase family in *Enterococcus faecalis*. Integrases are enzymes responsible for enabling genetic material from one organism to be integrated into the chromosome of a host organism (Groth and Calos, 2004). Hence this particular region of sequence may be an insert from a phage and it is likely the phage DNA was integrated into the chromosome of a common ancestor of the biovar *ovis Cp* isolates. It is probable that the other proteins in this region (and perhaps the first region) are annotated as hypothetical as they belong to families of phage proteins with little homology to proteins in public databases, and about which nothing is known with function still remaining unidentified. The third and largest section of sequence present in biovar *ovis Cp* isolates but missing from the other *Corynebacterium* also results from a phage. This region again contains many hypothetical proteins but also contains a phage-associated protein and significantly was flagged by IslandViewer as a GI in both *Cp* 42/02-A but also 3/99-5 in the previous chapter.

The second BRIG analysis used *C. diphtheriae* NCTC 13129 as a reference and this comparison indicated many small regions of sequence unique to *C. diphtheriae* from the other *Corynebacterium* strains investigated. The *C. diphtheriae* genome also contains a region of approximately 50 kb that has high homology with *C. ulcerans*, over almost the entire region, but that is absent from all of the *Cp* strains. This region of *C. diphtheriae* contains a prophage carrying the *tox* gene encoding diphtheria toxin, interestingly the *C. ulcerans* genome lacks the *tox* gene but does contain the majority of other genes from this region (including beta immunity and phage protein genes). This indicates that *C. ulcerans* has a very similar prophage at the same site to that in *C. diphtheriae*.

The genome content comparison of organisms (*Cp* 3/99-5, *C. ulcerans* NCTC 12077 and *C. diphtheriae* NCTC 13129) representing the diphtheria group shows that the majority of genes are common to all three genomes. However, all of the genomes also contain a considerable number of unique genes; *Cp* 3/99-5 contains the least with 199. Some of these unique genes will be responsible for the organisms' ability to cause different diseases in different host species. *Corynebacterium pseudotuberculosis* 3/99-5 and *C. ulcerans* NCTC 12077 share more genes than either of these genomes do with *C. diphtheriae* NCTC 13129.

This is reflective of a closer phylogenetic relationship between these two organisms than either with *C. diphtheriae.* Furthermore, the two genomes with the least genes in common are *Cp* 3/99-5 and *C. diphtheriae* NCTC 13129 with 1,743 shared genes. This is indicative that these two species have the most distant relationship of the three organisms.

The Venn diagram showing the genome content of the *Cp* strains depicts the high level of homology between the isolates. The *Cp* isolates each have a small number of 'unique' genes; below 100 with the exception of *Cp* 1/06-A which has 217. The *Cp* 1/06-A isolate has the largest genome of the four isolates included and is also the only strain belonging to biovar *equi*. It is likely that some of the genes found in this organism but not the other *Cp* strains are linked to host specificity. Indeed the same could be said for some of the 74 genes shared between the other three *Cp* isolates but that are absent from this equine strain.

Analysis of the *C. ulcerans* NCTC 12077, and *Cp* 42/02-A and 1/06-A genomes revealed all the sequences to be highly homologous, with the greatest differences observed between *C. ulcerans* and the *Cp* isolates. Interestingly, analysis of genome content revealed *Cp* 1/06-A, the only strain from biovar *equi,* shares the least genes amongst the *Cp* isolates. This strain would found to be more homologous than the ovine *Cp* isolates with *C. ulcerans.* Analysis with BRIG revealed three substantially sized regions of sequence which are present in the ovine *Cp* isolates but absent from *Cp* 1/06-A, *C. ulcerans* and *C. diphtheriae.* These regions largely consist of genes encoding hypothetical proteins, although there are some genes, such as that encoding an integrase, which suggest these sequences have diverged as a result of horizontal gene transfer from phage. Further comparative analysis of the *Cp* and *C. ulcerans* genomes will gain knowledge on the evolution, but also pathogenicity and host specificity of these organisms. Indeed the genomes sequenced here, although yet to be fully exploited, provide a wealth of information for studies going forward.

Chapter Five: Phylogenetic analyses of genus *Corynebacterium*

5.1 Introduction

Phylogenetics is the study of evolutionary relationships among a group of organisms determined from the comparison of molecular sequence data. These evolutionary relationships are represented by phylogenetic trees. A phylogenetic tree is a branching diagram which depicts the evolutionary relationship among different species, strains or other entities. The branching pattern of a phylogenetic tree is termed the tree topology, with the tree tips referred to as “terminals” or “taxa” (**Fig. 5.1**). Furthermore, a portion of phylogeny which includes a common ancestor and all descendants is known as a “clade”. Within a tree the branching points are called internal nodes and represent the last common ancestor of the two lineages descending from this point (Baum, 2008). Confidence in the topology at internal nodes is indicated by either a bootstrap support value (0-100) if analysis is performed using a Maximum Likelihood method or a posterior support value (0-1) if a Bayesian analysis is performed (Felsenstein, 1985). For both bootsrap and posterior support, the level of confidence increases with the support value. Branch length between nodes represents the number of changes that occurred in the sequences prior to the next level of separation.

Most phylogenetic trees are rooted meaning that they have a single node which corresponds to the most recent common ancestor of all of the taxa in the tree. The root of a tree can be calculated using an outgroup or known common ancestor.

Before a phylogenetic analysis can be performed and a tree estimated, a multiple sequence alignment (MSA) must be generated. An MSA is an alignment of three or more biological sequences, usually DNA or protein, and can be used to infer sequence homology. Phylogenetic analysis of an MSA can be conducted to assess the common evolutionary origins of the query sequences. As aligning sequences of a biologically relevant length can be complex and time consuming to do manually, computational algorithms are used to produce alignments. There are several available tools online to create MSAs including ClustalW (Thompson, Higgins and Gibson, 1994), MAFFT (Katoh *et al.*, 2002) and T-coffee (Notredame, Higgins and Heringa, 2000).

In order to optimise the phylogenetic analysis of an MSA, model selection is routinely performed prior to analysis. In the current work, model selection was performed by the

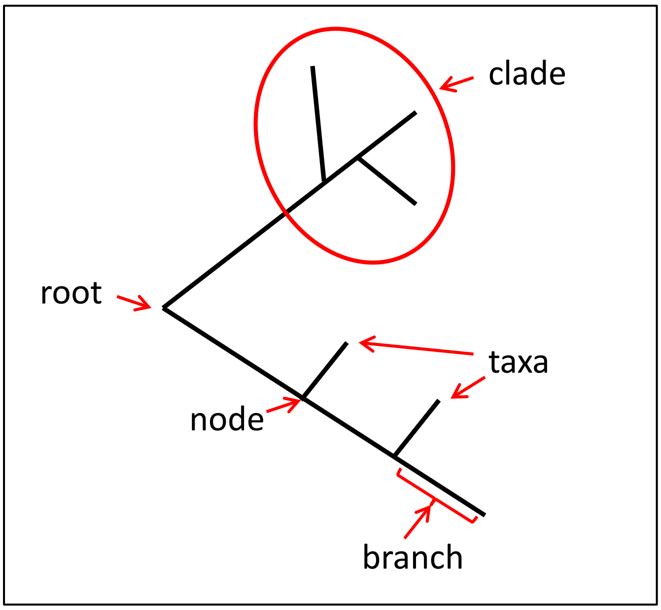


Figure 5.1: Phylogenetic terminology

The terms for describing parts of a phylogenetic tree are shown. Note taxa may also be referred to as terminals or leaves.

TOPALi software. TOPALi launches 24 nucleotide and 36 amino acid models for Mr Bayes and 56 and 40 respectively for PhyML (Milne *et al.*, 2009). The optimal model is then automatically selected for phylogenetic analyses.

Various phylogenetic methods have been described for the analyses of DNA and protein sequence alignments. The Neighbor-Joining method constructs the tree by sequentially finding pairs of neighbouring taxa connected by a single internal node (Vandamme, 2003). It is based on a distance matrix approach, where preference is given to the topology giving the shortest total branch length at each step of the algorithm (Brown, 2007). However, this methodology has been greatly superseded by those not relying on distance measures such as Maximum Parsimony or Maximum Likelihood. Maximum Parsimony decides between different tree topologies to find the one with the shortest evolutionary pathway, i.e. that requiring the smallest number of nucleotide changes from the common ancestor at the tree root to all the taxa being compared at the tree tips (Brown, 2007). Like Maximum Parsimony, Maximum Likelihood also examines all tree topologies however this method uses a probabilistic model to estimate the most likely tree topology. The algorithm calculates the probability of expecting each possible nucleotide (or amino acid) in the internal nodes, and from this infers the likeliness of the tree topology (Vandamme, 2003).
A further method, Bayesian-inference, produces phylogenetic trees in a manner closely related to the Maximum Likelihood method. Bayesian analysis takes the tree with the highest posterior probability as the best tree. This differs from Maximum Likelihood as topologies and branch lengths are treated as random variables as opposed to parameters (Rannala & Yang, 1996).

Phylogenetics has long been applied to bacteria to gain an insight into the evolutionary relationship both intra- and inter-species. Typically the sequence of the small-subunit ribosomal DNA (16S rRNA gene) is used as a phylogenetic marker in the study of bacteria (Janda and Abbott, 2007), however a considerable amount of studies using other loci have been conducted. Wu and Eisen (2008) describe ‘AMPHORA’, an automated pipeline for phylogenetic analysis and its use in constructing a genome tree of 578 bacterial species. AMPHORA employs the sequences of 31 housekeeping genes involved in either metabolism or information processing; they exist as a single copy and are universally distributed in bacteria.

More specifically Ruimy *et al.* (1995) describe the phylogeny of the *Corynebacterium* genus deduced from analyses using 16S rDNA sequences. From their work the authors deduced that several species that do not have all of the characteristics of the *Corynebacterium* genus should be included in the genus. The group determined that the presence of mycolic acids is not a requirement of the genus (as *Corynebacterium amycolatum* does not possess them), and also suggested the extension of the range of G+C content from 51-65% to 46-74%. It was concluded that analysis of the 16S rDNA sequence was likely the most straightforward method for determining if an organism belongs to the *Corynebacterium* genus (Ruimy *et al.*, 1995). Later Khamis *et al.* (2004) used *rpoB* gene sequencing to identify whether an organism belonged to the *Corynebacterium* genus. In this study, almost complete *rpoB* sequences were determined by PCR and genome walking, and phylogenetic analysis conducted using Neighbor-Joining, Maximum Parsimony and Maximum Likelihood methods. The authors found the *rpoB* sequence to be more polymorphic than that of the 16S rDNA and found a 434-452 bp *rpoB* fragment allowed accurate identification of all *Corynebacterium* species (Khamis *et al.*, 2004). The same group found that on comparison higher proportions of corynebacterial isolates were identified by partial *rpoB* gene sequence analysis than by analysis of the 16S rRNA gene (Khamis *et al.*, 2005).

Here a similar methodology to that of Wu and Eisen (2008) described above is applied to *Corynebacterium* species to aid an updated phylogenetic analysis of this genus. The 31 phylogenetic marker genes used by Wu and Eisen (2008) were accompanied by an additional two genes *hsp65* and *cpn60*. The sequences of these 33 loci were included from 17 *Corynebacterium* genomes and a further 18 related genomes from the CMNR group, which were used as an outgroup. A phylogenetic tree was estimated using a Bayesian codon position model.

Previously a comparison of *Cp* isolates showed a high level of homology between strains, regardless of host species or origin of the isolate (Connor *et al.*, 2007). In the current study, a further analysis was performed using 32 gene loci from the *Cp* 3/99-5 genome which were predicted to encode secreted proteins. This analysis aimed to further investigate the evolutionary relationship and distinguish between the different *Cp* isolates from different hosts and/or geographical locations.

5.2 Results

Phylogenetic analyses were conducted upon the four genome sequences attained during this project in addition to all those *Corynebacterial* genomes available on the NCBI database at the time the work commenced (genomes used are shown in Chapter Two).

Phylogenetic trees were constructed to visualise evolution between species of the *Corynebacterium* genus. Trees were estimated in TOPALi v.2.5 (Milne *et al.*, 2009) using either a maximum likelihood method implemented by PhyML (Guindon & Gascuel, 2003) or the Bayesian inference approach of MrBayes (Huelsenbeck & Ronquist, 2001).

5.2.1 Phylogenetic analyses of genus *Corynebacterium*

5.2.1.1 Preliminary analyses using a single locus

Preliminary analyses were performed using *pbpB* (Cp3995_0202) of the *Cp* 3/99-5 genome to allow an informed decision of method to use for subsequent analyses using several gene loci. The sequence of this gene locus was analysed in the *Corynebacterium* genus.

PhyML was used to analyse both a protein alignment and a nucleotide alignment of the locus and the resulting phylogenetic trees can be seen in **Fig. 5.2**. The nucleotide alignment was treated as simple DNA (so that all codon positions were treated the same), PhyML does not have the option of analysing a nucleotide alignment using codon positioning.
A Bayesian inference of phylogeny, using MrBayes, was performed upon nucleotide and protein alignments as for PhyML; however a second nucleotide analysis was also done assuming the DNA to be protein coding using a codon position model (**Fig. 5.2**).
These five analyses were then repeated using alignments that had been manually trimmed (**Fig. 5.3**) or trimmed using Gblocks software on either default settings (**Fig. 5.4**) or the least stringent settings possible (**Fig. 5.5**).

The trees produced from these analyses all have the same topology it is only the level of support that differs between them. For ease of comparison a table of supporting values for each node, based on the consensus tree (**Fig. 5.6**), is provided (**Table 5.1**). Generally the bootstrap values (PhyML) or posterior support values (MrBayes) supporting group branchings are high, indicating significant information within the data to have confidence that clades are distinct. Analysis using MrBayes produced trees of a higher resolution than those produced from analyses using PhyML. Generally trees were more resolved when the sequence alignments were trimmed prior to analysis. Also trees deduced from protein alignments have a lower resolution than those constructed from nucleotide alignments, and in most cases the trees estimated from Bayesian analysis treating a nucleotide alignment using a codon position (CP) model are less resolved than when the same analysis is performed treating the alignment as simple DNA.

The most resolved tree resulted from three different analyses and is completely resolved with the exception of Node 7 (**Fig. 5.6**) which has a posterior support of 0.94.

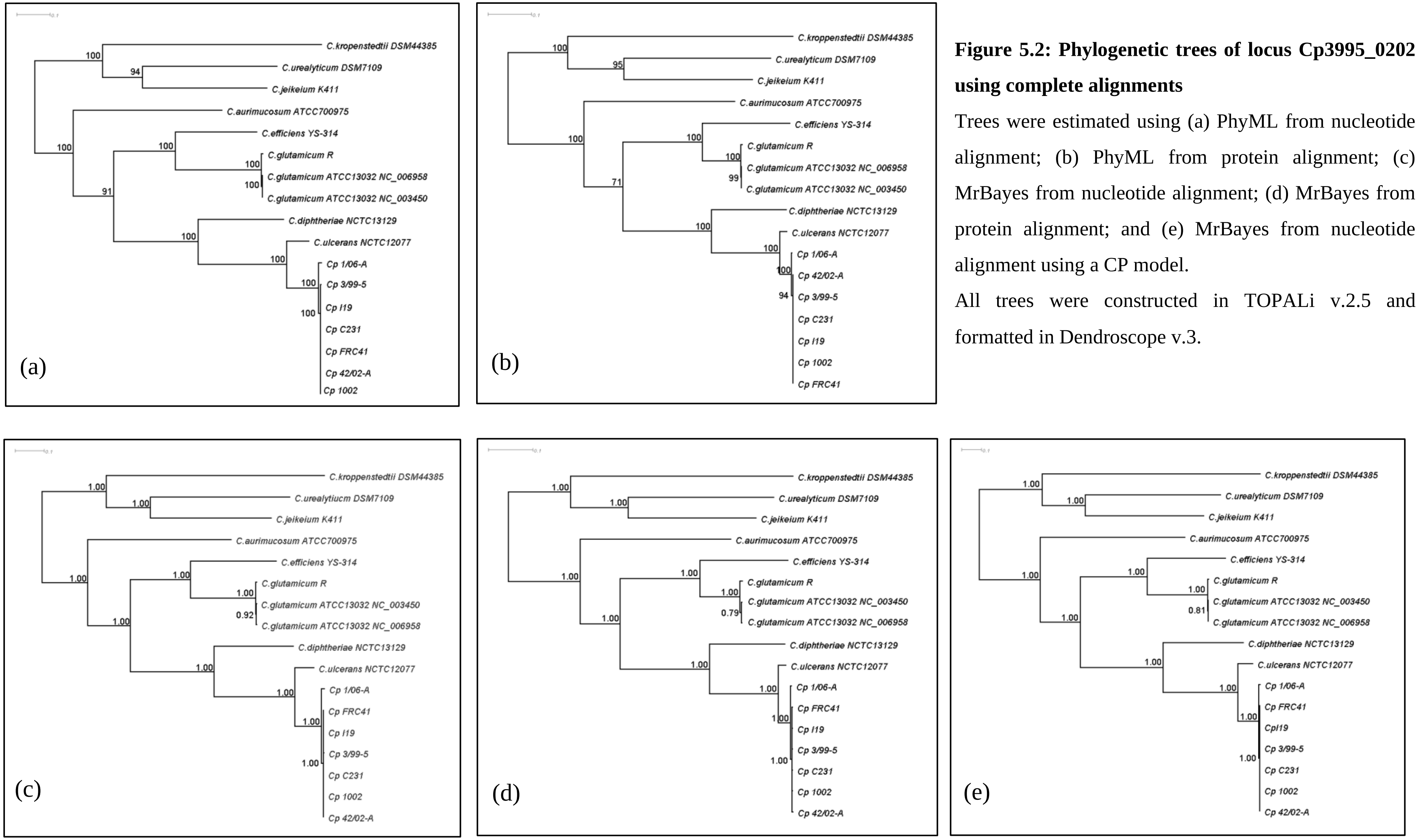


Figure 5.2: Phylogenetic trees of locus Cp3995_0202 using complete alignments

Trees were estimated using (a) PhyML from nucleotide alignment; (b) PhyML from protein alignment; (c) MrBayes from nucleotide alignment; (d) MrBayes from protein alignment; and (e) MrBayes from nucleotide alignment using a CP model.

All trees were constructed in TOPALi v.2.5 and formatted in Dendroscope v.3.

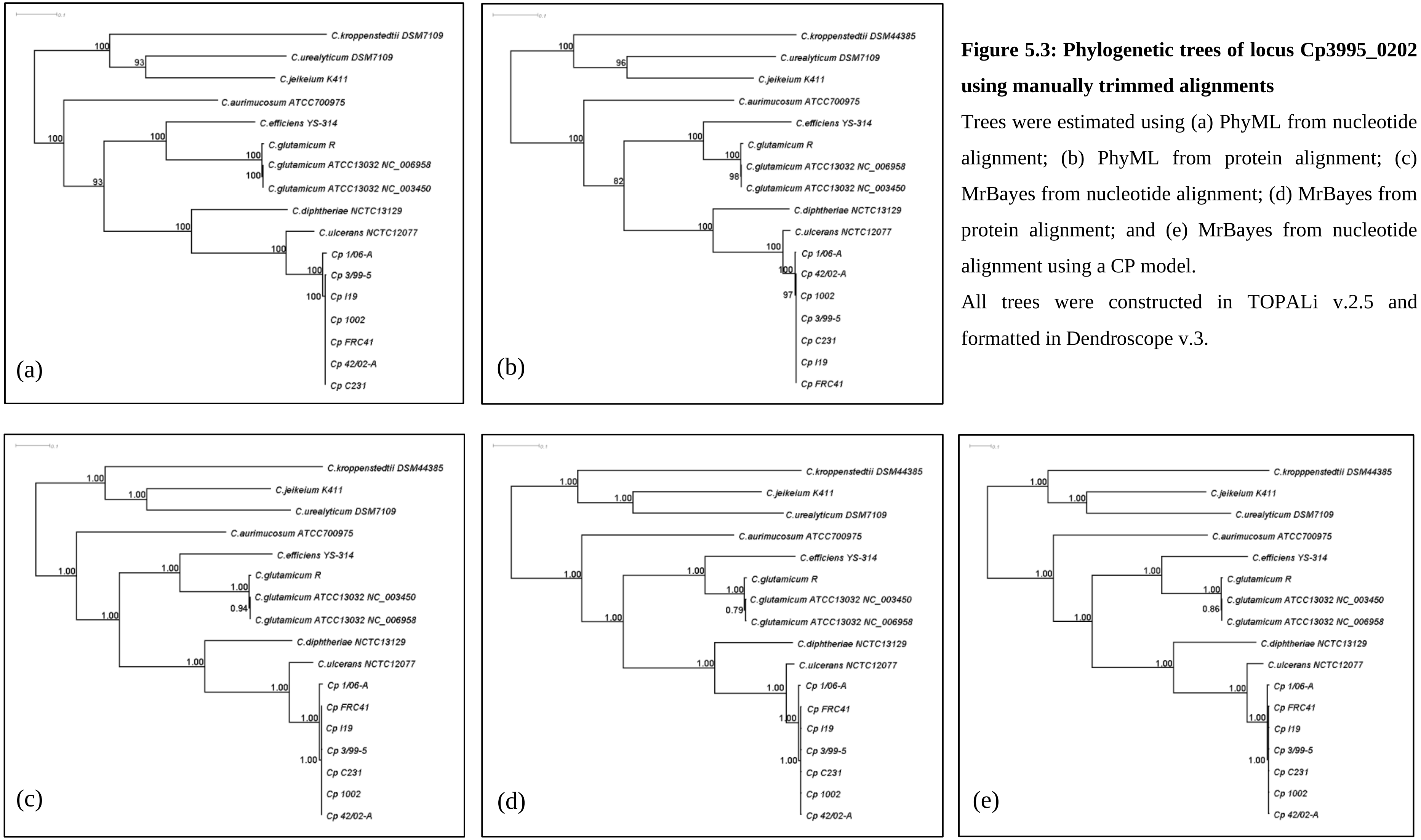


Figure 5.3: Phylogenetic trees of locus Cp3995_0202 using manually trimmed alignments

Trees were estimated using (a) PhyML from nucleotide alignment; (b) PhyML from protein alignment; (c) MrBayes from nucleotide alignment; (d) MrBayes from protein alignment; and (e) MrBayes from nucleotide alignment using a CP model.

All trees were constructed in TOPALi v.2.5 and formatted in Dendroscope v.3.

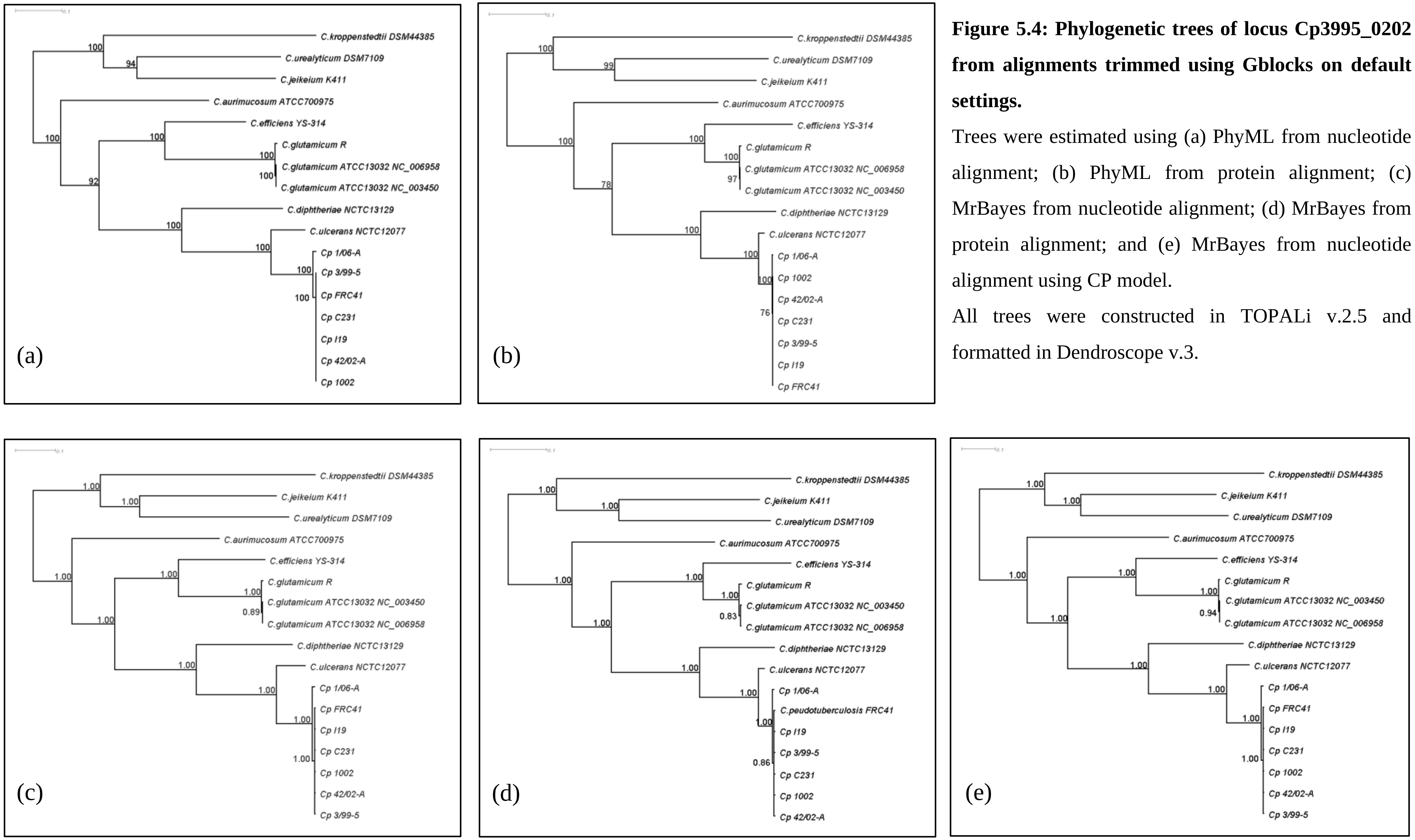


Figure 5.4: Phylogenetic trees of locus Cp3995_0202 from alignments trimmed using Gblocks on default settings.

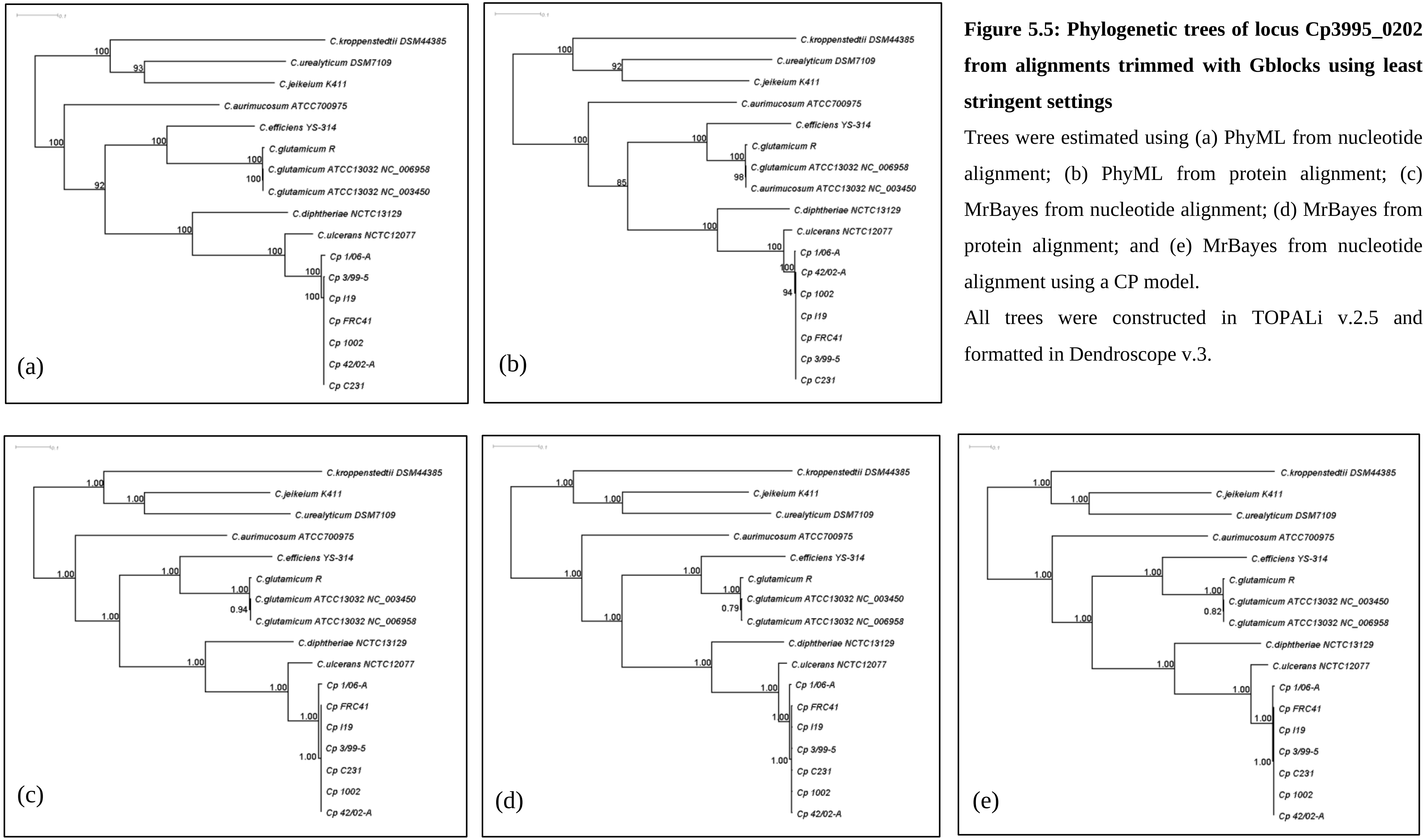
Trees were estimated using (a) PhyML from nucleotide alignment; (b) PhyML from protein alignment; (c) MrBayes from nucleotide alignment; (d) MrBayes from protein alignment; and (e) MrBayes from nucleotide alignment using CP model.

All trees were constructed in TOPALi v.2.5 and formatted in Dendroscope v.3.

Figure 5.5: Phylogenetic trees of locus Cp3995_0202 from alignments trimmed with Gblocks using least stringent settings

Trees were estimated using (a) PhyML from nucleotide alignment; (b) PhyML from protein alignment; (c) MrBayes from nucleotide alignment; (d) MrBayes from protein alignment; and (e) MrBayes from nucleotide alignment using a CP model.

All trees were constructed in TOPALi v.2.5 and formatted in Dendroscope v.3.



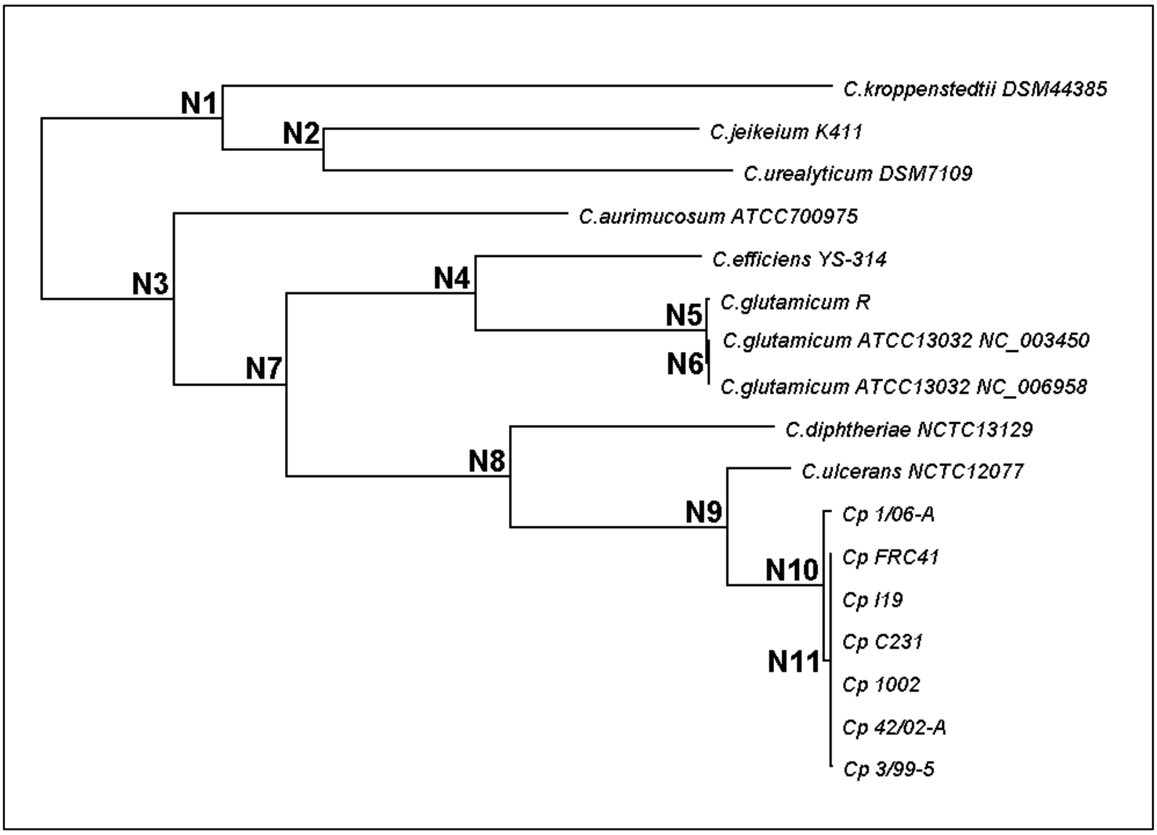


Figure 5.6: Consensus tree for locus Cp3995_0202

Node labels N1-N11 indicate the internal nodes for which support values are available and correspond to **Table 5.1** and the trees in **Figures 5.2-5.5**.

	Untrimmed					Manually trimmed					Gblocks trimmed (default)					Gblocks trimmed (least stringent)				
	PhyML		Mr Bayes			PhyML		Mr Bayes			PhyML		Mr Bayes			PhyML		Mr Bayes		
node	N	P	N	P	CP	N	P	N	P	CP	N	P	N	P	CP	N	P	N	P	CP
1	100	100	1	1	1	100	100	1	1	1	100	100	1	1	1	100	100	1	1	1
2	94	93	1	1	1	93	96	1	1	1	94	99	1	1	1	93	92	1	1	1
3	100	100	1	1	1	100	100	1	1	1	100	100	1	1	1	100	100	1	1	1
4	91	71	1	1	1	93	82	1	1	1	92	78	1	1	1	92	85	1	1	1
5	100	100	1	1	1	100	100	1	1	1	100	100	1	1	1	100	100	1	1	1
6	100	100	1	1	1	100	100	1	1	1	100	100	1	1	1	100	100	1	1	1
7	100	99	0.92	0.79	0.81	100	98	0.94	0.79	0.86	100	97	0.89	0.83	0.94	100	98	0.94	0.78	0.82
8	100	100	1	1	1	100	100	1	1	1	100	100	1	1	1	100	100	1	1	1
9	100	100	1	1	1	100	100	1	1	1	100	100	1	1	1	100	100	1	1	1
10	100	100	1	1	1	100	100	1	1	1	100	100	1	1	1	100	100	1	1	1
11	100	94	1	1	1	100	97	1	1	1	100	76	1	0.86	1	100	94	1	1	1

Table 5.1: Support values for nodes in each of the trees following analysis of locus Cp3995_0202.

Support values for each node, as indicated in the consensus tree (**Fig. 5.6**), are shown as either a bootstrap value for PhyML analyses or a posterior support value for Bayesian analyses. Values for trees deduced from analysis of protein alignments (P) and nucleotide alignments (N) as well as Bayesian codon position analysis (CP) are shown. Alignments remained complete (untrimmed) or were trimmed either manually or using Gblocks software on default or the least stringent settings.

5.2.1.2 Phylogenetic analysis using housekeeping loci

A nucleotide alignment of 33 loci from 17 *Corynebacterial* genomes and an outgroup consisting of 16 *Mycobacterial* genomes, *Rhodococcus jostii* RHA, and *Nocardia farcinica* IFM 10152 was created. Thirty-one of the 33 loci were as described by Wu and Eisen (2008) and sequences of *hsp65* and *cpn60* were also included. The alignment was trimmed using Gblocks using default settings. The alignment was then used to conduct a Bayesian phylogenetic analysis using a codon position model, the resulting tree is shown in **Fig. 5.7**. The topology of the tree is the same to those observed in **section 5.2.1.1**; however there is more distinction between the *Cp* isolates.

Due to similar branch lengths of those branches distinguishing between *Cp* isolates, it was recommended by Dr Frank Wright that it be checked that the set branch length prior parameter of the program was not having an effect on branch length within this clade. Hence, the analysis was repeated twice with branch length priors set to 1 and 100 (as opposed to the default of 10). As TOPALi does not allow access to alter the default priors, for these analyses Mr Bayes was run locally with the branch length prior altered. Both of these analyses resulted in the estimation of the same tree to that shown in **Fig. 5.7** (trees not shown).

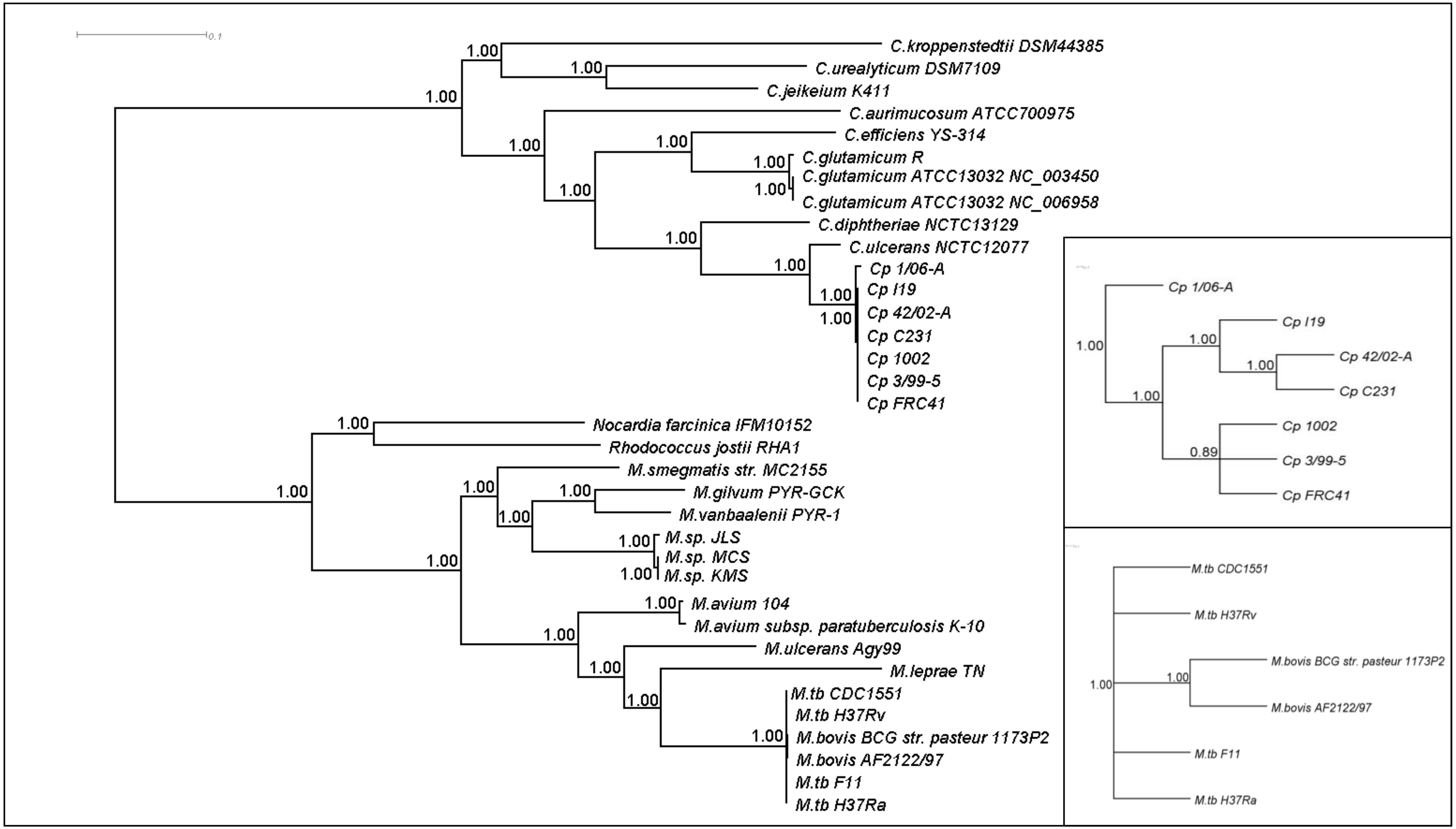


Figure 5.7: Phylogenetic tree showing evolutionary relationships of *Corynebacterium* species based on 33 housekeeping loci

The relationship of evolution amongst the *Corynebacterium* species is presented based on Bayesian analysis of housekeeping loci. Magnified inserts are shown of both the clade containing *Cp* isolates and the clade of *Mycobacterium tuberculosis* in the outgroup. The scale bars represent 0.1 substitutions per nucleotide site.

5.2.1.3 Phylogenetic analysis using loci encoding secreted proteins

A technique employing liquid chromatography electrospray ionisation tandem mass spectrophotometry (LC-ESI MS/MS) to analyse the proteins in an entire lane of an SDS-PAGE gel was used to detect secreted proteins from the supernatant fraction of a *Cp* 3/99-5 preparation. This analysis using criteria stated in Chapter Two, revealed a total of 76 secreted proteins (**Appendix Three**). Of the proteins detected, 32 were also predicted to be extracellular by the compiled bioinformatic analyses described in the previous chapter (**Fig. 5.8** and **Table 5.2**).

The presence or absence of these 32 gene loci was assessed in the genomes belonging to the genus *Corynebacterium*. Genes were only classed as present in a genome if they were a reciprocal best hit between the *Cp* 3/99-5 genome and the query genome. A presence/absence matrix for the secreted protein gene loci is shown in **Fig. 5.9**, from which a clear divide can be observed between the *Cp* and *C. ulcerans* isolates and the rest of the *Corynebacterium* species. The loci are largely conserved in the *Cp* isolates and *C. ulcerans*, only five gene loci are not present in all seven *Cp* strains and *C. ulcerans*, however the loci are considerably sparser in the other *Corynebacterial* genomes.

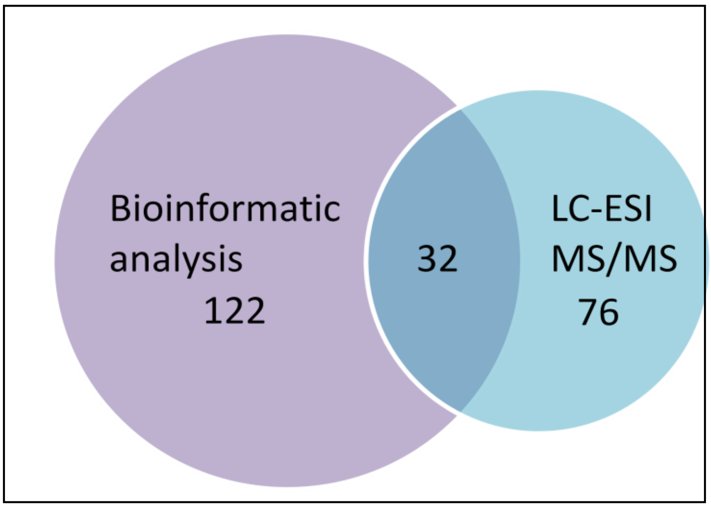


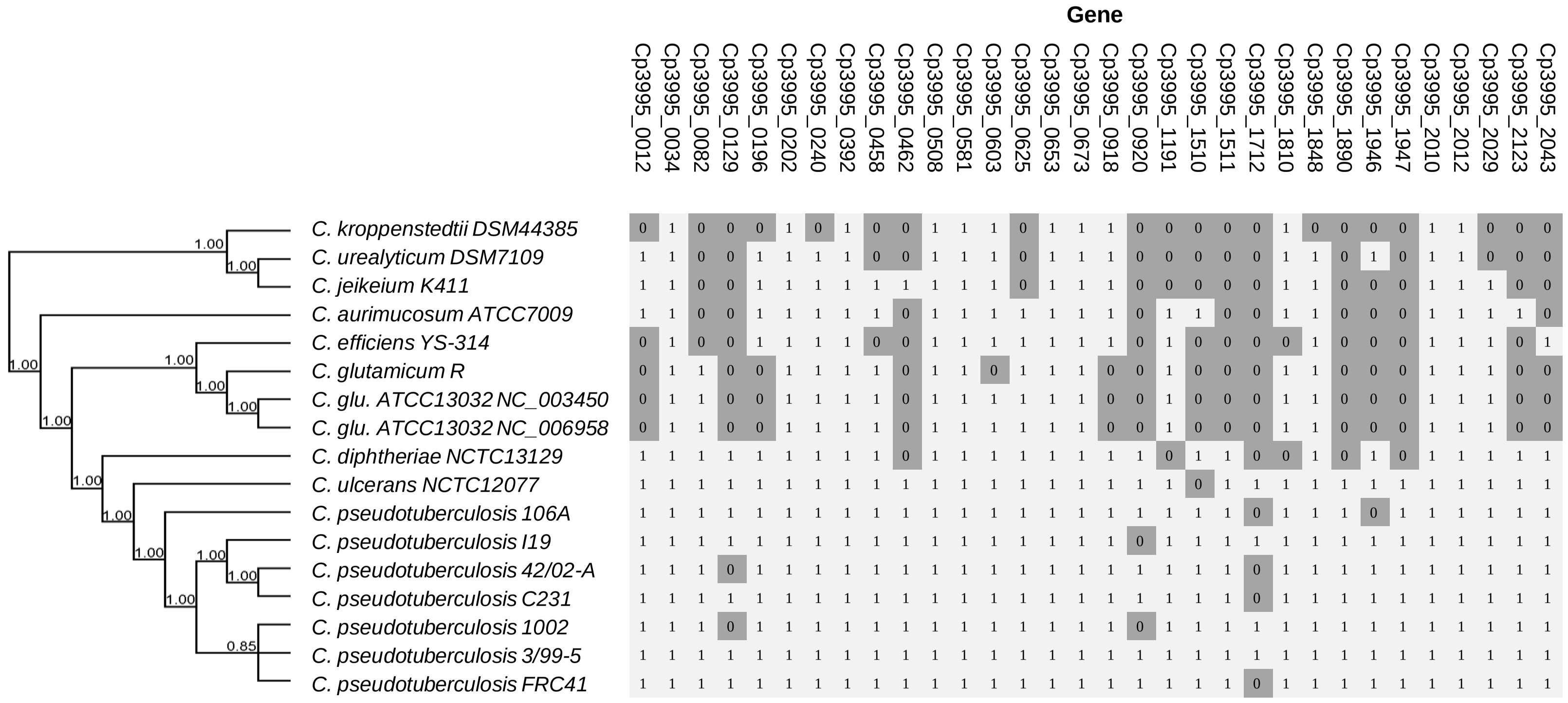
Figure 5.8: Identification of extracellular proteins of *Cp* 3/99-5

Genes encoding secreted proteins were used as loci in the phylogenetic analysis if they were identified by bioinformatic analysis of the *Cp* 3/99-5 genome but also if the corresponding protein was detected by LC-ESI MS/MS from the supernatant of a *Cp* 3/99-5 culture.

Table 5.2: Loci encoding extracellular proteins of *Cp* 3/99-5

Genes identified as encoding secreted proteins are shown along with the gene product. A large proportion of the gene products are hypothetical proteins with unknown functions.

CDS	Gene	Product
Cp3995_0012		Hypothetical protein
Cp3995_0034	*pbpA*	Penicillin-binding protein A
Cp3995_0082		Hypothetical protein
Cp3995_0129		Hypothetical protein
Cp3995_0196		Hypothetical protein
Cp3995_0202	*pbpB*	Penicillin binding protein transpeptidase
Cp3995_0240	*slpA*	Surface layer protein A
Cp3995_0392		L,D-transpeptidase catalytic domain, region YkuD
Cp3995_0458	*htaA*	Cell surface hemin receptor
Cp3995_0462	*htaC*	Hypothetical protein
Cp3995_0508	*lytR*	Transcriptional regulator LytR
Cp3995_0581		Hypothetical protein
Cp3995_0603	*rpfA*	Resuscitation-promoting factor
Cp3995_0625		Hypothetical protein
Cp3995_0653		Uncharacterized metalloprotease
Cp3995_0673	*pepD*	Serine protease
Cp3995_0918	*lpqC*	Poly(3-hydroxybutyrate) depolymerase
Cp3995_0920		Hypothetical protein
Cp3995_1191		Hypothetical protein
Cp3995_1510		Uncharacterized protein HtaC
Cp3995_1511		Cell-surface hemin receptor
Cp3995_1712		Hypothetical protein
Cp3995_1810	*lpqE*	Lipoprotein LpqE
Cp3995_1848	*lipY*	Exported lipase
Cp3995_1890		Hypothetical protein
Cp3995_1946		Membrane protein
Cp3995_1947	*cpp*	Serine protease CP40
Cp3995_2010	*cmtC*	Trehalose corynomycolyl transferase C
Cp3995_2012	*cmtB*	Trehalose corynomycolyl transferase B
Cp3995_2029		Peptidoglycan recognition protein
Cp3995_2043		Hypothetical protein
Cp3995_2123		Hypothetical protein



	Cp3995_0012	Cp3995_0034	Cp3995_0082	Cp3995_0129	Cp3995_0196	Cp3995_0202	Cp3995_0240	Cp3995_0392	Cp3995_0458	Cp3995_0462	Cp3995_0508	Cp3995_0581	Cp3995_0603	Cp3995_0625	Cp3995_0653	Cp3995_0673	Cp3995_0918	Cp3995_0920	Cp3995_1191	Cp3995_1510	Cp3995_1511	Cp3995_1712	Cp3995_1810	Cp3995_1848	Cp3995_1890	Cp3995_1946	Cp3995_1947	Cp3995_2010	Cp3995_2012	Cp3995_2029	Cp3995_2123	Cp3995_2043
C. kroppenstedtii DSM44385	0	1	0	0	0	1	0	1	0	0	1	1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0
C. urealyticum DSM7109	1	1	0	0	1	1	1	1	0	0	1	1	1	0	1	1	1	0	0	0	0	0	1	1	0	1	0	1	1	0	0	0
C. jeikeium K411	1	1	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	0	1	1	0	0	0	1	1	1	0	0
C. aurimucosum ATCC7009	1	1	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	0	0	1	1	0	0	0	1	1	1	1	0
C. efficiens YS-314	0	1	0	0	1	1	1	1	0	0	1	1	1	1	1	1	1	0	1	0	0	0	0	1	0	0	0	1	1	1	0	1
C. glutamicum R	0	1	1	0	0	1	1	1	1	0	1	1	0	1	1	1	0	0	1	0	0	0	1	1	0	0	0	1	1	1	0	0
C. glu. ATCC13032 NC_003450	0	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	0	0	1	1	0	0	0	1	1	1	0	0
C. glu. ATCC13032 NC_006958	0	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	0	0	1	1	0	0	0	1	1	1	0	0
C. diphtheriae NCTC13129	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	1	0	1	1	1	1	1
C. ulcerans NCTC12077	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
C. pseudotuberculosis 106A	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1
C. pseudotuberculosis I19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C. pseudotuberculosis 42/02-A	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
C. pseudotuberculosis C231	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
C. pseudotuberculosis 1002	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C. pseudotuberculosis 3/99-5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C. pseudotuberculosis FRC41	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1

Figure 5.9: Gene presence/absence matrix of the secreted protein loci with phylogenetic tree

The presence/absence of the 32 secreted protein loci was determined in the *Corynebacterium* genus using BLASTp; a gene was determined as present in a genome if it was a reciprocal best hit between the *Cp* 3/99-5 and query genome, and above the E-value cut-off of 1e-10. The presence or absence of a gene is indicated in this matrix by a '1' or '0' respectively (absent genes are alos shaded for ease of view). The phylogenetic tree inferred from Bayesian analysis of an alignment of secreted protein loci is also shown for illustrative purposes; however the branches of the tree are not to scale.

A nucleotide alignment of the 32 loci was created from the 17 *Corynebacterium* sequences and as with the previous analysis, the alignment trimmed with Gblocks. This alignment was then concatenated onto the previous alignment of housekeeping loci for the *Corynebacterium* genus using the FABOX Fasta Alignment Joiner (Villesen, 2007). A Bayesian codon position analysis was performed upon the resulting concatenated alignment in TOPALi (**Fig. 5.10**). The topology of the generated tree was the same as in **Fig. 5.7** from analysis using the housekeeping loci alone. The tree is almost completely resolved, and again the only subclade not resolved contains *Cp* isolates FRC41, 3/99-5 and 1002. However, the posterior support for this subclade is 0.85 which is slightly lower than that in the housekeeping loci tree (0.89).

The analysis was then repeated as before but the alignment of secreted protein loci was trimmed using Gblocks on the most relaxed settings the software would allow. The resulting Bayesian tree, shown in **Fig. 5.11**, is equivalent to those from the previous analyses. Using Gblocks on more relaxed settings has however further reduced the level of support for the *Cp* FRC41, 3/99-5 and 1002 subclade to 0.81.

For both trees estimated from concatenated alignments, the branch lengths within the clade containing just *Cp* isolates appear of equal lengths. The tree estimation for both concatenated alignments was repeated as in **section 5.2.1.2** using minimum and maximum branch length prior settings. Neither tree differed from the original employing the default branch length prior setting (trees not shown).

Using the information gained from tree estimation, predictions could be made as to the gain/loss of secreted protein loci genes within the *Corynebacterium* isolates. If gene loss/gain was not common then the simplest explanation of the evolutionary events to reach the current phylogenetic relationships is that gene loss occurred at certain internal nodes within the tree. The presence/absence matrix was adapted to show the prediction of these events in **Fig. 5.12**.

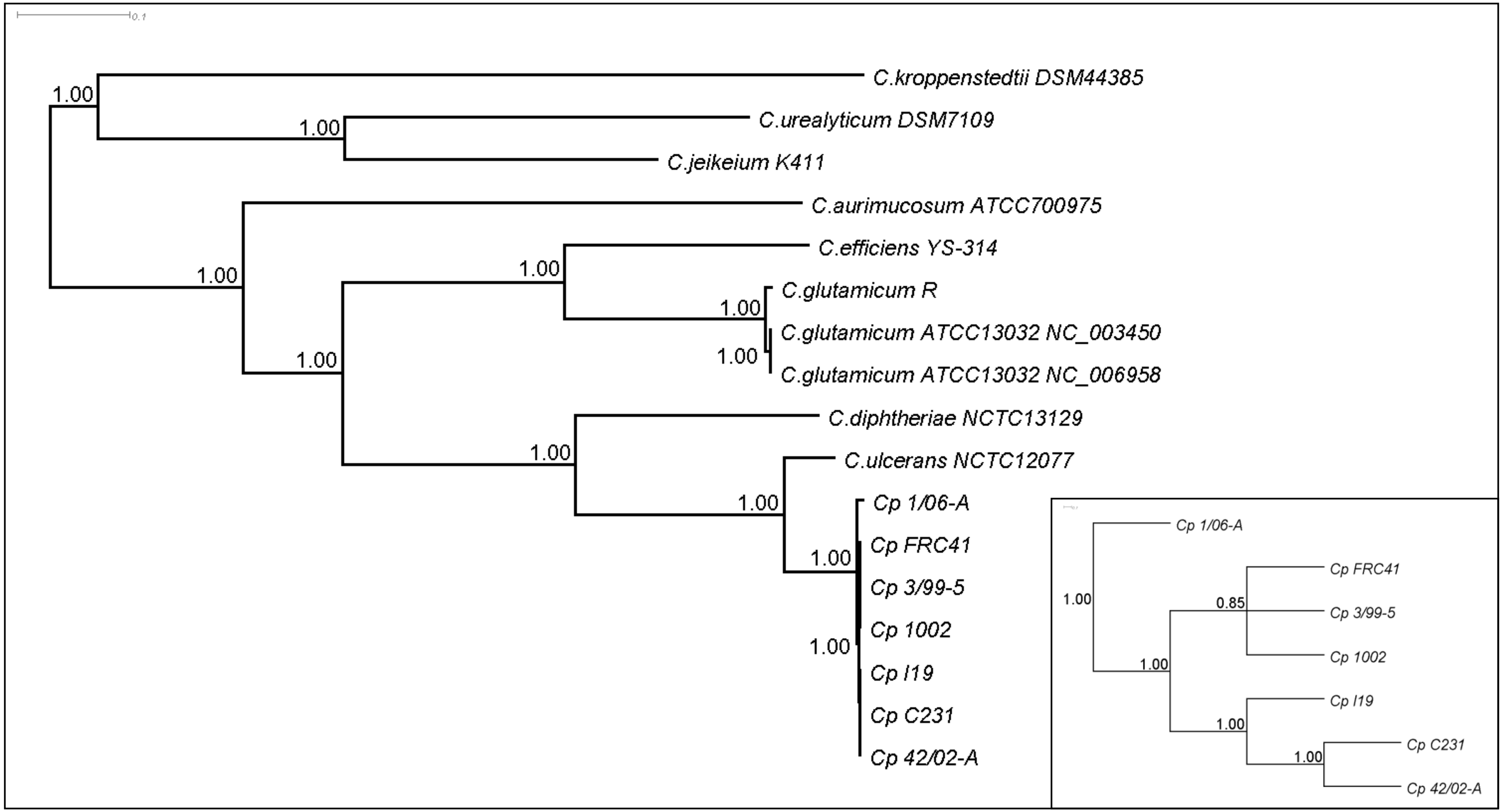


Figure 5.10: Phylogenetic tree based on the analysis of housekeeping loci concatenated with secreted protein loci

The tree was produced from Bayesian analysis of a concatenated alignment of (i) a default Gblocks trimmed alignment of the 33 housekeeping loci with (ii) a default Gblocks trimmed alignment of the 32 loci of secreted proteins. The relationship of the *Cp* isolates is shown at a greater magnification in the box at the bottom right of the figure. The tree is rooted as in **Figure 5.7** and the scale bars represents 0.1 substitutions per nucleotide site.

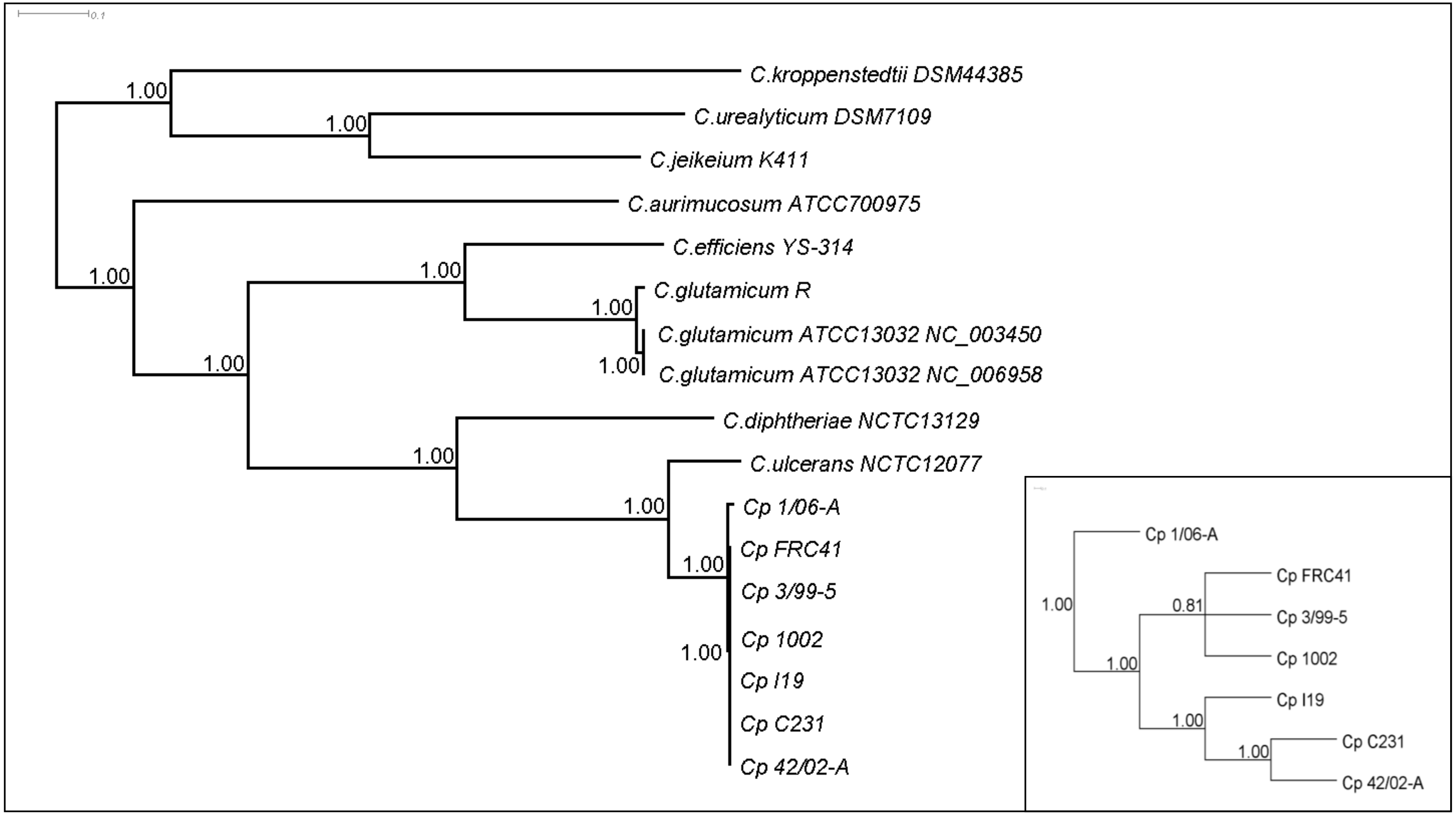
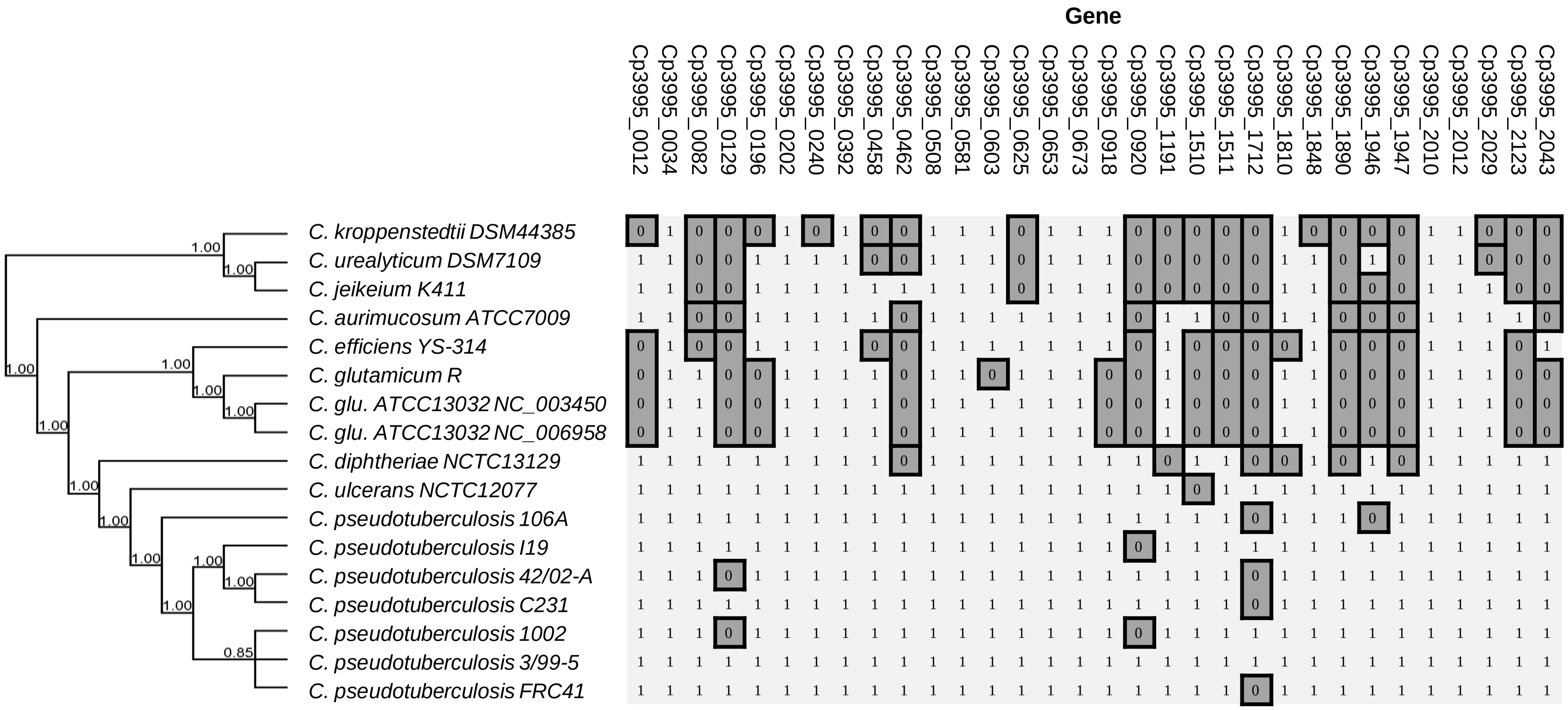


Figure 5.11: Phylogenetic tree based on the analysis of the housekeeping loci concatenated with a less trimmed alignment of secreted protein loci

The tree was produced from Bayesian analysis of a concatenated alignment of (i) a default Gblocks trimmed alignment of the 33 housekeeping loci with (ii) an alignment of the 32 loci of secreted proteins which was trimmed using Gblocks on the most relaxed settings. The relationship of the *Cp* isolates is shown at a greater magnificzation in the box at the bottom right of the figure. Again the tree is rooted as in **Figure 5.7**, and the scale bars represents 0.1 substitutions per nucleotide site.



	Cp3995_0012	Cp3995_0034	Cp3995_0082	Cp3995_0129	Cp3995_0196	Cp3995_0202	Cp3995_0240	Cp3995_0392	Cp3995_0458	Cp3995_0462	Cp3995_0508	Cp3995_0581	Cp3995_0603	Cp3995_0625	Cp3995_0653	Cp3995_0673	Cp3995_0918	Cp3995_0920	Cp3995_1191	Cp3995_1510	Cp3995_1511	Cp3995_1712	Cp3995_1810	Cp3995_1848	Cp3995_1890	Cp3995_1946	Cp3995_1947	Cp3995_2010	Cp3995_2012	Cp3995_2029	Cp3995_2123	Cp3995_2043
C. kroppenstedtii DSM44385	0	1	0	0	0	1	0	1	0	0	1	1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0
C. urealyticum DSM7109	1	1	0	0	1	1	1	1	0	0	1	1	1	0	1	1	1	0	0	0	0	0	1	1	0	1	0	1	1	0	0	0
C. jeikeium K411	1	1	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	0	1	1	0	0	0	1	1	1	0	0
C. aurimucosum ATCC7009	1	1	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	0	0	1	1	0	0	0	1	1	1	1	0
C. efficiens YS-314	0	1	0	0	1	1	1	1	0	0	1	1	1	1	1	1	1	0	1	0	0	0	0	1	0	0	0	1	1	1	0	1
C. glutamicum R	0	1	1	0	0	1	1	1	1	0	1	1	0	1	1	1	0	0	1	0	0	0	1	1	0	0	0	1	1	1	0	0
C. glu. ATCC13032 NC_003450	0	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	0	0	1	1	0	0	0	1	1	1	0	0
C. glu. ATCC13032 NC_006958	0	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	0	0	1	0	0	0	1	1	0	0	0	1	1	1	0	0
C. diphtheriae NCTC13129	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	0	0	1	0	1	0	1	1	1	1	1
C. ulcerans NCTC12077	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
C. pseudotuberculosis 106A	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1
C. pseudotuberculosis I19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C. pseudotuberculosis 42/02-A	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
C. pseudotuberculosis C231	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
C. pseudotuberculosis 1002	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C. pseudotuberculosis 3/99-5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
C. pseudotuberculosis FRC41	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1

Figure 5.12: Presence/absence matrix of secreted protein loci showing gene loss events.

The presence/absence matrix of the 32 secreted protein loci from **Fig. 5.9** is replicated with possible gene loss events indicated by the boxes with thick black borders. For example if we take Cp3995_0012, the 0's for *C. efficiens* and the three *C. glutamicum* isolates show that this gene is absent in these organisms, and the black box round these 0's represents that the gene loss event likely occurred in the ancestor of these four isolates.

Again, the phylogenetic is shown for illustrative purposes but the branches of the tree are not to scale.

5.2.2 Codon usage of loci used in phylogenetic analyses

The codon usage plot of all genes in the *Cp* 3/99-5 genome determined in Chapter Three is duplicated below but with the loci used for phylogenetic analyses in the current chapter highlighted (**Fig. 5.13**). In general the housekeeping gene loci have a lower effective number of codons (ENc) than the gene loci encoding secreted proteins. Also very few of these genes lie on the curve depicting random codon usage.

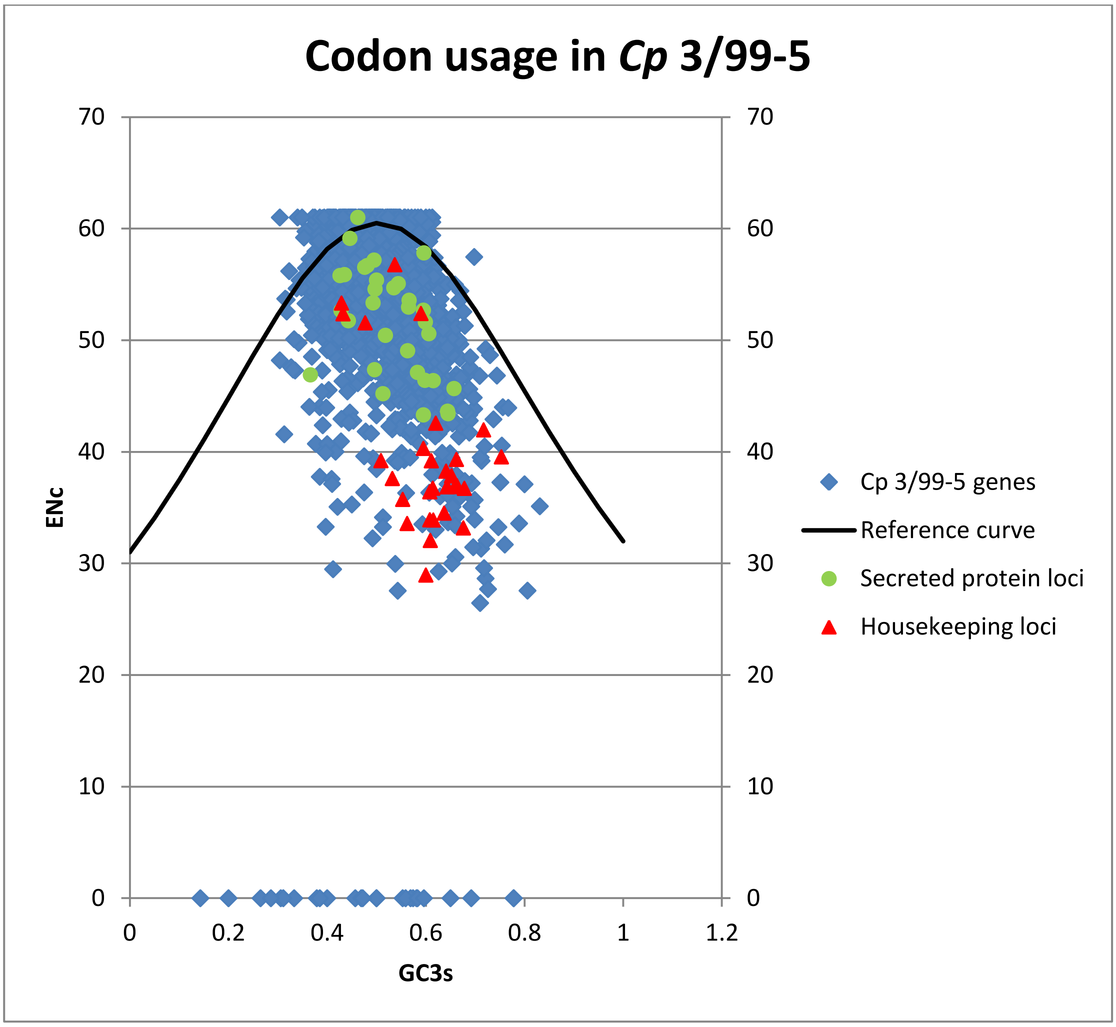


Figure 5.13: Codon usage of the phylogenetic loci

The relationship between the effective number of codons (Enc) and the G+C content at the third synonymous codon position (GC3s) as in **section 3.2.4.1** is shown but with the loci used for phylogenetic analysis highlighted; Housekeeping gene loci used in the current study are shown as red triangles and the secreted protein loci as green circles. The reference curve indicates the relationship between ENc and GC3s when codon usage is random. Generally the housekeeping loci have a greater codon usage bias than the secreted protein loci

5.3 Discussion

Gene *pbpB* (Cp3995_0202) which encodes a penicillin binding protein (PBP) was used to perform preliminary investigations in order to provide confidence in the most appropriate methods for sequence alignment and tree estimation of the relationships between several CMNR genomes. Initial studies were conducted using this secreted protein gene locus rather than a housekeeping locus as secreted protein loci sequences are generally more divergent so would give more obvious differences when determining the best methodology to use in subsequent analyses. Also Cp3995_0202 is one of only nine of the secreted protein loci that are present in all *Corynebacterium* strains included in the study.

PBPs are involved in the synthesis of peptidoglycan, the major component of bacterial cell walls. Peptidoglycan is made of glycan chains of alternating N-acetylglucosamine and N-acetylmuramic acid which are cross-linked by short peptides attached to the N-acetylmuramic acid. PBPs catalyse the polymerisation of the glycan strand (transglycosylation) and the cross-linking between glycan chains (transpeptidation) (Sauvage *et al.*, 2008). Cp3995_0202 has homology with many high M_r PBPs and functional and homology searches indicated it belongs to the PBP 1a family which consists of bifunctional transglycosylase/transpeptidase PBPs. High-M_r PBPs typically have several domains, such as those of PBP2b from *Streptococcus pneumoniae* which has a very small cytoplasmic domain, a predicted transmembrane region and a large periplasmic region (Pagliero *et al.*, 2004). Hence PBPs may not be truly extracellular in the way that, for example, a secreted toxin is. So this protein may not have been the most appropriate example of an extracellular protein to use in the preliminary studies conducted here. However, the protein contains a signal peptide, was predicted to be extracellular by the bioinformatic tools used, and was confirmed to be present in the supernatant of a *Cp* 3/99-5 culture by mass spectrometry, therefore it was classed as extracellular. The extracellular location of the protein could have been further confirmed by raising antibodies which recognise it and using these to probe all cellular fractions (of particular importance the membrane and extracellular fractions) of a *Cp* 3/99-5 culture. Another way to further substantiate the protein's location would be to tag it with, for example, a FLAG™ tag (Sigma-Aldrich Co.). Adding a FLAG™ tag would allow the protein to be followed with an antibody against the FLAG™ sequence, the FLAG™ tag could in turn be detected using western blotting or immunoaffinity chromatography (Einhauer & Jungbauer, 2001).

All trees estimated from the alignment of *pbpB* sequences were of identical topology, only the level of support at some internal nodes differed between trees. The comparison of supporting values for all internal nodes in all trees, shown in **Table 5.1**, indicated the most resolved tree resulted from Bayesian analysis of (i) a manually trimmed nucleotide alignment, (ii) a default Gblocks trimmed nucleotide alignment using a *CP* model, and (iii) a nucleotide alignment trimmed using Gblocks on the least stringent settings.

Overall, tree estimations were of higher resolution when analysis was performed on nucleotide alignments than protein alignments. However, there are two occasions when an individual node (N2) is more resolved using the protein sequence. In both cases this, results from PhyML analysis, one using a manually trimmed alignment and the other using a Gblocks default trimmed alignment. Protein alignments do not account for nucleotide differences which result in a synonymous codon change, so will not distinguish between isolates as well as the equivalent nucleotide alignment.

Some authors consider the best approach to dealing with regions of problematic alignment is to remove these sections prior to tree analysis. However, the opposing argument is that in doing so, important information may be lost. Talavera and Castresana (2007) reported a study in which the previously developed Gblocks program (Castresana, 2000) was used to determine whether removing blocks of problematic alignment leads to more accurate trees. The authors showed that in general the removal of blocks did yield better trees for the divergent sequences tested. Furthermore, in the current study trimming the alignment produced a tree of higher resolution than the analysis performed on an untrimmed alignment. Here, data obtained using the single locus for preliminary investigation revealed that the automated Gblocks alignment trimmer could be used to produce trees of a similar or slightly improved resolution to when trimming was performed manually. The automated process allows the elimination of poorly aligned regions at a much faster rate than the manual trimming, and it also removes the chance of human error.

Also from preliminary investigation, Bayesian analysis appeared superior to the Maximum Likelihood approach of PhyML. Although no methods employed in preliminary work distinguished isolates within a species, the trees were all of high resolution. All trees resulting from MrBayes analysis were so highly resolved that only a single node (N7) had a posterior support of less than 1.0. However PhyML analyses yielded trees that had between two and four nodes with bootstrap values below 100 (depending on the type of alignment used).

Generally Bayesian analysis of a nucleotide alignment using a codon positioning model produced trees that were less resolved than the same analysis performed treating the alignment as simple DNA. Despite this it is likely that analyses using a codon position model have rendered more accurate trees as this method more realistically treats the three codon positions differently. The significance of nucleotide substitutions varies at each of the three codon positions; for example, a substitution at position three is much more likely to be non-synonymous than a similar change at position 2. Importantly, codon position models have previously been reported as underused, but have also been shown to explain evolution better than homogenous models (Shapiro *et al.,* 2006).

The use of codon position models may be more applicable for genomes with codon bias and the codon usage of all *Cp* 3/99-5 genes was determined in **section 3.2.4.1** and was found to be biased for the majority of genes. In the current chapter, the codon usage of those loci used for phylogenetic analyses was highlighted amongst that of all of the *Cp* 3/99-5 genes as determined in Chapter Three. The housekeeping loci predominantly have a greater codon usage bias than the secreted protein loci. Interestingly, many studies have demonstrated a positive correlation between the degree of codon bias and gene expression level (Goetz *et al.,* 2005; Gouy & Gautier, 1982; Hiraoka *et al.,* 2009). This may indicate that the housekeeping loci used here are more highly expressed than the secreted protein loci, however, this would need to be substantiated by a laboratory study into the transcriptomics of the *Cp* 3/99-5 genes.

The phylogenetic relationship of *Corynebacterium* species was inferred from Bayesian analysis of an alignment of housekeeping genes which had been trimmed using Gblocks software. An outgroup of 18 genomes belonging to the CMNR group was included to ensure suitable rooting of the *Corynebacterium* tree and warrant accurate distances at the initial tree branching. The resulting tree is almost entirely resolved with only a single node within the *Cp* group of isolates that has below the maximum level of support (with a posterior support value of 0.89). Interestingly not only have the housekeeping loci been sufficient to differentiate the *Corynebacterium* genus on a species level, but also to some extent within a species. The methodology distinguished between *Cp* isolates, and the resulting subclades consisting of *Cp* strains may be explained by the host species. The only equine isolate (and sole member of the *equi* biovar) is singled out from all of the other *Cp* strains which all belong to the *ovis* biovar. These isolates are separated into two subclades; the first contains *Cp* I19, a bovine isolate originating from Israel, which is distinguished from a further subclade of two ovine isolates (*Cp* 42/02-A and *Cp* C231). The second consists of the caprine isolate *Cp* 1002, the ovine 3/99-5 isolate originating from the UK, and a strain isolated from a case of human lymphadenitis (*Cp* FRC41). Interestingly *Cp* 3/99-5 has not been grouped with the other ovine *Cp* isolates but rather with the only caprine isolate, this may reflect the history of the bacterium in the UK, which is thought to have originated from a goat in 1990 imported from Europe (Lloyd *et al.*, 1990). The isolation and genome sequencing of *Cp* FRC41 was reported by Trost *et al.* (2010), however, the authors do not discuss the background of the 12 year old girl from which the strain was isolated. It may be that the girl was in frequent contact with common hosts of *Cp*, and in particular goats. Indeed this would explain the position of the isolate within the phylogenetic tree, however without sufficient information little conclusion can be drawn on the positioning of *Cp* FRC41 in comparison to the other *Cp* isolates.

During the identification of *Cp* 3/99-5 extracellular proteins there was a low overlap of proteins identified between the computational and laboratory based methods. The low overlap of *Cp* 3/99-5 extracellular proteins may be explained by the limitations of the techniques used to identify them. Bioinformatic approaches are basing predictions on translated nucleotide sequence and have limitations in their capabilities, which are reflected in the discrepancies in results between different prediction software observed in Chapter Three. The LC-ESI MS/MS analysis was only performed under one set of conditions; it is likely that not all extracellular proteins may have been expressed. Some proteins are only transcribed and translated under a particular set of conditions such as a specific temperature or iron concentration. Also even if proteins were expressed they may not have been detected; the mass spectrophotometer used in the LC-ESI MS/MS procedure has detection limits which may result in proteins expressed at very low levels remaining undetected.

Assessment of the presence or absence of the secreted loci within the genus *Corynebacterium* revealed a distinction between the diphtheria group of organisms, and to a greater extent the *C. ulcerans* and *Cp* isolates, from the remaining *Corynebacterium* species. Assuming the gene loss/gain is not common, then the simplest explanation of the clade containing *C. ulcerans* and the seven *Cp* isolates is that the ancestor had all 32 loci and that subsequently gene loss occurred: *C. ulcerans* (one gene, Cp3995_1510), *Cp* 1/06-A (two genes, Cp3995_1712 and Cp3995_1946), *Cp* FRC41 (one gene, Cp3995_1712), *Cp* 1002 (two genes, Cp3995_0129 and Cp3995_0920), *Cp* I19 (one gene, Cp3995_0920), *Cp* 42/02-A (two genes, Cp3995_0129 and Cp3995_1712) and *Cp* C231 (one gene, Cp3995_1712). Since gene loss events are quite rare, it is likely that these gene loss events occurred in the ancestor of some subclades so we might expect that there were only five gene loss events (Cp3995_0129, Cp3995_0920, Cp3995_1510, Cp3995_1712 and Cp3995_1946). Although the phylogenetic tree does not wholly agree with a "parsimonious" set of gene loss events, there are several occurrences which do. For example the loss of Cp3995_1712 from *Cp* isolates 42/02-A and C231 likely occurred in the common ancestor of these two strains.

For the secreted protein loci, the analysis was performed on a concatenation of a Gblocks trimmed alignment of housekeeping loci and an alignment of secreted protein loci that had been trimmed using Gblocks on the least stringent settings. For the later, Gblocks was used on the least stringent settings as the secreted protein loci are absent from some genomes, and this difference would not be taken into account when stricter settings were used. Loci which are not in all genomes would be treated as problematic regions in the alignment and would be completely removed by the Gblocks program. Using the least stringent settings that the automated trimmer would allow meant more of these regions were included in the final alignment for phylogenetic analysis.

The phylogenetic tree of corynebacteria produced from the concatenated alignment of housekeeping loci and secreted protein loci has the same topology as that resulting from the housekeeping loci alone. The only difference is that the posterior support value of the subclade containing *Cp* isolates FRC41, 3/99-5 and 1002 is slightly lower in the tree produced from the concatenated alignment (0.85 compared to 0.89 in the housekeeping loci tree). The secreted proteins have not further distinguished between or within a species, reflecting the high homology of this genus of bacteria.

Phylogenetic analysis of whole genomes would give a truer representation of the relationships between different species, however this would be hugely time consuming and require an excessive amount of computer power. The current study may arguably be an improvement on many studies as it includes a large outgroup, modern methodology and a large number of loci. This work could be improved from the inclusion of a greater number of bacterial isolates. However there are limitations due to the number of loci used, either a genome sequence needs to be obtained for each isolate or the sequences of all loci, which if high numbers of isolates were to be included would ultimately be considerably labour intensive, time consuming and costly. This is likely one of the reasons behind most phylogenetic studies to date being focussed on a single or very few loci. One such study conducted by Retamal *et al.* (2011) investigated a hypervariable segment of the *rpoB* gene and concluded that it's analysis could be used as a diagnostic to differentiate *Cp* strains at a subspecies level. They produced a tree constructed using the neighbour-joining method using a single *C. ulcerans* sequence as an outgroup. However, the current work has indicated Bayesian inference produces more accurate trees. Also a single isolate as an outgroup may not give an accurate indication of the root of the tree.

In the current study, trimming of alignments yielded better phylogenetic trees than equivalent alignments that remained untrimmed, indicating that removal of some problematic sequence can improve phylogenetic analyses. Furthermore, the use of an automated trimming tool, Gblocks, reduced human error during the trimming process and was used to produce slightly improved phylogenetic trees. Resolution of phylogenetic trees also improved when nucleotide sequences were employed rather than protein sequences. In addition, from this work, Bayesian analysis of these alignments can be concluded as a superior method to Maximum Likelihood, with the use of a codon position model deemed more appropriate than a homogenous one (which treats all nucleotides in a codon equally). As a result of these findings, Bayesian analysis of the *Corynebacterium* genus was performed on a Gblocks (default) trimmed nucleotide alignment of housekeeping loci. The analysis indicated the evolutionary relationships between strains on a species level, but did not distinguish strains of the same species. The inclusion of secreted protein loci did not distinguish the isolates any further. However, this study does provide an updated and arguably more accurate phylogeny of the *Corynebacterium* genus than previously reported.

Chapter Six: Identification and analysis of proteins with potential diagnostic applications

6.1 Introduction

As briefly mentioned in Chapter Three, one of the benefits of genome sequencing is the identification of proteins which may have downstream applications in diagnostics or immunotherapy. For a protein to have uses in diagnostics and/or vaccinology it will ideally be conserved amongst all isolates of a species and have immunogenic properties.

Almost all existing vaccines were developed based on traditional vaccinology methods. Such methods require knowledge and rely on known features of a pathogen. Typically only a few candidates could be experimentally tested at a time. In contrast genome sequencing allows access to the entire antigenic repertoire of a pathogen, thus the genomic era has allowed for a completely novel approach to vaccine development. Sequence-based 'Reverse Vaccinology' approaches are becoming more common and aim to predict protective antigens using an *in silico* analysis of a pathogen's genome to identify genes encoding potential vaccine candidates (Rappuoli, 2000). The technique was first used successfully to identify potential vaccine candidates against serogroup B meningococcus (Pizza *et al*., 2000) following the publication of the complete genome sequence of the virulent *Neisseria meningitidis* serogroup B strain MC58 (Tettelin *et al.*, 2010). Approximately 600 new putative secreted or surface-exposed proteins were identified from the genome sequence and 350 of these candidate proteins were used to immunise mice (Pizza *et al.*, 2000). Of these, 28 protein antigens with bactericidal activity were identified and from these five were included in the multicomponent vaccine 5CVMB (5 component vaccine against MenB). The 5CVMB vaccine consists of NadA, factor H binding protein (fHbp; also known as GNA1870) as a fusion protein with GNA2091, and neisserial heparin-binding antigen (NHBA, also known as GNA2132) which is also present as a fusion but with GNA1030 (Giuliani *et al.*, 2006). The sera from mice immunised with 5CVMB were tested in a bactericidal antibody assay and were found to kill 78% of 85 representative worldwide isolates (Giuliani *et al.*, 2006). Phase 2 trials of the vaccine with an outer membrane vesicle component found serum bactericidal antibody titres of ≥ 1/4 in 63–100% of young infants and 96–100% of children 6–8 months old (Sadarangani and Pollard, 2010). These promising results show the potential of utilising genome sequences in the attempt to identify protein candidates for vaccine development.

Reverse vaccinology has also been applied to several other pathogens, including Group B streptococcus. Maione *et al.* (2005) analysed eight Group B streptococcus genome sequences, and cloned and tested 312 surface proteins as vaccines. Four of the proteins elicited protection in mice, and their combination was highly protective against all circulating serotypes.

A further example of successful reverse vaccinology is the identification a vaccine antigen that protects mice from infection by *Chlamydia pneumoniae.* Thorpe *et al.* (2007) tested five proteins previously identified by genomics and proteomics, and discovered one antigen, LcrE, was highly immunogenic.

Although the study of genomes has been more successful in the identification of vaccine candidates, it is not unreasonable that a similar technique be used to identify candidates for diagnostic development. The attributes necessary for a successful diagnostic candidate are extremely similar to those necessary for a successful vaccine candidate; in both cases the protein must elicit a host immune response and are likely to be secreted or surface-bound.

In terms of *Cp* infection, available diagnostic tests are limited, and the only commercially available ELISA in the UK is the ELITEST CLA (Chapter One). The ELITEST CLA (available from Hyphen Biomed, France) uses recombinant PLD to detect anti-PLD IgG antibodies in ovine and caprine sera from animals with CLA. This test has a specificity of 95% and sensitivity of 85%, due to which it is operated on a flock/group basis and is not good for individual animals (Dr Michael Fontaine, personal communication).

Here, the aims were to use the genome sequence of *Cp* to aid identification of novel antigenic proteins in order to improve the CLA diagnostic test available to farmers/stock holders in the UK. This would be either by augmenting the existing test via an improved sensitivity or specificity, or by replacing it with a new superior diagnostic.

6.2 Results

6.2.1 Identification of targets for downstream work

Targets are summarised in **Table 6.1** and were identified by different approaches discussed below.

Target proteins were chosen by analysing all the available data on every predicted protein encoded within the genome. Candidates were shortlisted if they were predicted to be extracellular or contain a cell-surface anchoring motif, as determined in **section 3.2.3.2**, and thought to be unique to *Cp* (established in the first instance by looking for homology with the closest sequenced relative *C. diphtheriae*). Once all criteria had been met, a panel of candidate proteins were produced: 23 proteins with no homology to *C. diphtheriae* were identified; although homology to other Gram-positive bacteria and in some cases other *Corynebacteria spp.* (such as the non-pathogenic *C. efficiens* and *C. glutamicum*) was found. In addition a further 24 proteins which had some homology to proteins in *C. diphtheriae*, but that were considered to have diverged considerably from the *C. diphtheriae* homologue, were identified. This list was reduced further with a view to identifying novel proteins with unknown functions, and proteins related to the pathogenicity. Three proteins, the predicted products of Cp3995_0167, Cp3995_0518 and Cp3995_0570 were picked for further investigation.

From the proteins that were predicted to be secreted by bioinformatic analysis and also determined to be secreted by LS-ECI MS/MS as described in **section 5.2.1.2**, a further three proteins were included: Cp3995_1191, Cp3995_1510, Cp3995_1712.

A previously described protein, CP40, encoded by Cp3995_1947 was also included in the study. This protein is known to be immunogenic, although interestingly there are no studies into the use of it as a diagnostic. There are also no reports in the literature of a *Cp* isolate lacking CP40, so it acts a positive control to the other targets discussed above.

Table 6.1: Targets to be produced as recombinant proteins and assessed as diagnostic markers

CDS	Gene	Size/ bp	Product	Molecular weight of product/ kDa
Cp3995_0167		1557	Hypothetical protein	55.6
Cp3995_0518		768	Hypothetical protein	26.6
Cp3995_0570	*sprT*	675	Trypsin-like serine protease	23.5
Cp3995_1191		342	Hypothetical protein-possible WXG motif	11.6
Cp3995_1510		900	Uncharacterized protein HtaC	31.9
Cp3995_1712		636	Hypothetical protein	23.4
Cp3995_1947	*cpp*	1140	Serine protease CP40	42.9

6.2.2 Target conservation

Conservation of the target proteins was assessed by PCR and selected sequencing amongst a panel of multinational *Cp* isolates and a single *C. ulcerans* strain purchased from the NCTC. The isolates originated from 9 different countries and include 37 ovine, 6 caprine, 3 equine, and 3 bovine *Cp* isolates. With the exception of *cpp*, all of the genes were found to be conserved in all of the isolates (**Table 6.2**). Despite several attempts to amplify the sequence with two different sets of primers, the *cpp* gene was only found to be conserved in approximately 60% of isolates. Genomic DNA of the strains from which *cpp* could not be isolated was prepared using the phenol extraction method described in Chapter Two and further attempts were then made to amplify the gene from these isolates but with no success.

For each of the targets the PCR products of up to eight strains were sequenced by Eurofins MWG. The eight strains (asterisked in **Table 6.2**) represented 6 countries and at least one of each host species available. Sequencing was successful for all PCR products except that of Cp3995_0167 from *C. ulcerans* which failed on more than one occasion on both the forward and reverse strands. The sequences for each target were aligned using Clone Manager software and are shown in **Figures 6.1-6.7**. For all targets the PCR products from different strains share a very high level of homology with differences largely being confined to the equine 1/06-L and bovine 2/06-K isolates. Generally the ends of the sequences do not align well.

Table 6.2: Conservation of target genes in isolate panel.

Presence/absence of target genes was determined by PCR amplification. PCR products were sequenced for those isolates asterisked. Colours refer to host species; green are ovine isolates, blue are caprine, red are equine, yellow are bovine and *C. ulcerans* in white is a human isolate. A negative control is shown in grey.

		Cp3995_0167	Cp3995_0518	Cp3995_0570	Cp3995_1191	Cp3995_1510	Cp3995_1712	Cp3995_1947
Negative control		-	-	-	-	-	-	-
🟩 **3/99-5***	UK	+	+	+	+	+	+	+
🟩 NCTC 3450	S. America	+	+	+	+	+	+	+
🟩 40/01-1	N. Ireland	+	+	+	+	+	+	+
🟩 40/01-2	N. Ireland	+	+	+	+	+	+	+
🟩 40/01-3	N. Ireland	+	+	+	+	+	+	+
🟩 40/01-4	N. Ireland	+	+	+	+	+	+	+
🟩 40/01-5	N. Ireland	+	+	+	+	+	+	+
🟩 40/01-6	N. Ireland	+	+	+	+	+	+	+
🟩 40/01-7	N. Ireland	+	+	+	+	+	+	-
🟩 40/01-8	N. Ireland	+	+	+	+	+	+	+
🟩 40/01-9	Eire	+	+	+	+	+	+	+
🟩 41/01-1	N. Ireland	+	+	+	+	+	+	+
🟩 41/01-2	N. Ireland	+	+	+	+	+	+	-
🟩 13/02-A	N. Ireland	+	+	+	+	+	+	+
🟩 **13/02-B***	N. Ireland	+	+	+	+	+	+	-
🟩 **28/02-A***	Netherlands	+	+	+	+	+	+	+
🟩 28/02-B	Netherlands	+	+	+	+	+	+	+
🟩 28/02-C	Netherlands	+	+	+	+	+	+	+
🟦 **37/02-A***	Canada	+	+	+	+	+	+	-
🟩 37/02-B	Canada	+	+	+	+	+	+	+
🟩 37/02-C	Canada	+	+	+	+	+	+	+
🟩 37/02-D	Canada	+	+	+	+	+	+	+
🟩 37/02-E	Canada	+	+	+	+	+	+	+
🟩 38/02-A	Canada	+	+	+	+	+	+	+
🟩 38/02-B	Canada	+	+	+	+	+	+	+
🟩 38/02-C	Canada	+	+	+	+	+	+	+
🟦 38/02-D	Canada	+	+	+	+	+	+	+
🟩 38/02-E	Canada	+	+	+	+	+	+	+
🟩 38/02-F	Canada	+	+	+	+	+	+	+
🟩 38/02-G	Canada	+	+	+	+	+	+	-
🟩 **42/02-A***	Australia	+	+	+	+	+	+	+
🟩 42/02-B	Australia	+	+	+	+	+	+	-
🟩 42/02-C	Australia	+	+	+	+	+	+	+
🟩 42/02-D	Australia	+	+	+	+	+	+	+
🟩 42/02-E	Australia	+	+	+	+	+	+	+
🟩 42/02-F	Australia	+	+	+	+	+	+	-
🟩 42/02-G	Australia	+	+	+	+	+	+	+
🟩 42/02-H	Australia	+	+	+	+	+	+	+
🟩 42/02-I	Australia	+	+	+	+	+	+	-
🟦 42/02-J	Australia	+	+	+	+	+	+	-
🟦 42/02-K	Australia	+	+	+	+	+	+	-
🟦 42/02-L	Australia	+	+	+	+	+	+	-
🟦 42/02-M	Australia	+	+	+	+	+	+	-
🟩 48/02	Netherlands	+	+	+	+	+	+	-
🟥 1/06-A	USA	+	+	+	+	+	+	-
🟥 **1/06-L***	USA	+	+	+	+	+	+	-
🟥 3/99-4	UK	+	+	+	+	+	+	-
🟨 2/06-C	USA	+	+	+	+	+	+	-
🟨 2/06-G	USA	+	+	+	+	+	+	-
🟨 **2/06-K***	USA	+	+	+	+	+	+	-
C. .ulcerans* NCTC 12077	UK	+??	+	-	-	-	-	-

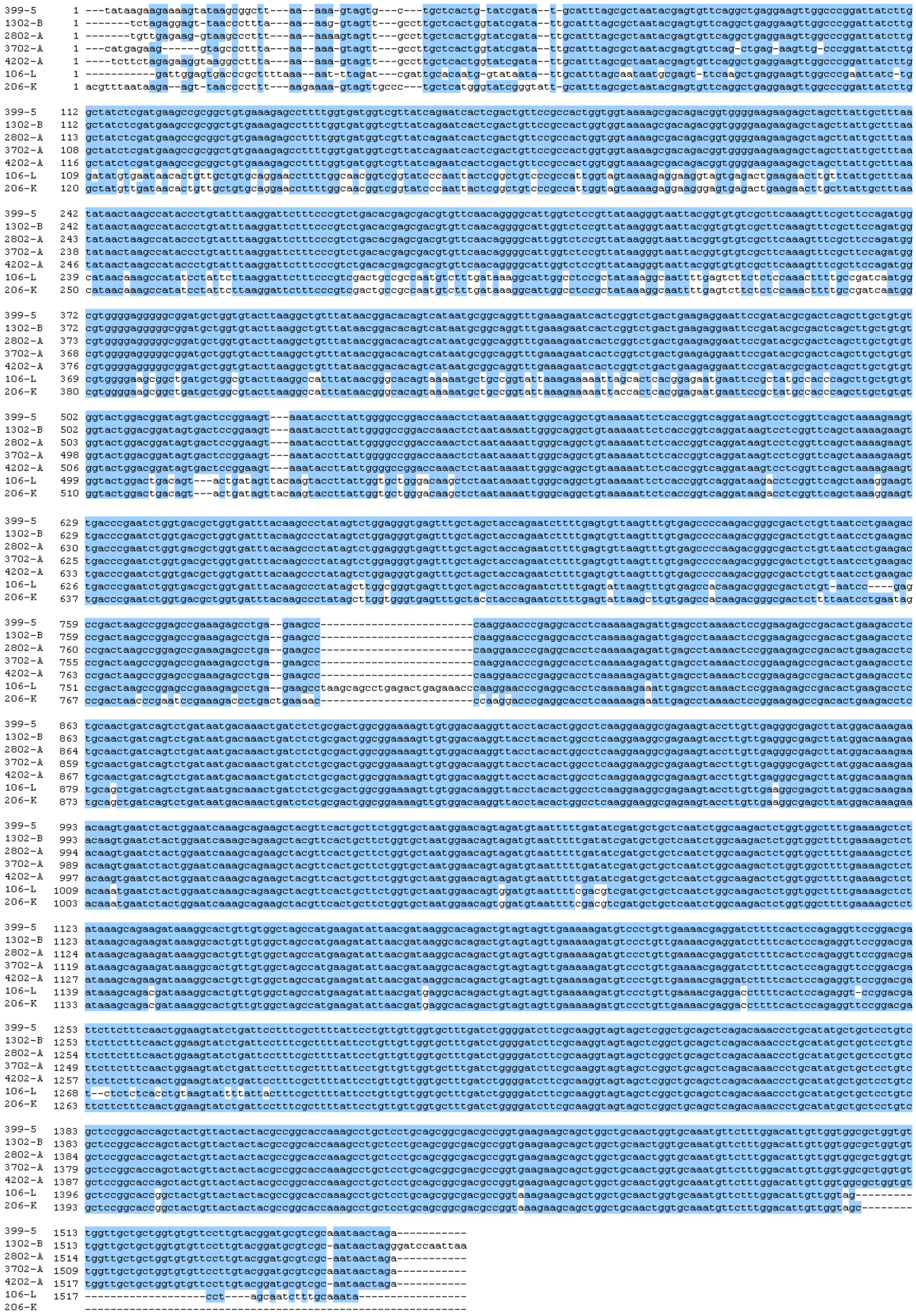
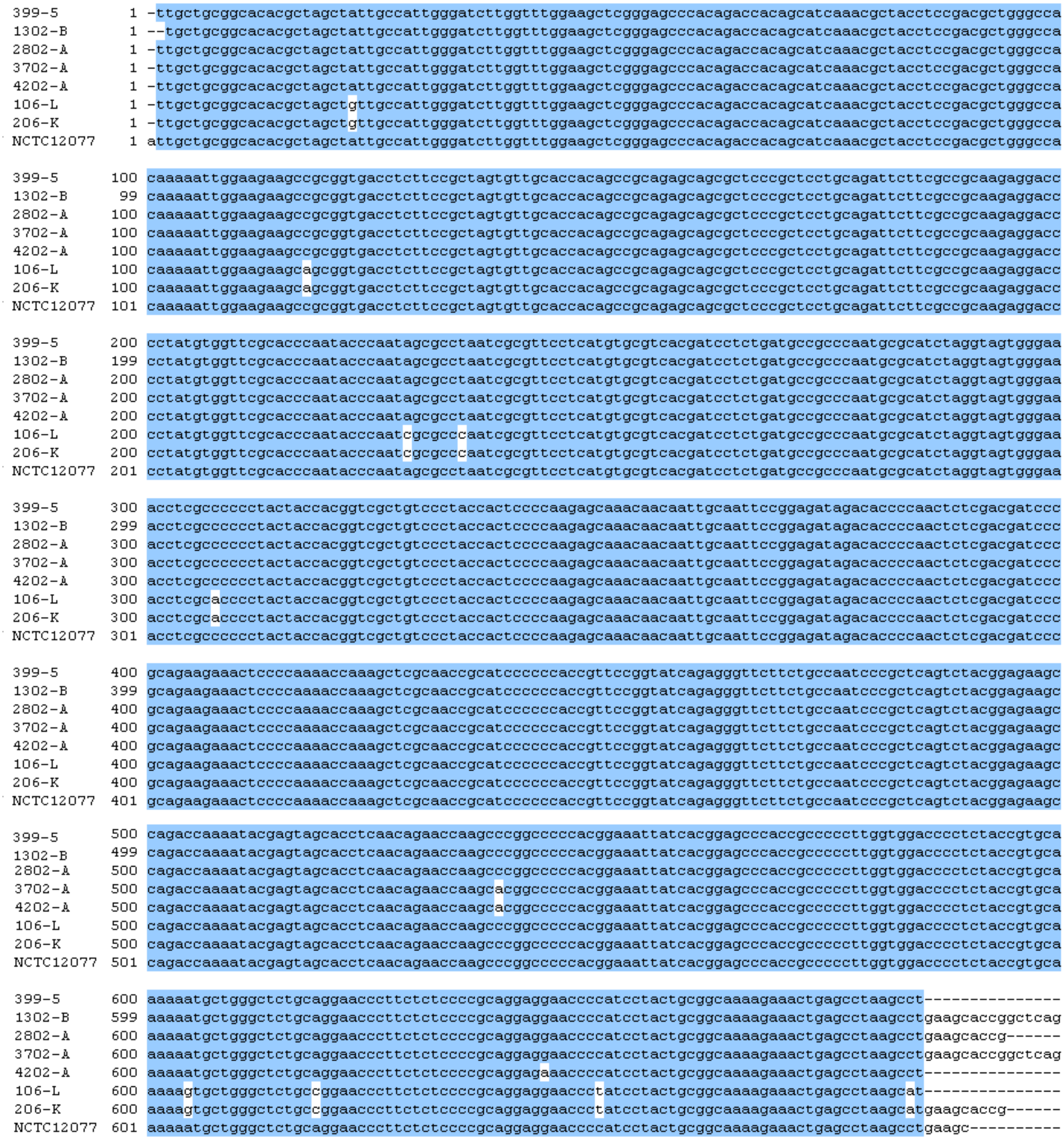


Figure 6.1: Comparison of Cp3995_0167 nucleotide sequences

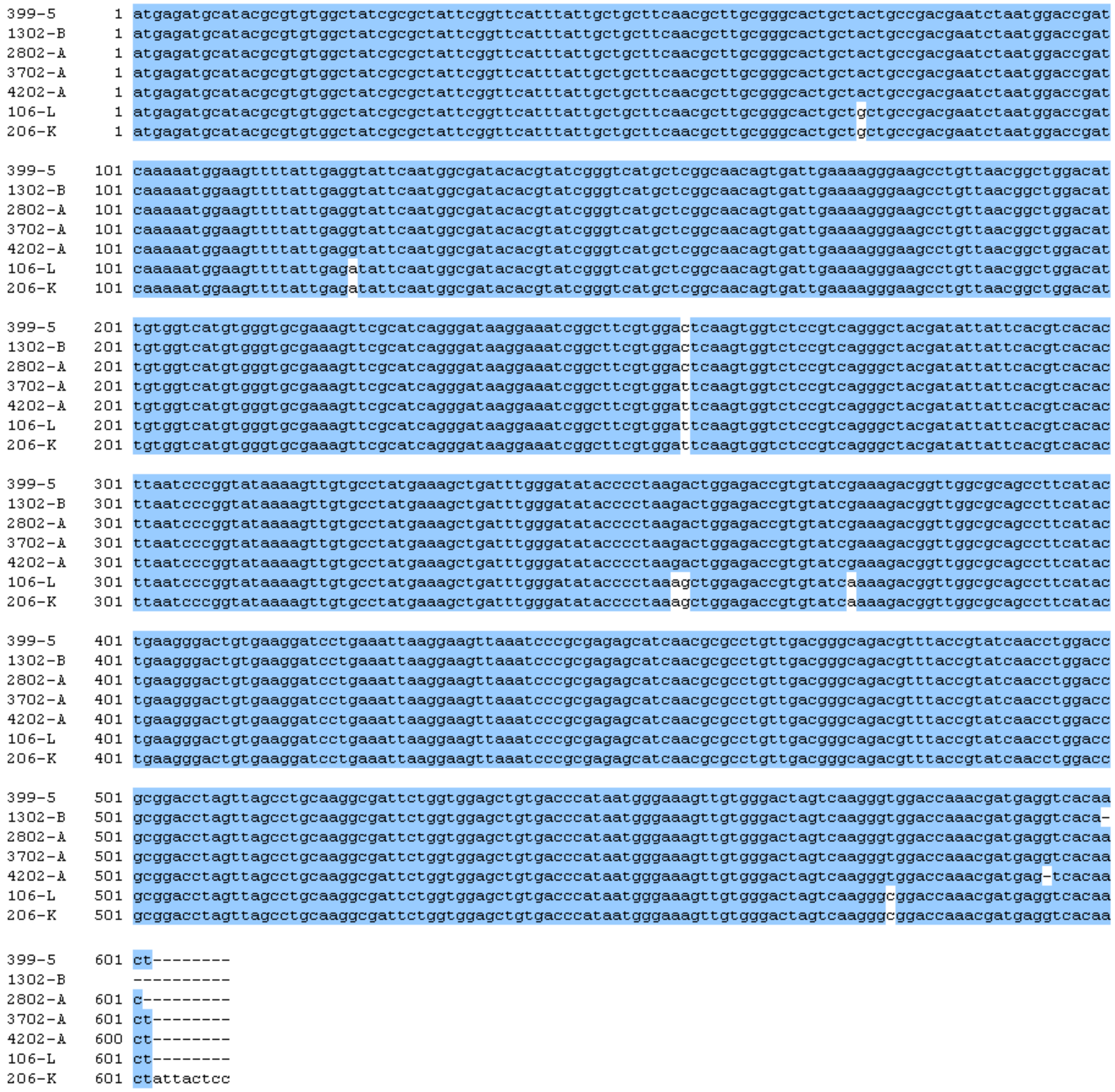
Primers were designed to screen for the presence of Cp3995_ 0167 (encoding a hypothetical protein) in a panel of 50 isolates, PCR products from eight of these strains were then sequenced by Eurofins MWG, however sequencing of the *C. ulcerans* PCR product failed. Edited sequences of the remaining seven shown here were obtained from aligning both forward and reverse sequences. The edited sequences were then aligned using Clone Manager and the image copied from this software.



```
399-5        1 -ttgctgcggcacacgctagctattgccattgggatcttggtttggaagctcgggagcccacagaccacagcatcaaacgctacctccgacgctgggcca
1302-B       1 --tgctgcggcacacgctagctattgccattgggatcttggtttggaagctcgggagcccacagaccacagcatcaaacgctacctccgacgctgggcca
2802-A       1 -ttgctgcggcacacgctagctattgccattgggatcttggtttggaagctcgggagcccacagaccacagcatcaaacgctacctccgacgctgggcca
3702-A       1 -ttgctgcggcacacgctagctattgccattgggatcttggtttggaagctcgggagcccacagaccacagcatcaaacgctacctccgacgctgggcca
4202-A       1 -ttgctgcggcacacgctagctattgccattgggatcttggtttggaagctcgggagcccacagaccacagcatcaaacgctacctccgacgctgggcca
106-L        1 -ttgctgcggcacacgctagctgttgccattgggatcttggtttggaagctcgggagcccacagaccacagcatcaaacgctacctccgacgctgggcca
206-K        1 -ttgctgcggcacacgctagctgttgccattgggatcttggtttggaagctcgggagcccacagaccacagcatcaaacgctacctccgacgctgggcca
NCTC12077    1 attgctgcggcacacgctagctattgccattgggatcttggtttggaagctcgggagcccacagaccacagcatcaaacgctacctccgacgctgggcca

399-5      100 caaaaattggaagaagccgcggtgacctcttccgctagtgttgcaccacagccgcagagcagcgctcccgctcctgcagattcttcgccgcaagaggacc
1302-B      99 caaaaattggaagaagccgcggtgacctcttccgctagtgttgcaccacagccgcagagcagcgctcccgctcctgcagattcttcgccgcaagaggacc
2802-A     100 caaaaattggaagaagccgcggtgacctcttccgctagtgttgcaccacagccgcagagcagcgctcccgctcctgcagattcttcgccgcaagaggacc
3702-A     100 caaaaattggaagaagccgcggtgacctcttccgctagtgttgcaccacagccgcagagcagcgctcccgctcctgcagattcttcgccgcaagaggacc
4202-A     100 caaaaattggaagaagccgcggtgacctcttccgctagtgttgcaccacagccgcagagcagcgctcccgctcctgcagattcttcgccgcaagaggacc
106-L      100 caaaaattggaagaagcagcggtgacctcttccgctagtgttgcaccacagccgcagagcagcgctcccgctcctgcagattcttcgccgcaagaggacc
206-K      100 caaaaattggaagaagcagcggtgacctcttccgctagtgttgcaccacagccgcagagcagcgctcccgctcctgcagattcttcgccgcaagaggacc
NCTC12077  101 caaaaattggaagaagccgcggtgacctcttccgctagtgttgcaccacagccgcagagcagcgctcccgctcctgcagattcttcgccgcaagaggacc

399-5      200 cctatgtggttcgcacccaatacccaatagcgcctaatcgcgttcctcatgtgcgtcacgatcctctgatgccgcccaatgcgcatctaggtagtgggaa
1302-B     199 cctatgtggttcgcacccaatacccaatagcgcctaatcgcgttcctcatgtgcgtcacgatcctctgatgccgcccaatgcgcatctaggtagtgggaa
2802-A     200 cctatgtggttcgcacccaatacccaatagcgcctaatcgcgttcctcatgtgcgtcacgatcctctgatgccgcccaatgcgcatctaggtagtgggaa
3702-A     200 cctatgtggttcgcacccaatacccaatagcgcctaatcgcgttcctcatgtgcgtcacgatcctctgatgccgcccaatgcgcatctaggtagtgggaa
4202-A     200 cctatgtggttcgcacccaatacccaatagcgcctaatcgcgttcctcatgtgcgtcacgatcctctgatgccgcccaatgcgcatctaggtagtgggaa
106-L      200 cctatgtggttcgcacccaatacccaatcgcgcccaatcgcgttcctcatgtgcgtcacgatcctctgatgccgcccaatgcgcatctaggtagtgggaa
206-K      200 cctatgtggttcgcacccaatacccaatcgcgcccaatcgcgttcctcatgtgcgtcacgatcctctgatgccgcccaatgcgcatctaggtagtgggaa
NCTC12077  201 cctatgtggttcgcacccaatacccaatagcgcctaatcgcgttcctcatgtgcgtcacgatcctctgatgccgcccaatgcgcatctaggtagtgggaa

399-5      300 acctcgcccccctactaccacggtcgctgtccctaccactccccaagagcaaacaacaattgcaattccggagatagacaccccaactctcgacgatccc
1302-B     299 acctcgcccccctactaccacggtcgctgtccctaccactccccaagagcaaacaacaattgcaattccggagatagacaccccaactctcgacgatccc
2802-A     300 acctcgcccccctactaccacggtcgctgtccctaccactccccaagagcaaacaacaattgcaattccggagatagacaccccaactctcgacgatccc
3702-A     300 acctcgcccccctactaccacggtcgctgtccctaccactccccaagagcaaacaacaattgcaattccggagatagacaccccaactctcgacgatccc
4202-A     300 acctcgcccccctactaccacggtcgctgtccctaccactccccaagagcaaacaacaattgcaattccggagatagacaccccaactctcgacgatccc
106-L      300 acctcgcacccctactaccacggtcgctgtccctaccactccccaagagcaaacaacaattgcaattccggagatagacaccccaactctcgacgatccc
206-K      300 acctcgcacccctactaccacggtcgctgtccctaccactccccaagagcaaacaacaattgcaattccggagatagacaccccaactctcgacgatccc
NCTC12077  301 acctcgcccccctactaccacggtcgctgtccctaccactccccaagagcaaacaacaattgcaattccggagatagacaccccaactctcgacgatccc

399-5      400 gcagaagaaactccccaaaaccaaagctcgcaaccgcatccccccaccgttccggtatcagagggttcttctgccaatcccgctcagtctacggagaagc
1302-B     399 gcagaagaaactccccaaaaccaaagctcgcaaccgcatccccccaccgttccggtatcagagggttcttctgccaatcccgctcagtctacggagaagc
2802-A     400 gcagaagaaactccccaaaaccaaagctcgcaaccgcatccccccaccgttccggtatcagagggttcttctgccaatcccgctcagtctacggagaagc
3702-A     400 gcagaagaaactccccaaaaccaaagctcgcaaccgcatccccccaccgttccggtatcagagggttcttctgccaatcccgctcagtctacggagaagc
4202-A     400 gcagaagaaactccccaaaaccaaagctcgcaaccgcatccccccaccgttccggtatcagagggttcttctgccaatcccgctcagtctacggagaagc
106-L      400 gcagaagaaactccccaaaaccaaagctcgcaaccgcatccccccaccgttccggtatcagagggttcttctgccaatcccgctcagtctacggagaagc
206-K      400 gcagaagaaactccccaaaaccaaagctcgcaaccgcatccccccaccgttccggtatcagagggttcttctgccaatcccgctcagtctacggagaagc
NCTC12077  401 gcagaagaaactccccaaaaccaaagctcgcaaccgcatccccccaccgttccggtatcagagggttcttctgccaatcccgctcagtctacggagaagc

399-5      500 cagaccaaaatacgagtagcacctcaacagaaccaagcccggcccccacggaaattatcacggagcccaccgcccccttggtggacccctctaccgtgca
1302-B     499 cagaccaaaatacgagtagcacctcaacagaaccaagcccggcccccacggaaattatcacggagcccaccgcccccttggtggacccctctaccgtgca
2802-A     500 cagaccaaaatacgagtagcacctcaacagaaccaagcccggcccccacggaaattatcacggagcccaccgcccccttggtggacccctctaccgtgca
3702-A     500 cagaccaaaatacgagtagcacctcaacagaaccaagcacggcccccacggaaattatcacggagcccaccgcccccttggtggacccctctaccgtgca
4202-A     500 cagaccaaaatacgagtagcacctcaacagaaccaagcacggcccccacggaaattatcacggagcccaccgcccccttggtggacccctctaccgtgca
106-L      500 cagaccaaaatacgagtagcacctcaacagaaccaagcccggcccccacggaaattatcacggagcccaccgcccccttggtggacccctctaccgtgca
206-K      500 cagaccaaaatacgagtagcacctcaacagaaccaagcccggcccccacggaaattatcacggagcccaccgcccccttggtggacccctctaccgtgca
NCTC12077  501 cagaccaaaatacgagtagcacctcaacagaaccaagcccggcccccacggaaattatcacggagcccaccgcccccttggtggacccctctaccgtgca

399-5      600 aaaaatgctgggctctgcaggaacccttctctccccgcaggaggaaccccatcctactgcggcaaaagaaactgagcctaagcct---------------
1302-B     599 aaaaatgctgggctctgcaggaacccttctctccccgcaggaggaaccccatcctactgcggcaaaagaaactgagcctaagcctgaagcaccggctcag
2802-A     600 aaaaatgctgggctctgcaggaacccttctctccccgcaggaggaaccccatcctactgcggcaaaagaaactgagcctaagcctgaagcaccg------
3702-A     600 aaaaatgctgggctctgcaggaacccttctctccccgcaggaggaaccccatcctactgcggcaaaagaaactgagcctaagcctgaagcaccggctcag
4202-A     600 aaaaatgctgggctctgcaggaacccttctctccccgcaggagaaaccccatcctactgcggcaaaagaaactgagcctaagcct---------------
106-L      600 aaaagtgctgggctctgccggaacccttctctccccgcaggaggaaccctatcctactgcggcaaaagaaactgagcctaagcat---------------
206-K      600 aaaagtgctgggctctgccggaacccttctctccccgcaggaggaaccctatcctactgcggcaaaagaaactgagcctaagcatgaagcaccg------
NCTC12077  601 aaaaatgctgggctctgcaggaacccttctctccccgcaggaggaaccccatcctactgcggcaaaagaaactgagcctaagcctgaagc----------
```

Figure 6.2: Comparison of Cp3995_0518 nucleotide sequences

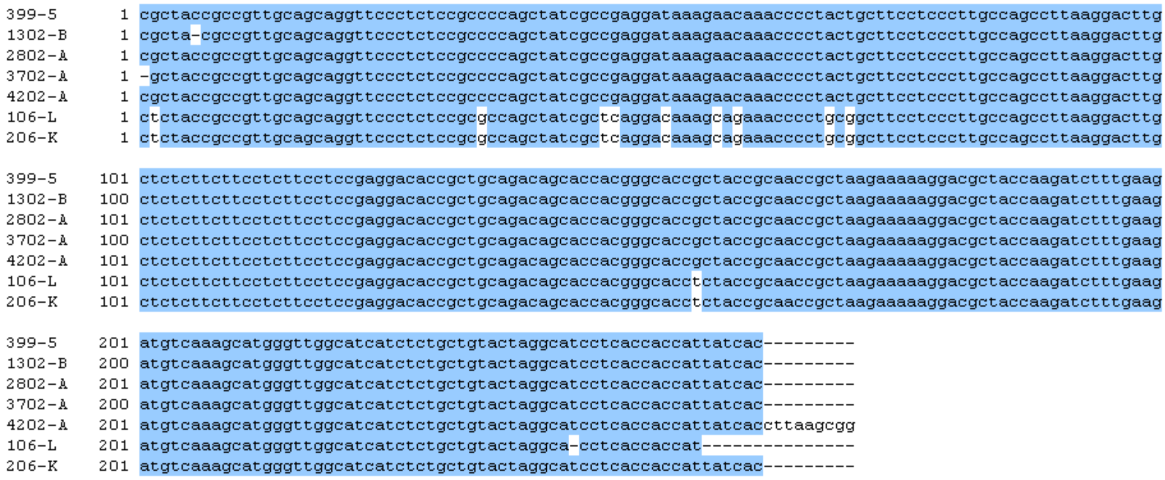
Cp3995_0518 encodes a hypothetical protein, and primers were designed to screen for the presence of the gene in a panel of 50 isolates, PCR products from eight of these strains were then sequenced by Eurofins MWG. Edited sequences shown were obtained from aligning both forward and reverse sequences. The edited sequences were then aligned using Clone Manager and the image copied from this software.



```
399-5      1 atgagatgcatacgcgtgtggctatcgcgctattcggttcatttattgctgcttcaacgcttgcgggcactgctactgccgacgaatctaatggaccgat
1302-B     1 atgagatgcatacgcgtgtggctatcgcgctattcggttcatttattgctgcttcaacgcttgcgggcactgctactgccgacgaatctaatggaccgat
2802-A     1 atgagatgcatacgcgtgtggctatcgcgctattcggttcatttattgctgcttcaacgcttgcgggcactgctactgccgacgaatctaatggaccgat
3702-A     1 atgagatgcatacgcgtgtggctatcgcgctattcggttcatttattgctgcttcaacgcttgcgggcactgctactgccgacgaatctaatggaccgat
4202-A     1 atgagatgcatacgcgtgtggctatcgcgctattcggttcatttattgctgcttcaacgcttgcgggcactgctactgccgacgaatctaatggaccgat
106-L      1 atgagatgcatacgcgtgtggctatcgcgctattcggttcatttattgctgcttcaacgcttgcgggcactgctgctgccgacgaatctaatggaccgat
206-K      1 atgagatgcatacgcgtgtggctatcgcgctattcggttcatttattgctgcttcaacgcttgcgggcactgctgctgccgacgaatctaatggaccgat

399-5    101 caaaaatggaagttttattgaggtattcaatggcgatacacgtatcgggtcatgctcggcaacagtgattgaaaagggaagcctgttaacggctggacat
1302-B   101 caaaaatggaagttttattgaggtattcaatggcgatacacgtatcgggtcatgctcggcaacagtgattgaaaagggaagcctgttaacggctggacat
2802-A   101 caaaaatggaagttttattgaggtattcaatggcgatacacgtatcgggtcatgctcggcaacagtgattgaaaagggaagcctgttaacggctggacat
3702-A   101 caaaaatggaagttttattgaggtattcaatggcgatacacgtatcgggtcatgctcggcaacagtgattgaaaagggaagcctgttaacggctggacat
4202-A   101 caaaaatggaagttttattgaggtattcaatggcgatacacgtatcgggtcatgctcggcaacagtgattgaaaagggaagcctgttaacggctggacat
106-L    101 caaaaatggaagttttattgagatattcaatggcgatacacgtatcgggtcatgctcggcaacagtgattgaaaagggaagcctgttaacggctggacat
206-K    101 caaaaatggaagttttattgagatattcaatggcgatacacgtatcgggtcatgctcggcaacagtgattgaaaagggaagcctgttaacggctggacat

399-5    201 tgtggtcatgtgggtgcgaaagttcgcatcagggataaggaaatcggcttcgtggactcaagtggtctccgtcagggctacgatattattcacgtcacac
1302-B   201 tgtggtcatgtgggtgcgaaagttcgcatcagggataaggaaatcggcttcgtggactcaagtggtctccgtcagggctacgatattattcacgtcacac
2802-A   201 tgtggtcatgtgggtgcgaaagttcgcatcagggataaggaaatcggcttcgtggactcaagtggtctccgtcagggctacgatattattcacgtcacac
3702-A   201 tgtggtcatgtgggtgcgaaagttcgcatcagggataaggaaatcggcttcgtggattcaagtggtctccgtcagggctacgatattattcacgtcacac
4202-A   201 tgtggtcatgtgggtgcgaaagttcgcatcagggataaggaaatcggcttcgtggattcaagtggtctccgtcagggctacgatattattcacgtcacac
106-L    201 tgtggtcatgtgggtgcgaaagttcgcatcagggataaggaaatcggcttcgtggattcaagtggtctccgtcagggctacgatattattcacgtcacac
206-K    201 tgtggtcatgtgggtgcgaaagttcgcatcagggataaggaaatcggcttcgtggattcaagtggtctccgtcagggctacgatattattcacgtcacac

399-5    301 ttaatcccggtataaaagttgtgcctatgaaagctgatttgggatatacccctaagactggagaccgtgtatcgaaagacggttggcgcagccttcatac
1302-B   301 ttaatcccggtataaaagttgtgcctatgaaagctgatttgggatatacccctaagactggagaccgtgtatcgaaagacggttggcgcagccttcatac
2802-A   301 ttaatcccggtataaaagttgtgcctatgaaagctgatttgggatatacccctaagactggagaccgtgtatcgaaagacggttggcgcagccttcatac
3702-A   301 ttaatcccggtataaaagttgtgcctatgaaagctgatttgggatatacccctaagactggagaccgtgtatcgaaagacggttggcgcagccttcatac
4202-A   301 ttaatcccggtataaaagttgtgcctatgaaagctgatttgggatatacccctaagactggagaccgtgtatcgaaagacggttggcgcagccttcatac
106-L    301 ttaatcccggtataaaagttgtgcctatgaaagctgatttgggatatacccctaaagctggagaccgtgtatcaaaagacggttggcgcagccttcatac
206-K    301 ttaatcccggtataaaagttgtgcctatgaaagctgatttgggatatacccctaaagctggagaccgtgtatcaaaagacggttggcgcagccttcatac

399-5    401 tgaagggactgtgaaggatcctgaaattaaggaagttaaatcccgcgagagcatcaacgcgcctgttgacgggcagacgtttaccgtatcaacctggacc
1302-B   401 tgaagggactgtgaaggatcctgaaattaaggaagttaaatcccgcgagagcatcaacgcgcctgttgacgggcagacgtttaccgtatcaacctggacc
2802-A   401 tgaagggactgtgaaggatcctgaaattaaggaagttaaatcccgcgagagcatcaacgcgcctgttgacgggcagacgtttaccgtatcaacctggacc
3702-A   401 tgaagggactgtgaaggatcctgaaattaaggaagttaaatcccgcgagagcatcaacgcgcctgttgacgggcagacgtttaccgtatcaacctggacc
4202-A   401 tgaagggactgtgaaggatcctgaaattaaggaagttaaatcccgcgagagcatcaacgcgcctgttgacgggcagacgtttaccgtatcaacctggacc
106-L    401 tgaagggactgtgaaggatcctgaaattaaggaagttaaatcccgcgagagcatcaacgcgcctgttgacgggcagacgtttaccgtatcaacctggacc
206-K    401 tgaagggactgtgaaggatcctgaaattaaggaagttaaatcccgcgagagcatcaacgcgcctgttgacgggcagacgtttaccgtatcaacctggacc

399-5    501 gcggacctagttagcctgcaaggcgattctggtggagctgtgacccataatgggaaagttgtgggactagtcaagggtggaccaaacgatgaggtcacaa
1302-B   501 gcggacctagttagcctgcaaggcgattctggtggagctgtgacccataatgggaaagttgtgggactagtcaagggtggaccaaacgatgaggtcaca-
2802-A   501 gcggacctagttagcctgcaaggcgattctggtggagctgtgacccataatgggaaagttgtgggactagtcaagggtggaccaaacgatgaggtcacaa
3702-A   501 gcggacctagttagcctgcaaggcgattctggtggagctgtgacccataatgggaaagttgtgggactagtcaagggtggaccaaacgatgaggtcacaa
4202-A   501 gcggacctagttagcctgcaaggcgattctggtggagctgtgacccataatgggaaagttgtgggactagtcaagggtggaccaaacgatgag-tcacaa
106-L    501 gcggacctagttagcctgcaaggcgattctggtggagctgtgacccataatgggaaagttgtgggactagtcaagggcggaccaaacgatgaggtcacaa
206-K    501 gcggacctagttagcctgcaaggcgattctggtggagctgtgacccataatgggaaagttgtgggactagtcaagggcggaccaaacgatgaggtcacaa

399-5    601 ct--------
1302-B       ----------
2802-A   601 c---------
3702-A   601 ct--------
4202-A   600 ct--------
106-L    601 ct--------
206-K    601 ctattactcc
```

Figure 6.3: Comparison of Cp3995_0570 nucleotide sequences

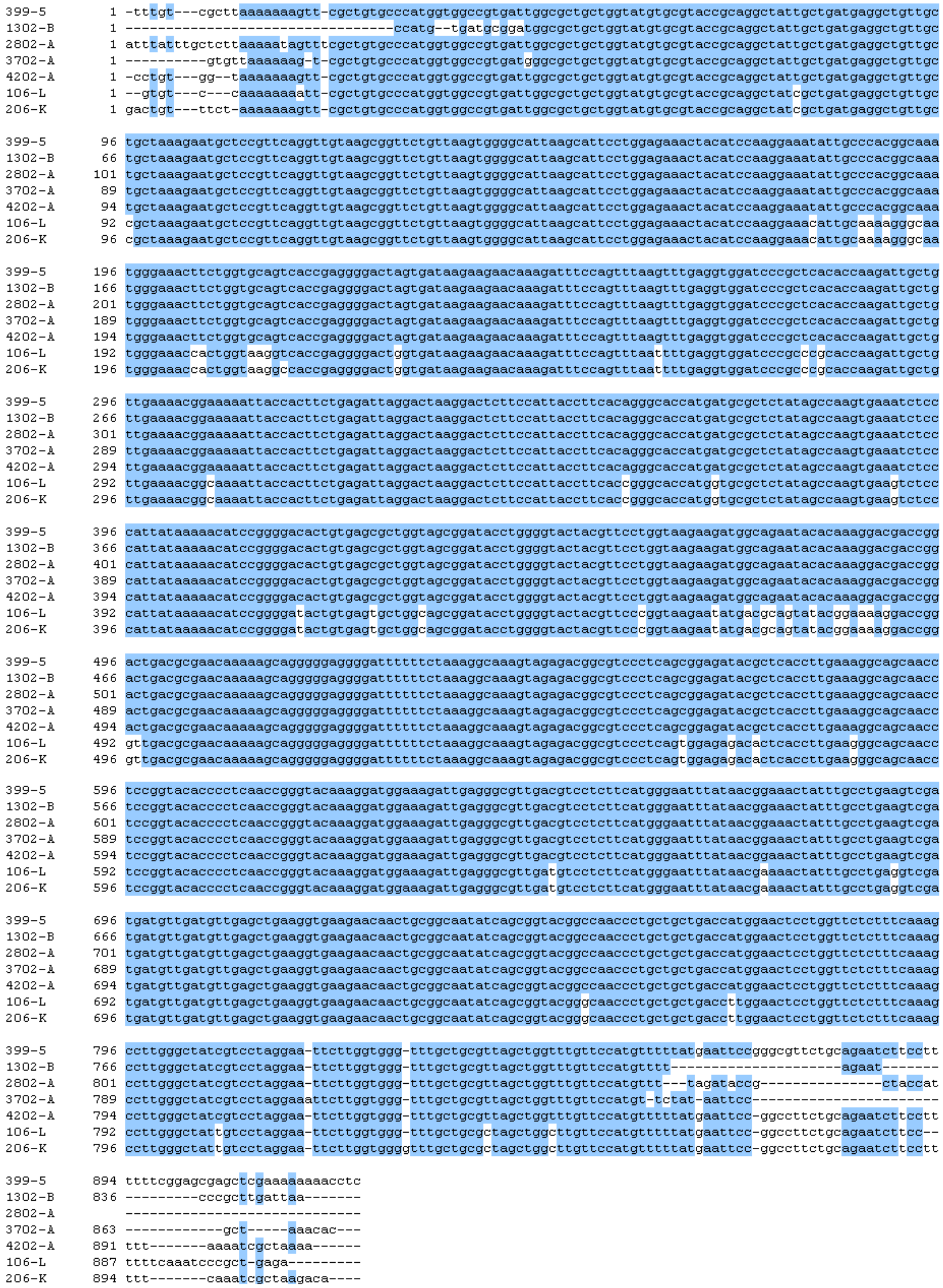
Primers were designed to screen for the presence of target gene Cp3995_0570 (*sprT*) that encodes a trypsin-like serine protease. Subsequently PCR products from seven strains were then sequenced by Eurofins MWG. Edited sequences shown were obtained from aligning both forward and reverse sequences. Edited sequences were then aligned using Clone Manager.



```
399-5       1 cgctaccgccgttgcagcaggttccctctccgccccagctatcgccgaggataaagaacaaacccctactgcttcctcccttgccagccttaaggacttg
1302-B      1 cgcta-cgccgttgcagcaggttccctctccgccccagctatcgccgaggataaagaacaaacccctactgcttcctcccttgccagccttaaggacttg
2802-A      1 cgctaccgccgttgcagcaggttccctctccgccccagctatcgccgaggataaagaacaaacccctactgcttcctcccttgccagccttaaggacttg
3702-A      1 -gctaccgccgttgcagcaggttccctctccgccccagctatcgccgaggataaagaacaaacccctactgcttcctcccttgccagccttaaggacttg
4202-A      1 cgctaccgccgttgcagcaggttccctctccgccccagctatcgccgaggataaagaacaaacccctactgcttcctcccttgccagccttaaggacttg
106-L       1 ctctaccgccgttgcagcaggttccctctccgcgccagctatcgctcaggacaaagcagaaacccctgcggcttcctcccttgccagccttaaggacttg
206-K       1 ctctaccgccgttgcagcaggttccctctccgcgccagctatcgctcaggacaaagcagaaacccctgcggcttcctcccttgccagccttaaggacttg

399-5     101 ctctcttcttcctcttcctccgaggacaccgctgcagacagcaccacgggcaccgctaccgcaaccgctaagaaaaaggacgctaccaagatctttgaag
1302-B    100 ctctcttcttcctcttcctccgaggacaccgctgcagacagcaccacgggcaccgctaccgcaaccgctaagaaaaaggacgctaccaagatctttgaag
2802-A    101 ctctcttcttcctcttcctccgaggacaccgctgcagacagcaccacgggcaccgctaccgcaaccgctaagaaaaaggacgctaccaagatctttgaag
3702-A    100 ctctcttcttcctcttcctccgaggacaccgctgcagacagcaccacgggcaccgctaccgcaaccgctaagaaaaaggacgctaccaagatctttgaag
4202-A    101 ctctcttcttcctcttcctccgaggacaccgctgcagacagcaccacgggcaccgctaccgcaaccgctaagaaaaaggacgctaccaagatctttgaag
106-L     101 ctctcttcttcctcttcctccgaggacaccgctgcagacagcaccacgggcacctctaccgcaaccgctaagaaaaaggacgctaccaagatctttgaag
206-K     101 ctctcttcttcctcttcctccgaggacaccgctgcagacagcaccacgggcacctctaccgcaaccgctaagaaaaaggacgctaccaagatctttgaag

399-5     201 atgtcaaagcatgggttggcatcatctctgctgtactaggcatcctcaccaccattatcac---------
1302-B    200 atgtcaaagcatgggttggcatcatctctgctgtactaggcatcctcaccaccattatcac---------
2802-A    201 atgtcaaagcatgggttggcatcatctctgctgtactaggcatcctcaccaccattatcac---------
3702-A    200 atgtcaaagcatgggttggcatcatctctgctgtactaggcatcctcaccaccattatcac---------
4202-A    201 atgtcaaagcatgggttggcatcatctctgctgtactaggcatcctcaccaccattatcaccttaagcgg
106-L     201 atgtcaaagcatgggttggcatcatctctgctgtactaggca-cctcaccaccat---------------
206-K     201 atgtcaaagcatgggttggcatcatctctgctgtactaggcatcctcaccaccattatcac---------
```

Figure 6.4: Comparison of Cp3995_1191 nucleotide sequences

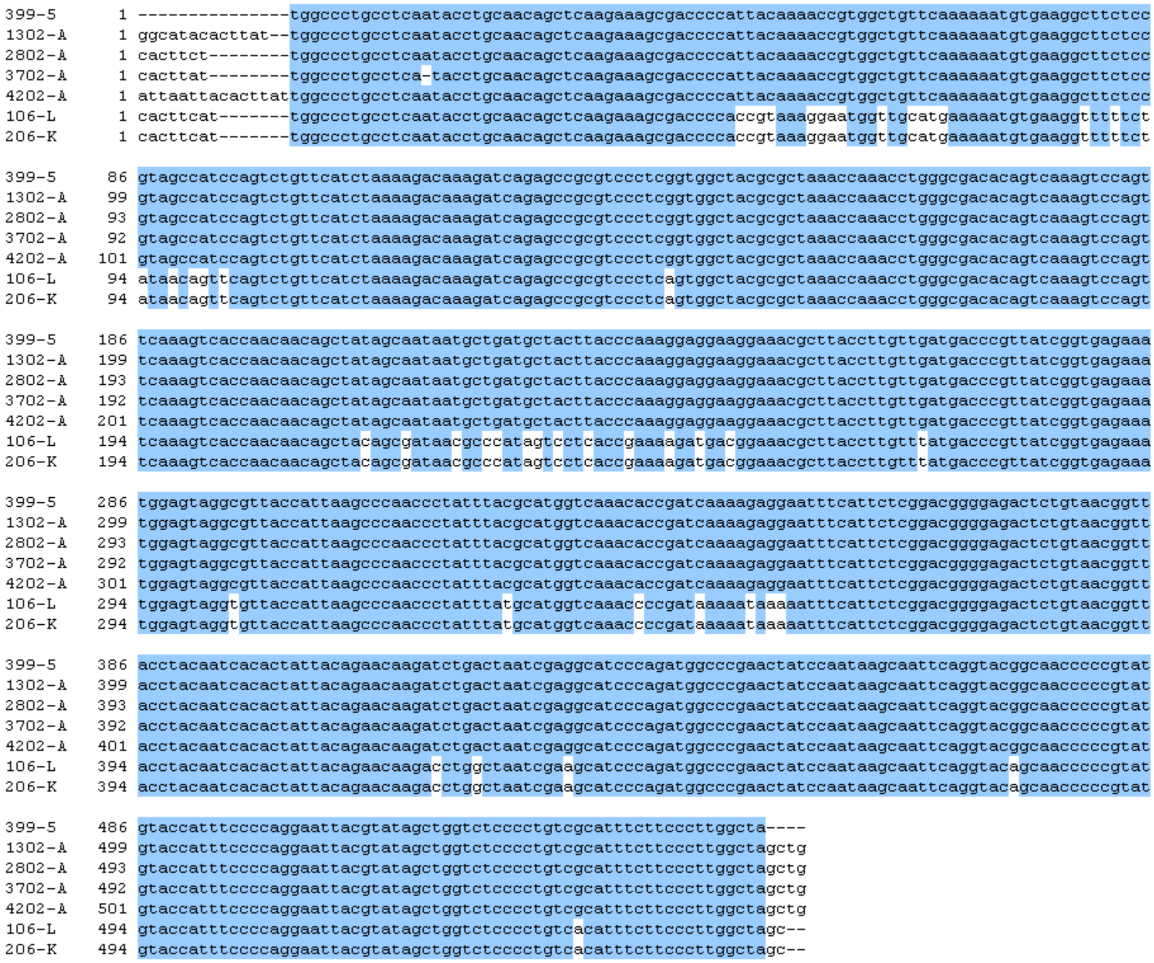
A panel of 50 isolates were screened for the presence of Cp3995_1191, a gene encoding a hypothetical protein containing a possible WXG motif. The PCR products from seven of these strains were then sequenced by Eurofins MWG. Edited sequences of the remaining seven shown here were obtained from aligning both forward and reverse sequences. The edited sequences were then aligned using Clone Manager and the image copied from this software.



```
399-5      1 -tttgt---cgcttaaaaaaagtt-cgctgtgcccatggtggccgtgattggcgctgctggtatgtgcgtaccgcaggctattgctgatgaggctgttgc
1302-B     1 ---------------------------------ccatg--tgatgcggatggcgctgctggtatgtgcgtaccgcaggctattgctgatgaggctgttgc
2802-A     1 atttatttgctcttaaaaatagtttcgctgtgcccatggtggccgtgattggcgctgctggtatgtgcgtaccgcaggctattgctgatgaggctgttgc
3702-A     1 ----------gtgttaaaaaag-t-cgctgtgcccatggtggccgtgatgggcgctgctggtatgtgcgtaccgcaggctattgctgatgaggctgttgc
4202-A     1 -cctgt---gg--taaaaaaagtt-cgctgtgcccatggtggccgtgattggcgctgctggtatgtgcgtaccgcaggctattgctgatgaggctgttgc
106-L      1 --gtgt---c---caaaaaaaatt-cgctgtgcccatggtggccgtgattggcgctgctggtatgtgcgtaccgcaggctatcgctgatgaggctgttgc
206-K      1 gactgt---ttct-aaaaaaagtt-cgctgtgcccatggtggccgtgattggcgctgctggtatgtgcgtaccgcaggctatcgctgatgaggctgttgc

399-5     96 tgctaaagaatgctccgttcaggttgtaagcggttctgttaagtggggcattaagcattcctggagaaactacatccaaggaaatattgcccacggcaaa
1302-B    66 tgctaaagaatgctccgttcaggttgtaagcggttctgttaagtggggcattaagcattcctggagaaactacatccaaggaaatattgcccacggcaaa
2802-A   101 tgctaaagaatgctccgttcaggttgtaagcggttctgttaagtggggcattaagcattcctggagaaactacatccaaggaaatattgcccacggcaaa
3702-A    89 tgctaaagaatgctccgttcaggttgtaagcggttctgttaagtggggcattaagcattcctggagaaactacatccaaggaaatattgcccacggcaaa
4202-A    94 tgctaaagaatgctccgttcaggttgtaagcggttctgttaagtggggcattaagcattcctggagaaactacatccaaggaaatattgcccacggcaaa
106-L     92 cgctaaagaatgctccgttcaggttgtaagcggttctgttaagtggggcattaagcattcctggagaaactacatccaaggaaacattgcaaaagggcaa
206-K     96 cgctaaagaatgctccgttcaggttgtaagcggttctgttaagtggggcattaagcattcctggagaaactacatccaaggaaacattgcaaaagggcaa

399-5    196 tgggaaacttctggtgcagtcaccgaggggactagtgataagaagaacaaagatttccagtttaagtttgaggtggatcccgctcacaccaagattgctg
1302-B   166 tgggaaacttctggtgcagtcaccgaggggactagtgataagaagaacaaagatttccagtttaagtttgaggtggatcccgctcacaccaagattgctg
2802-A   201 tgggaaacttctggtgcagtcaccgaggggactagtgataagaagaacaaagatttccagtttaagtttgaggtggatcccgctcacaccaagattgctg
3702-A   189 tgggaaacttctggtgcagtcaccgaggggactagtgataagaagaacaaagatttccagtttaagtttgaggtggatcccgctcacaccaagattgctg
4202-A   194 tgggaaacttctggtgcagtcaccgaggggactagtgataagaagaacaaagatttccagtttaagtttgaggtggatcccgctcacaccaagattgctg
106-L    192 tgggaaaccactggtaaggtcaccgaggggactggtgataagaagaacaaagatttccagtttaatttgaggtggatcccgcccgcaccaagattgctg
206-K    196 tgggaaaccactggtaaggccaccgaggggactggtgataagaagaacaaagatttccagtttaatttgaggtggatcccgcccgcaccaagattgctg

399-5    296 ttgaaaacggaaaaattaccacttctgagattaggactaaggactcttccattaccttcacagggcaccatgatgcgctctatagccaagtgaaatctcc
1302-B   266 ttgaaaacggaaaaattaccacttctgagattaggactaaggactcttccattaccttcacagggcaccatgatgcgctctatagccaagtgaaatctcc
2802-A   301 ttgaaaacggaaaaattaccacttctgagattaggactaaggactcttccattaccttcacagggcaccatgatgcgctctatagccaagtgaaatctcc
3702-A   289 ttgaaaacggaaaaattaccacttctgagattaggactaaggactcttccattaccttcacagggcaccatgatgcgctctatagccaagtgaaatctcc
4202-A   294 ttgaaaacggaaaaattaccacttctgagattaggactaaggactcttccattaccttcacagggcaccatgatgcgctctatagccaagtgaaatctcc
106-L    292 ttgaaaacggcaaaattaccacttctgagattaggactaaggactcttccattaccttcaccgggcaccatggtgcgctctatagccaagtgaagtctcc
206-K    296 ttgaaaacggcaaaattaccacttctgagattaggactaaggactcttccattaccttcaccgggcaccatggtgcgctctatagccaagtgaagtctcc

399-5    396 cattataaaaacatccggggacactgtgagcgctggtagcggatacctggggtactacgttcctggtaagaagatggcagaatacacaaaggacgaccgg
1302-B   366 cattataaaaacatccggggacactgtgagcgctggtagcggatacctggggtactacgttcctggtaagaagatggcagaatacacaaaggacgaccgg
2802-A   401 cattataaaaacatccggggacactgtgagcgctggtagcggatacctggggtactacgttcctggtaagaagatggcagaatacacaaaggacgaccgg
3702-A   389 cattataaaaacatccggggacactgtgagcgctggtagcggatacctggggtactacgttcctggtaagaagatggcagaatacacaaaggacgaccgg
4202-A   394 cattataaaaacatccggggacactgtgagcgctggtagcggatacctggggtactacgttcctggtaagaagatggcagaatacacaaaggacgaccgg
106-L    392 cattataaaaacatccggggatactgtgagtgctggcagcggatacctggggtactacgttcccggtaagaatatgacgcagtatacggaaaaggaccgg
206-K    396 cattataaaaacatccggggatactgtgagtgctggcagcggatacctggggtactacgttcccggtaagaatatgacgcagtatacggaaaaggaccgg

399-5    496 actgacgcgaacaaaaagcaggggggaggggatttttctaaaggcaaagtagagacggcgtccctcagcggagatacgctcaccttgaaaggcagcaacc
1302-B   466 actgacgcgaacaaaaagcaggggggaggggatttttctaaaggcaaagtagagacggcgtccctcagcggagatacgctcaccttgaaaggcagcaacc
2802-A   501 actgacgcgaacaaaaagcaggggggaggggatttttctaaaggcaaagtagagacggcgtccctcagcggagatacgctcaccttgaaaggcagcaacc
3702-A   489 actgacgcgaacaaaaagcaggggggaggggatttttctaaaggcaaagtagagacggcgtccctcagcggagatacgctcaccttgaaaggcagcaacc
4202-A   494 actgacgcgaacaaaaagcaggggggaggggatttttctaaaggcaaagtagagacggcgtccctcagcggagatacgctcaccttgaaaggcagcaacc
106-L    492 gttgacgcgaacaaaaagcaggggggaggggatttttctaaaggcaaagtagagacggcgtccctcagtggagagacactcaccttgaagggcagcaacc
206-K    496 gttgacgcgaacaaaaagcaggggggaggggatttttctaaaggcaaagtagagacggcgtccctcagtggagagacactcaccttgaagggcagcaacc

399-5    596 tccggtacacccctcaaccgggtacaaaggatggaaagattgagggcgttgacgtcctcttcatgggaatttataacggaaactatttgcctgaagtcga
1302-B   566 tccggtacacccctcaaccgggtacaaaggatggaaagattgagggcgttgacgtcctcttcatgggaatttataacggaaactatttgcctgaagtcga
2802-A   601 tccggtacacccctcaaccgggtacaaaggatggaaagattgagggcgttgacgtcctcttcatgggaatttataacggaaactatttgcctgaagtcga
3702-A   589 tccggtacacccctcaaccgggtacaaaggatggaaagattgagggcgttgacgtcctcttcatgggaatttataacggaaactatttgcctgaagtcga
4202-A   594 tccggtacacccctcaaccgggtacaaaggatggaaagattgagggcgttgacgtcctcttcatgggaatttataacggaaactatttgcctgaagtcga
106-L    592 tccggtacacccctcaaccgggtacaaaggatggaaagattgagggcgttgatgtcctcttcatgggaatttataacgaaaactatttgcctgaggtcga
206-K    596 tccggtacacccctcaaccgggtacaaaggatggaaagattgagggcgttgatgtcctcttcatgggaatttataacgaaaactatttgcctgaggtcga

399-5    696 tgatgttgatgttgagctgaaggtgaagaacaactgcggcaatatcagcggtacggccaaccctgctgctgaccatggaactcctggttctctttcaaag
1302-B   666 tgatgttgatgttgagctgaaggtgaagaacaactgcggcaatatcagcggtacggccaaccctgctgctgaccatggaactcctggttctctttcaaag
2802-A   701 tgatgttgatgttgagctgaaggtgaagaacaactgcggcaatatcagcggtacggccaaccctgctgctgaccatggaactcctggttctctttcaaag
3702-A   689 tgatgttgatgttgagctgaaggtgaagaacaactgcggcaatatcagcggtacggccaaccctgctgctgaccatggaactcctggttctctttcaaag
4202-A   694 tgatgttgatgttgagctgaaggtgaagaacaactgcggcaatatcagcggtacggccaaccctgctgctgaccatggaactcctggttctctttcaaag
106-L    692 tgatgttgatgttgagctgaaggtgaagaacaactgcggcaatatcagcggtacgggcaaccctgctgctgaccttggaactcctggttctctttcaaag
206-K    696 tgatgttgatgttgagctgaaggtgaagaacaactgcggcaatatcagcggtacgggcaaccctgctgctgaccttggaactcctggttctctttcaaag

399-5    796 ccttgggctatcgtcctaggaa-ttcttggtggg-tttgctgcgttagctggtttgttccatgtttttatgaattccgggcgttctgcagaatcttcctt
1302-B   766 ccttgggctatcgtcctaggaa-ttcttggtggg-tttgctgcgttagctggtttgttccatgtttt---------------------agaat-------
2802-A   801 ccttgggctatcgtcctaggaa-ttcttggtggg-tttgctgcgttagctggtttgttccatgttt---tagataccg---------------ctaccat
3702-A   789 ccttgggctatcgtcctaggaaattcttggtggg-tttgctgcgttagctggtttgttccatgt-tctat-aattcc-----------------------
4202-A   794 ccttgggctatcgtcctaggaa-ttcttggtggg-tttgctgcgttagctggtttgttccatgtttttatgaattcc-ggccttctgcagaatcttcctt
106-L    792 ccttgggctattgtcctaggaa-ttcttggtggg-tttgctgcgctagctggcttgttccatgtttttatgaattcc-ggccttctgcagaatcttcc--
206-K    796 ccttgggctattgtcctaggaa-ttcttggtggggtttgctgcgctagctggcttgttccatgtttttatgaattcc-ggccttctgcagaatcttcctt

399-5    894 ttttcggagcgagctcgaaaaaaaacctc
1302-B   836 ---------cccgcttgattaa-------
2802-A       -----------------------------
3702-A   863 ------------gct-----aaacac---
4202-A   891 ttt-------aaaatcgctaaaa------
106-L    887 ttttcaaatcccgct-gaga---------
206-K    894 ttt-------caaatcgctaagaca----
```

Figure 6.5: Comparison of Cp3995_1510 nucleotide sequences

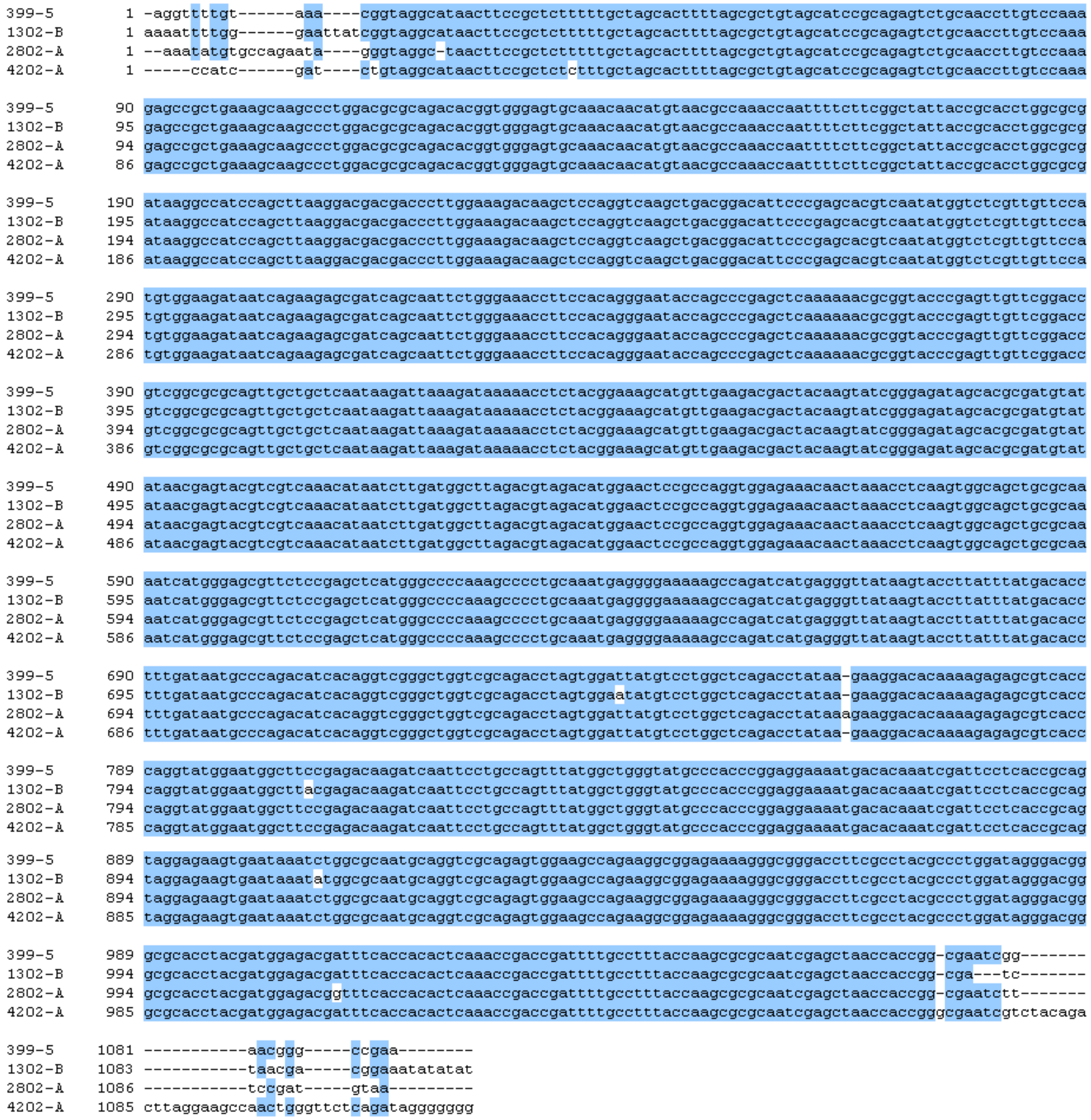
PCR products of seven *Cp* isolates screened for Cp3995_1510 (encoding the uncharacterised protein HtaC) were sequenced by Eurofins MWG. The nucleotide sequences were then edited using an alignment of forward and reverse sequences and the resulting sequences aligned using Clone Manger, the image generated by the software is reproduced here.



```
399-5      1 ---------------tggccctgcctcaatacctgcaacagctcaagaaagcgaccccattacaaaaccgtggctgttcaaaaaatgtgaaggcttctcc
1302-A     1 ggcatacacttat--tggccctgcctcaatacctgcaacagctcaagaaagcgaccccattacaaaaccgtggctgttcaaaaaatgtgaaggcttctcc
2802-A     1 cacttct--------tggccctgcctcaatacctgcaacagctcaagaaagcgaccccattacaaaaccgtggctgttcaaaaaatgtgaaggcttctcc
3702-A     1 cacttat--------tggccctgcctca-tacctgcaacagctcaagaaagcgaccccattacaaaaccgtggctgttcaaaaaatgtgaaggcttctcc
4202-A     1 attaattacacttattggccctgcctcaatacctgcaacagctcaagaaagcgaccccattacaaaaccgtggctgttcaaaaaatgtgaaggcttctcc
106-L      1 cacttcat-------tggccctgcctcaatacctgcaacagctcaagaaagcgaccccaccgtaaaggaatggttgcatgaaaaatgtgaaggtttttct
206-K      1 cacttcat-------tggccctgcctcaatacctgcaacagctcaagaaagcgaccccaccgtaaaggaatggttgcatgaaaaatgtgaaggtttttct

399-5     86 gtagccatccagtctgttcatctaaaagacaaagatcagagccgcgtccctcggtggctacgcgctaaaccaaacctgggcgacacagtcaaagtccagt
1302-A    99 gtagccatccagtctgttcatctaaaagacaaagatcagagccgcgtccctcggtggctacgcgctaaaccaaacctgggcgacacagtcaaagtccagt
2802-A    93 gtagccatccagtctgttcatctaaaagacaaagatcagagccgcgtccctcggtggctacgcgctaaaccaaacctgggcgacacagtcaaagtccagt
3702-A    92 gtagccatccagtctgttcatctaaaagacaaagatcagagccgcgtccctcggtggctacgcgctaaaccaaacctgggcgacacagtcaaagtccagt
4202-A   101 gtagccatccagtctgttcatctaaaagacaaagatcagagccgcgtccctcggtggctacgcgctaaaccaaacctgggcgacacagtcaaagtccagt
106-L     94 ataacagttcagtctgttcatctaaaagacaaagatcagagccgcgtccctcagtggctacgcgctaaaccaaacctgggcgacacagtcaaagtccagt
206-K     94 ataacagttcagtctgttcatctaaaagacaaagatcagagccgcgtccctcagtggctacgcgctaaaccaaacctgggcgacacagtcaaagtccagt

399-5    186 tcaaagtcaccaacaacagctatagcaataatgctgatgctacttacccaaaggaggaaggaaacgcttaccttgttgatgacccgttatcggtgagaaa
1302-A   199 tcaaagtcaccaacaacagctatagcaataatgctgatgctacttacccaaaggaggaaggaaacgcttaccttgttgatgacccgttatcggtgagaaa
2802-A   193 tcaaagtcaccaacaacagctatagcaataatgctgatgctacttacccaaaggaggaaggaaacgcttaccttgttgatgacccgttatcggtgagaaa
3702-A   192 tcaaagtcaccaacaacagctatagcaataatgctgatgctacttacccaaaggaggaaggaaacgcttaccttgttgatgacccgttatcggtgagaaa
4202-A   201 tcaaagtcaccaacaacagctatagcaataatgctgatgctacttacccaaaggaggaaggaaacgcttaccttgttgatgacccgttatcggtgagaaa
106-L    194 tcaaagtcaccaacaacagctacagcgataacgcccatagtcctcaccgaaaagatgacggaaacgcttaccttgtttatgacccgttatcggtgagaaa
206-K    194 tcaaagtcaccaacaacagctacagcgataacgcccatagtcctcaccgaaaagatgacggaaacgcttaccttgtttatgacccgttatcggtgagaaa

399-5    286 tggagtaggcgttaccattaagcccaaccctatttacgcatggtcaaacaccgatcaaaagaggaatttcattctcggacggggagactctgtaacggtt
1302-A   299 tggagtaggcgttaccattaagcccaaccctatttacgcatggtcaaacaccgatcaaaagaggaatttcattctcggacggggagactctgtaacggtt
2802-A   293 tggagtaggcgttaccattaagcccaaccctatttacgcatggtcaaacaccgatcaaaagaggaatttcattctcggacggggagactctgtaacggtt
3702-A   292 tggagtaggcgttaccattaagcccaaccctatttacgcatggtcaaacaccgatcaaaagaggaatttcattctcggacggggagactctgtaacggtt
4202-A   301 tggagtaggcgttaccattaagcccaaccctatttacgcatggtcaaacaccgatcaaaagaggaatttcattctcggacggggagactctgtaacggtt
106-L    294 tggagtaggtgttaccattaagcccaaccctatttatgcatggtcaaaccccgataaaaataaaaatttcattctcggacggggagactctgtaacggtt
206-K    294 tggagtaggtgttaccattaagcccaaccctatttatgcatggtcaaaccccgataaaaataaaaatttcattctcggacggggagactctgtaacggtt

399-5    386 acctacaatcacactattacagaacaagatctgactaatcgaggcatcccagatggcccgaactatccaataagcaattcaggtacggcaacccccgtat
1302-A   399 acctacaatcacactattacagaacaagatctgactaatcgaggcatcccagatggcccgaactatccaataagcaattcaggtacggcaacccccgtat
2802-A   393 acctacaatcacactattacagaacaagatctgactaatcgaggcatcccagatggcccgaactatccaataagcaattcaggtacggcaacccccgtat
3702-A   392 acctacaatcacactattacagaacaagatctgactaatcgaggcatcccagatggcccgaactatccaataagcaattcaggtacggcaacccccgtat
4202-A   401 acctacaatcacactattacagaacaagatctgactaatcgaggcatcccagatggcccgaactatccaataagcaattcaggtacggcaacccccgtat
106-L    394 acctacaatcacactattacagaacaagacctggctaatcgaagcatcccagatggcccgaactatccaataagcaattcaggtacagcaacccccgtat
206-K    394 acctacaatcacactattacagaacaagacctggctaatcgaagcatcccagatggcccgaactatccaataagcaattcaggtacagcaacccccgtat

399-5    486 gtaccatttccccaggaattacgtatagctggtctcccctgtcgcatttcttcccttggcta----
1302-A   499 gtaccatttccccaggaattacgtatagctggtctcccctgtcgcatttcttcccttggctagctg
2802-A   493 gtaccatttccccaggaattacgtatagctggtctcccctgtcgcatttcttcccttggctagctg
3702-A   492 gtaccatttccccaggaattacgtatagctggtctcccctgtcgcatttcttcccttggctagctg
4202-A   501 gtaccatttccccaggaattacgtatagctggtctcccctgtcgcatttcttcccttggctagctg
106-L    494 gtaccatttccccaggaattacgtatagctggtctcccctgtcacatttcttcccttggctagc--
206-K    494 gtaccatttccccaggaattacgtatagctggtctcccctgtcacatttcttcccttggctagc--
```

Figure 6.6: Comparison of Cp3995_1712 PCR products

Primers were designed to screen for the presence of target gene Cp3995_1712 (encoding a hypothetical protein), PCR products from seven strains were then sequenced by Eurofins MWG. Edited sequences shown were obtained from aligning both forward and reverse sequences. Edited sequences were then aligned using Clone Manager, and the image generated by the software is shown here.



```
399-5       1 -aggtttgt------aaa----cggtaggcataacttccgctctttttgctagcacttttagcgctgtagcatccgcagagtctgcaaccttgtccaaa
1302-B      1 aaaatttgg------gaattatcggtaggcataacttccgctctttttgctagcacttttagcgctgtagcatccgcagagtctgcaaccttgtccaaa
2802-A      1 --aaatatgtgccagaata----gggtaggc-taacttccgctctttttgctagcacttttagcgctgtagcatccgcagagtctgcaaccttgtccaaa
4202-A      1 -----ccatc------gat----ctgtaggcataacttccgctctctttgctagcacttttagcgctgtagcatccgcagagtctgcaaccttgtccaaa

399-5      90 gagccgctgaaagcaagccctggacgcgcagacacggtgggagtgcaaacaacatgtaacgccaaaccaatttcttcggctattaccgcacctggcgcg
1302-B     95 gagccgctgaaagcaagccctggacgcgcagacacggtgggagtgcaaacaacatgtaacgccaaaccaatttcttcggctattaccgcacctggcgcg
2802-A     94 gagccgctgaaagcaagccctggacgcgcagacacggtgggagtgcaaacaacatgtaacgccaaaccaatttcttcggctattaccgcacctggcgcg
4202-A     86 gagccgctgaaagcaagccctggacgcgcagacacggtgggagtgcaaacaacatgtaacgccaaaccaatttcttcggctattaccgcacctggcgcg

399-5     190 ataaggccatccagcttaaggacgacgacccttggaaagacaagctccaggtcaagctgacggacattcccgagcacgtcaatatggtctcgttgttcca
1302-B    195 ataaggccatccagcttaaggacgacgacccttggaaagacaagctccaggtcaagctgacggacattcccgagcacgtcaatatggtctcgttgttcca
2802-A    194 ataaggccatccagcttaaggacgacgacccttggaaagacaagctccaggtcaagctgacggacattcccgagcacgtcaatatggtctcgttgttcca
4202-A    186 ataaggccatccagcttaaggacgacgacccttggaaagacaagctccaggtcaagctgacggacattcccgagcacgtcaatatggtctcgttgttcca

399-5     290 tgtggaagataatcagaagagcgatcagcaattctgggaaaccttccacagggaataccagcccgagctcaaaaaacgcggtacccgagttgttcggacc
1302-B    295 tgtggaagataatcagaagagcgatcagcaattctgggaaaccttccacagggaataccagcccgagctcaaaaaacgcggtacccgagttgttcggacc
2802-A    294 tgtggaagataatcagaagagcgatcagcaattctgggaaaccttccacagggaataccagcccgagctcaaaaaacgcggtacccgagttgttcggacc
4202-A    286 tgtggaagataatcagaagagcgatcagcaattctgggaaaccttccacagggaataccagcccgagctcaaaaaacgcggtacccgagttgttcggacc

399-5     390 gtcggcgcgcagttgctgctcaataagattaaagataaaaacctctacggaaagcatgttgaagacgactacaagtatcgggagatagcacgcgatgtat
1302-B    395 gtcggcgcgcagttgctgctcaataagattaaagataaaaacctctacggaaagcatgttgaagacgactacaagtatcgggagatagcacgcgatgtat
2802-A    394 gtcggcgcgcagttgctgctcaataagattaaagataaaaacctctacggaaagcatgttgaagacgactacaagtatcgggagatagcacgcgatgtat
4202-A    386 gtcggcgcgcagttgctgctcaataagattaaagataaaaacctctacggaaagcatgttgaagacgactacaagtatcgggagatagcacgcgatgtat

399-5     490 ataacgagtacgtcgtcaaacataatcttgatggcttagacgtagacatggaactccgccaggtggagaaacaactaaacctcaagtggcagctgcgcaa
1302-B    495 ataacgagtacgtcgtcaaacataatcttgatggcttagacgtagacatggaactccgccaggtggagaaacaactaaacctcaagtggcagctgcgcaa
2802-A    494 ataacgagtacgtcgtcaaacataatcttgatggcttagacgtagacatggaactccgccaggtggagaaacaactaaacctcaagtggcagctgcgcaa
4202-A    486 ataacgagtacgtcgtcaaacataatcttgatggcttagacgtagacatggaactccgccaggtggagaaacaactaaacctcaagtggcagctgcgcaa

399-5     590 aatcatgggagcgttctccgagctcatgggccccaaagcccctgcaaatgaggggaaaaagccagatcatgagggttataagtaccttatttatgacacc
1302-B    595 aatcatgggagcgttctccgagctcatgggccccaaagcccctgcaaatgaggggaaaaagccagatcatgagggttataagtaccttatttatgacacc
2802-A    594 aatcatgggagcgttctccgagctcatgggccccaaagcccctgcaaatgaggggaaaaagccagatcatgagggttataagtaccttatttatgacacc
4202-A    586 aatcatgggagcgttctccgagctcatgggccccaaagcccctgcaaatgaggggaaaaagccagatcatgagggttataagtaccttatttatgacacc

399-5     690 tttgataatgcccagacatcacaggtcgggctggtcgcagacctagtggattatgtcctggctcagacctataa-gaaggacacaaaagagagcgtcacc
1302-B    695 tttgataatgcccagacatcacaggtcgggctggtcgcagacctagtggaatatgtcctggctcagacctataa-gaaggacacaaaagagagcgtcacc
2802-A    694 tttgataatgcccagacatcacaggtcgggctggtcgcagacctagtggattatgtcctggctcagacctataaagaaggacacaaaagagagcgtcacc
4202-A    686 tttgataatgcccagacatcacaggtcgggctggtcgcagacctagtggattatgtcctggctcagacctataa-gaaggacacaaaagagagcgtcacc

399-5     789 caggtatggaatggcttccgagacaagatcaattcctgccagtttatggctgggtatgcccacccggaggaaaatgacacaaatcgattcctcaccgcag
1302-B    794 caggtatggaatggcttacgagacaagatcaattcctgccagtttatggctgggtatgcccacccggaggaaaatgacacaaatcgattcctcaccgcag
2802-A    794 caggtatggaatggcttccgagacaagatcaattcctgccagtttatggctgggtatgcccacccggaggaaaatgacacaaatcgattcctcaccgcag
4202-A    785 caggtatggaatggcttccgagacaagatcaattcctgccagtttatggctgggtatgcccacccggaggaaaatgacacaaatcgattcctcaccgcag

399-5     889 taggagaagtgaataaatctggcgcaatgcaggtcgcagagtggaagccagaaggcggagaaaagggcgggaccttcgcctacgccctggatagggacgg
1302-B    894 taggagaagtgaataaatatggcgcaatgcaggtcgcagagtggaagccagaaggcggagaaaagggcgggaccttcgcctacgccctggatagggacgg
2802-A    894 taggagaagtgaataaatctggcgcaatgcaggtcgcagagtggaagccagaaggcggagaaaagggcgggaccttcgcctacgccctggatagggacgg
4202-A    885 taggagaagtgaataaatctggcgcaatgcaggtcgcagagtggaagccagaaggcggagaaaagggcgggaccttcgcctacgccctggatagggacgg

399-5     989 gcgcacctacgatggagacgatttcaccacactcaaaccgaccgattttgcctttaccaagcgcgcaatcgagctaaccaccgg-cgaatcgg-------
1302-B    994 gcgcacctacgatggagacgatttcaccacactcaaaccgaccgattttgcctttaccaagcgcgcaatcgagctaaccaccgg-cga---tc-------
2802-A    994 gcgcacctacgatggagacggtttcaccacactcaaaccgaccgattttgcctttaccaagcgcgcaatcgagctaaccaccgg-cgaatctt-------
4202-A    985 gcgcacctacgatggagacgatttcaccacactcaaaccgaccgattttgcctttaccaagcgcgcaatcgagctaaccaccgggcgaatcgtctacaga

399-5    1081 -----------aacggg-----ccgaa--------
1302-B   1083 -----------taacga-----cggaaatatatat
2802-A   1086 -----------tccgat-----gtaa---------
4202-A   1085 cttaggaagccaactgggttctcagatagggggg
```

Figure 6.7: Comparison of Cp3995_1947 PCR products

A panel of 50 isolates were screened for the presence of Cp3995_1947 (*cpp*) encoding the serine protease CP40; PCR products from four of these strains were then sequenced by Eurofins MWG. Edited sequences were obtained from aligning both forward and reverse sequences, and these edited sequences were then aligned using Clone Manager software.

6.2.3 Amplification and cloning of target genes

The target genes were successfully amplified by PCR using Hot Start DNA Polymerase (Novagen) and the primers in **Table 2.6**. The PCR products were visualised by gel electrophoresis (**Fig. 6.8**) and then extracted from the gel and cloned into blunt-end vector pCR®II-TOPO®. The resulting construct was digested with the appropriate restriction endonuclease combinations as indicated in **Table 2.6**. The DNA fragment produced (containing target gene) was subsequently ligated into a similarly digested pET-15b vector (**Table 2.2**). Cloning of all targets into pET-15b was confirmed by restriction digest. Constructs were also successfully made using Champion™ pET SUMO.

6.2.4 Expression and purification of recombinant proteins

The pET-15b/target and Champion™ pET SUMO/target constructs were used to transform One Shot® BL21(DE3) *E. coli* (Invitrogen). For each construct, LB broths were inoculated with the transformed *E. coli* and cultured at each of three temperatures; 26°C, 30°C and 37°C. Once cultures reached an OD_{600nm} of *ca*. 0.6, recombinant protein expression was induced using 1.0 mM IPTG. The pET-15b/target and Champion™ pET SUMO/target constructs were also used to transform One Shot® BL21(DE3)pLysS *E. coli* (Invitrogen), and again cultured at the three temperatures and induced with 1.0 mM IPTG. Furthermore, all of the cultures described above were repeated as previously but using 0.5 mM IPTG to induce recombinant protein expression.

General observations were made that the cultures with induced expression of Cp3995_0167 and Cp3995_1712 grew much slower than other cultures, however, time points were not taken. Attempted expression of these proteins was unsuccessful, in fact of all of the targets, only products of Cp3995_0570 (*sprT*) and Cp3995_1947 (*cpp*) were successfully over-expressed. Both were over-expressed by One Shot® BL21(DE3) *E. coli* that had been transformed with pET-15b/target constructs, following incubation at 37°C and protein induction by 1.0 mM IPTG. Lower temperatures and IPTG concentration did not improve expression levels of the two recombinant proteins. Cp3995_0570 and Cp3995_1947 were confirmed to be the desired targets initially by Western blot using anti-His (C-Term)-HRP antibody but also by Matrix-assisted laser desorption/ionization (MALDI) analysis.

Purification of the recombinant proteins was performed using nickel affinity columns, initially under native conditions. However, purification using native conditions was unsuccessful regardless of both the temperature the transformants had been cultured at and concentration of IPTG used to induce protein expression. The same purification technique was then used but with urea-based buffers. These denaturing conditions resulted in successful enrichment of both recombinant proteins and the purified proteins are shown in a Western blot using anti-His (C-Term)-HRP antibody in **Fig. 6.9**.

.

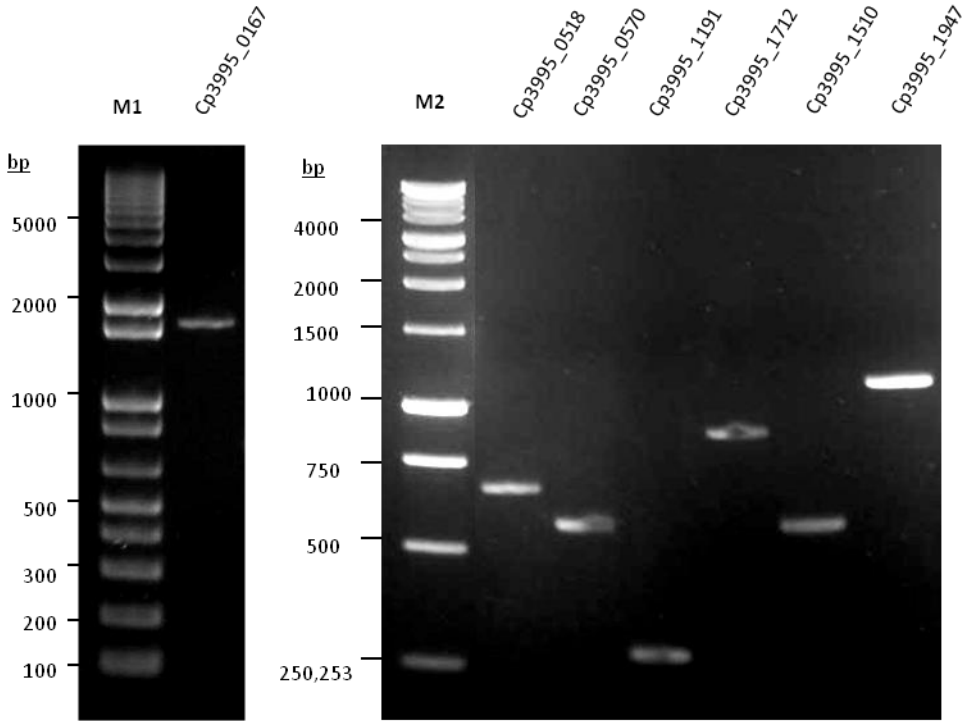


Figure 6.8: Amplification of target genes

Genes encoding the target proteins were amplified by PCR using primers designed with restriction endonuclease recognition sites at the start for cloning purposes. DNA marker ladders used were Invitrogen's Trackit™ 1 Kb Plus DNA Ladder (**M1**) and Promega's 1Kb DNA Ladder (**M2**).

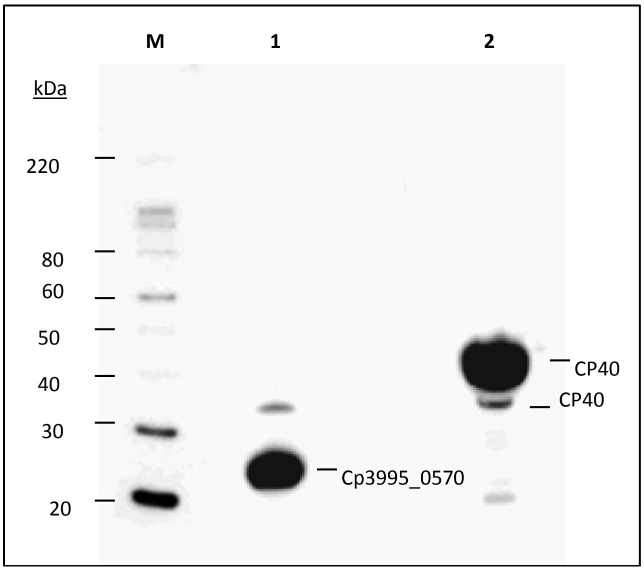


Figure 6.9: Western blot of the recombinant proteins

Target protein expression was induced in transformed *E. coli* BL21 (DE3) using IPTG and expressed proteins bound to a nitrocellulose membrane and probed with anti-His (C-Term)-HRP antibody (1:5,000). Recombinant proteins were subsequently purified using nickel affinity columns. Sample lanes correspond to *E. coli* transformed with the pET-15b/Cp3995_0570 construct (Lane **1**) or the pET-15b/*cpp* construct (Lane **2**) and the molecular weight marker (**M**) was Magic Mark™ XP Western Protein Standard (Invitrogen).

6.2.5 Seroreactivity of target proteins

The seroreactivity of the recombinant target proteins was assessed by ELISA. Each recombinant protein was used to coat an ELISA plate which was then probed with a panel of sheep sera. The sheep sera panel consisted of ten biological replicates of ovine sera from sheep with CLA (CLA positive) and ten from sheep naïve to *Cp* (CLA negative). Sera had previously been defined as CLA positive or negative as determined by PLD and whole-cell ELISA (Flockhart 2011, unpublished data). The trypsin-like protease (Cp3995_0570) did not distinguish between negative and positive sera and the optical density (450 nm) varied greatly between biological replicates in each group (**Figs. 6.10**). A western blot of the protein was performed using three negative and three positive ovine sera. Single bands were visible at approximately 23 kDa for each serum used (**Fig. 6.11**).

The CP40 recombinant protein also failed to distinguish between CLA negative and positive sera (**Fig. 6.10**). There is also a large standard deviation of the biological replicates for both groups of sera, although this is not as great as for the previous recombinant.

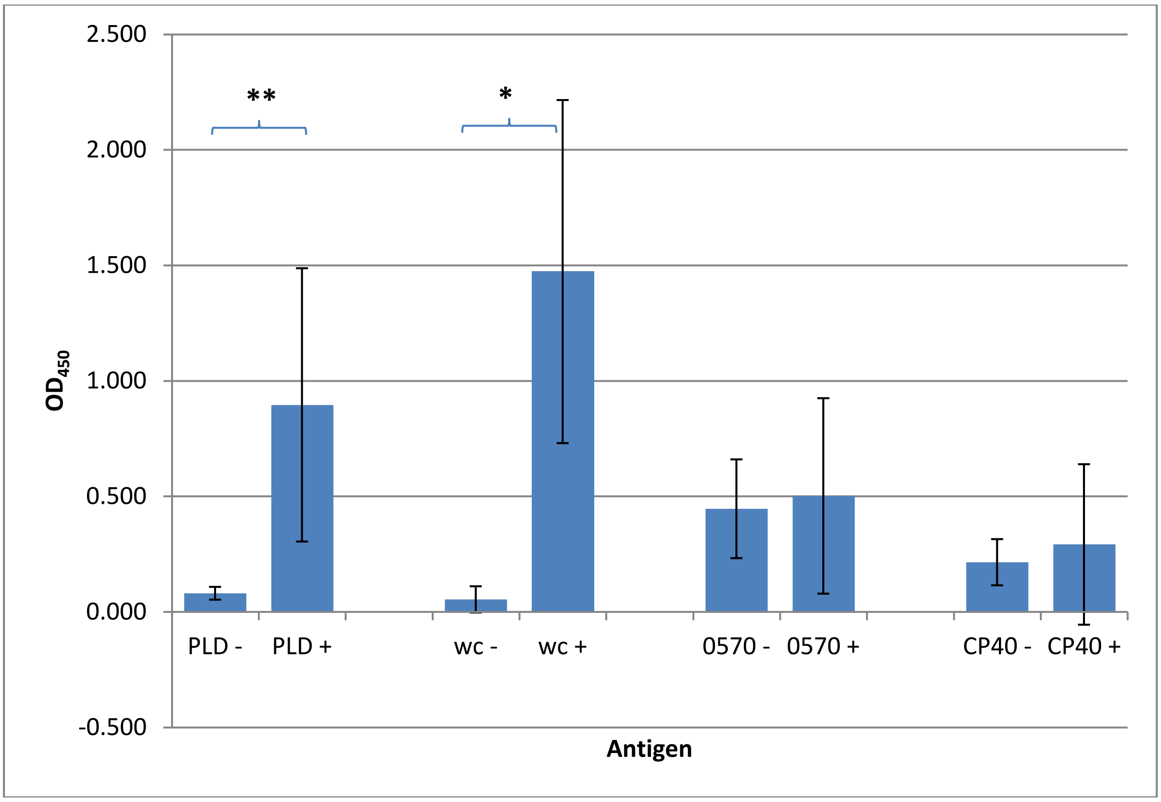


Figure 6.10: Recombinant protein ELISA data using a panel of ovine sera

Seroreactivity of the recombinant proteins Cp3995_0570 (**0570**) and CP40 were assessed by ELISA using ovine sera from both sheep with CLA or sheep naïve to *Cp*. Sera used had previously been defined as CLA positive or negative by both PLD and whole cell (**wc**) ELISA and this data is replicated here (Flockhart 2011, unpublished data). The mean absorbance at an OD of 450 nm is shown and the standard deviation of ten biological replicates for each CLA positive (+) and negative (-) sera is indicated by the error bars. Statistical significance between the positive and negative groups for each antigen was determined using a t-test and is indicated; ** represents $P \leq 0.005$ and * represents $P \leq 0.02$.

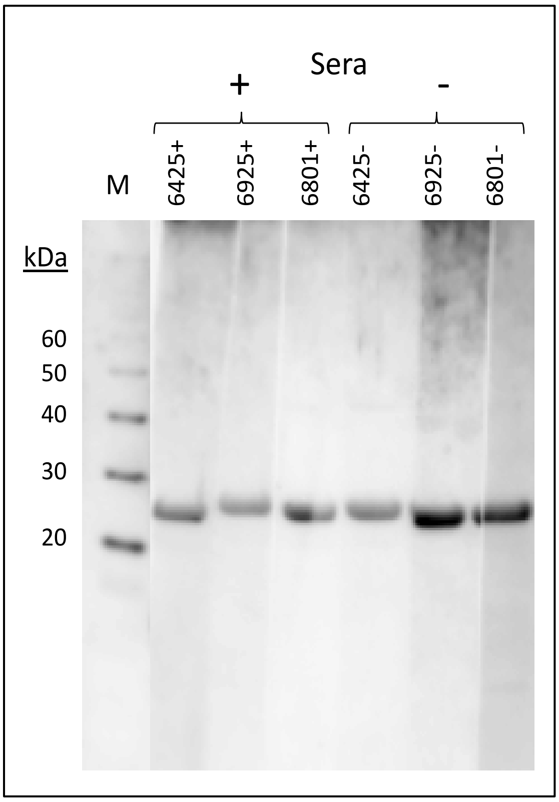


Figure 6.11: Western blot of analysis of recombinant protein Cp3995_0570

Primary antibodies to probe the membrane-bound protein were either serum (1:200) from sheep with CLA (+) or sheep naïve to *Cp* (-). Although there appear slight discrepancies in the size of the protein bands, this is actually due to the membrane being cut into strips for primary antibody probing before being re-assembled. Sample lanes correspond to individual sheep serum and the molecular weight marker used (**M**) was Magic Mark™ XP Western Protein Standard (Invitrogen).

6.3 Discussion

The targets chosen were all extracellular or membrane-bound as these are the proteins that are exposed to the host's immune system at initial. Three of the targets, Cp3995_0167, Cp3995_0518 and Cp3995_1712 are hypothetical proteins with unknown functions. Hypothetical proteins are always interesting to investigate further as they are novel and reports of studies on hypothetical proteins have implicated them in roles such as antibiotic resistance (Nan *et al.*, 2009) and lead to the identification of possible vaccine candidates (Bashir *et al.*, 2010). Cp3995_1191 is also a hypothetical protein but contains a possible WXG motif (amino acid sequence Trp-Xaa-Gly), which is a signature sequence of ESAT-6-like proteins (Pallen, 2002). *Mycobacterium tuberculosis* secretes ESAT-6, a virulence factor that triggers cell-mediated immune responses (Burts *et al.*, 2005). Distant homologues of ESAT-6 have been identified in several Gram-positive bacteria including *Bacillus spp*., *Staphylococcus aureus* and *Clostridium acetobutylicum* (Pallen, 2002). Two serine proteases were included in the study, one, CP40, has been previously studied and is known to be immunogenic, the other is a trypsin-like serine protease encoded by *sprT*. This trypsin-like protease contains a Hap active site; the Hap or 'Haemophilus adhesion and penetration' family contains proteins concerned with the interaction of epithelial cells (St Geme, de la Morena and Falkow, 1994). Serine proteases have a range of roles in many biological processes but are also necessary for virulence in many bacterial pathogens, such as *M. tuberculosis* and *Streptococcus suis* (Upadhye *et al*., 2009; Bonifait *et al.,* 2010).

The possible use of serine proteases in diagnostics has been studied in other Gram-positive organisms; Wankhade *et al.* (2005) found that a chymotrypsin-like serine protease of *M. tuberculosis* has lipase activity and concluded the secreted antigen to have drug target potential. The remaining protein target, Cp3995_1510 (HtaC), is an uncharacterised protein encoded by a gene which is part of a cluster concerned with utilising haemin as an iron source. Of the encoded proteins identified as extracellular by combined bioinformatic and proteomic analysis in **section 5.2.1.3**, Cp3995_1510 was the only gene that was found to be present in all *Cp* isolates but not in *C. ulcerans*.

For use in a successful diagnostic, ideally there would be no cross-reactivity of the protein with sera from hosts infected with any other bacterial species than *Cp*, so preferably the target would be conserved amongst all *Cp* isolates but absent from the rest of the genus.

The conservation of the of the target genes was assessed using a presence/absence analysis determined by PCR amplification in a panel of 50 *Cp* isolates but also in *C. ulcerans* strain NCTC 12077. With the exception of *cpp*, the target genes were found to be present in all *Cp* isolates screened. This confirms the genes to be conserved in this organism irrespective of host species or geographical origin. Two of the targets (Cp3995_0167 and Cp3995_0518) were also conserved in the *C. ulcerans* isolate. Originally these genes were chosen to study due to their lack of homology with other Corynebacteria (as indicated by the absence of these organisms in BLAST searches conducted in Chapter Three). The presence of these genes in *C. ulcerans* but not *C. diphtheriae* reflects the closer phylogenetic relationship between *Cp* and *C. ulcerans* than either of these organisms with *C. diphtheriae*. Although Cp3995_0167 and Cp3995_0518 were found in *C. ulcerans*, they may still have some potential for use as a diagnostic as it would be very rare to find a *C. ulcerans* infection in a ruminant, as *C. ulcerans* usually infects cats, humans and cattle (Hommez *et al.*, 1999; de Zoysa *et al.*, 2005) but has never been reported in an ovine or caprine host.

Despite repeated attempts to amplify *cpp* (Cp3995_1947) using two different primer pairs from colony DNA preps and phenol gDNA extractions, the gene could not be amplified in a significant amount of isolates and therefore were assumed absent from these genomes. Significantly, there are no reports in the literature of other *Cp* isolates lacking this gene. These results question the use of this gene in the development of a diagnostic tool or vaccine; if some strains truly lack *cpp*, a vaccine/diagnostic based on this gene will not be effective in those strains. However, of more concern, analysis of the *Cp* 1/06-A and *C. ulcerans* NCTC 12077 genomes sequenced in Chapter Four revealed that the *cpp* gene is actually present in these genomes despite the fact that the gene could not be amplified here. The primers used for amplification were designed from the genome sequences and one primer pair was specifically designed in a conserved region of identical sequence between *C. ulcerans* and *Cp* 1/06-A. These results suggest that the current screening of *cpp* as unreliable, and therefore no conclusions could be made from the screening of this gene. Due to the inconclusive nature of this data, attempts were still made to express *cpp* as a recombinant protein.

For each target, the PCR products of eight Corynebacterial isolates were sequenced and, once aligned, these sequences indicated the degree of homology of the target proteins in different strains. The eight isolates consisted of the single *C. ulcerans* strain and seven *Cp* strains chosen to represent a range of both different host species and countries of origin.

The sequences for any given target were highly homologous, with the vast majority of differences being in the bovine and equine sequences. For all targets, the bovine and equine sequences are very similar with single nucleotide polymorphisms (SNPs) in one of these two isolates being the same in the other, despite differing from all other isolates. In this respect the target sequences show a clear difference of isolates belonging to the distinct biovars of *ovis* and *equi*. PCR products of all eight *Cp* isolates were successfully sequenced for each of the targets (except *cpp,* for which only four were amplified). Although two of the targets were amplified from *C. ulcerans*, only that of Cp3995_0518 could be sequenced and this sequence is identical to the ovine and caprine *Cp* isolates, however, again there are several SNPs in the equine and bovine sequences. The sequencing of Cp3995_0167 from *C. ulcerans* was unsuccessful despite this gene being present in the genome. The reason for this is unclear; one possibility is that a secondary structure, such as hairpin formation, caused the sequencing reaction to fail. Also in this study target genes were sequenced using the same primers that were used to amplify the genes by PCR. However, ideally primers for amplification would have been designed outside the gene itself and sequencing primers inside this, targeting the whole gene and allowing the sequencing of the complete gene on both forward and reverse strands. This would have abolished the problem currently experienced with degeneration of the sequences towards the start and stop codons. That said the use of the same primers for amplification and sequencing utilised a resource already available for a second time and so reduced time and financial costs. Also the current strategy was sufficient for confirming the PCR products were orthologues of the target, and not a non-specific product which by coincidence happened to be of a similar size (bp) to the desired target.

The targets were amplified and cloned into pET-15b and pET SUMO and the constructs checked by restriction digest which confirmed the inserts and that the targets had been cloned successfully. The constructs were then used to chemically transform One Shot® BL21(DE3) and One Shot® BL21(DE3)pLysS *E. coli*. Target expression was then induced by addition of IPTG. Expression and purification of recombinant protein targets proved challenging and only two of the targets, Cp3995_0570 and CP40, were over-expressed despite using two different vectors and expression strains, as well as several different incubation conditions. Lack of expression could have resulted for a number of reasons. Although constructs were checked by restriction digest, this only confirms the presence and direction of the insert and could not rule out other problems during the cloning process such as a the possibility of introducing a stop codon within the gene. Had the insertion site of the constructs been sequenced this would have offered more insight into the reasoning behind the lack of protein expression. If time constraints allowed, it would have been worthwhile to use different expression hosts as some proteins express better in certain hosts. The target proteins may have been harmful/toxic to the *E. coli* expression strains used here, perhaps the slow growth noted in Cp3995_0167 and Cp3995_1712 cultures is due to the recombinant proteins having a toxic effect on the host's growth. It is also possible that more of the targets were expressed but that expression was on such a low level that staining with Simply Blue™ Safe Stain (Invitrogen) did not visually distinguish the recombinant protein from the proteome of the expression host. Such low expression could have been confirmed using LC-ESI-MS/MS to analyse the proteins run through a lane of an SDS-PA gel.

Initially purification of recombinant proteins CP40 and Cp3995_0570 was unsuccessful; however upon further analysis the proteins were discovered to be present in the cell debris pellet post cell lysis. This indicated the proteins to be insoluble and a likely explanation is that the proteins were deposited in the cytoplasm of the bacteria as aggregates or inclusion bodies. This is often a complication of *E. coli* expression systems. Changing the temperature and/or the concentration of IPTG used did nothing to improve the solubility of either recombinant protein. However, both CP40 and Cp3995_0570 were purified using urea. The buffers will have denatured the proteins and it is unlikely that the proteins re-folded. Denaturing recombinant proteins like this is not ideal as denaturing a protein could destroy conformational epitopes. If these epitopes were destroyed, antibodies in the sera used to probe the recombinant in the ELISAs may not bind to the protein. This would suggest a lack of immune recognition which would not actually be the case *in vivo*. Despite the results seen here, CP40 has previously been expressed as a recombinant protein with enzymatic activity. Wilson, Brandon and Walker (1995) used plasmid vector pGEX2 to express CP40 as a fusion protein and successfully purified by the method of Frangioni and Neel (1993). If this vector had been available, recombinant CP40 could have been produced using this method.

The seroreactivity of the recombinant proteins was assessed using a small panel of ten CLA negative and ten positive sheep sera. The sera had originated from CLA-free and CLA-infected sheep and were previously confirmed as negative and positive respectively using both PLD and whole-cell ELISAs (Flockhart 2011, unpublished data). The difference between OD readings for sera from *Cp*-naïve and *Cp*-infected animals was significant using both PLD and whole cell ELISAs. Recombinant PLD ELISA distinguished between the sera types with a probability of $P \leq 0.005$ and the whole cell ELISA with $P \leq 0.02$. However, the trypsin-like protease encoded by *sprT* did not distinguish between negative and positive sera, and the data had a very large standard deviation as a result of a wide range of values, particularly for the positive sera. Hence the data shows that this protein would not be suitable for use as a diagnostic. However, as denaturation would destroy conformational epitopes, which are only exposed if the protein is correctly folded, it remains unknown whether this would be true of the protein in a native state. To confirm that it was the target protein that was causing any differences in OD and not non-specific readings resulting from the ELISA process (such as the blocking agent), a western blot was performed. A representative subset of three positive and three negative sera were used and bands were observed at approximately 23 kDa for all six sera. Although two of the positive sera appear to have slightly more intense bands, this method (unlike the ELISAs) is not quantitative and so cannot be used to draw conclusions regarding this. The western blot does confirm that antibodies in the positive and negative sera are reacting to the serine protease. The inability of this trypsin-like protease to distinguish between CLA negative and positive sera could be due to pre-exposure of the antigens of this protein. The only sequence of relevance in the public databases with homology to Cp3995_0570 is a hypothetical protein of *Corynebacterium bovis* DSM 20582. The sheep from which sera used in the current study originated may have been exposed to *Corynebacterium bovis* in the farm environment, as it is a pathogenic bacterium that causes mastitis and pyelonephritis in cattle (Watts *et al.,* 2000).

Interestingly in the current study, CP40 also did not distinguish between the negative and positive sera and OD (450 nm) values for this protein again had a large standard deviation (although not as great as for Cp3995_0570). The CP40 sequence has homology to other bacteria including some to *Entercoccus faecalis.* Hence it is also possible that sheep had been previously exposed to CP40 antigens as *Entercoccus faecalis* for example can be found in soil, water and plants (Kayser, 2003), and has been reported to cause mastitis in sheep (Sanciu *et al.,* 2012). Despite the results observed here, Frangioni and Neel (1993) found CP40 to be immunogenic. It may be that had the protein been produced as a recombinant protein by their previously described method, the ELISA could have indicated the native protein suitable as a diagnostic.

In the current work the genome was used to attempt to identify novel diagnostic targets, however, there are alternative ways this could be approached; indeed other methods have been described to identify diagnostic candidates. For example, the antigenic serine protease CP40 was identified using antibody-secreting cells (ASC). ASC probes (generated by culture of ASC obtained from the local drainage lymph node at the site of infection) displayed specificity restricted to a CP40 antigen (Walker *et al.*, 1994). Also proteomics could have been used to identify immunogenic proteins by 2-dimensional polyacrylamide gel electrophoresis (2D-PAGE) and immunoblot using sera from *Cp*-infected animals. (This could have been done by probing the secreted, or even whole, *Cp* proteome). The antigens would then be identified by mass spectrophotometry such as matrix-assisted laser desorption/ionization mass spectrophotometry (MALDI-MS). DelVecchio *et al.* (2006) used this approach to conduct an immunoproteomic analysis of the spore antigens of the *Bacillus cereus* group and successfully identified differentially expressed and immunogenic spore proteins. A further study to employ 2D-PAGE followed by mass spectrophotometry aimed to identify proteins with potential diagnostic value for bovine paratuberculosis (Cho *et al.*, 2006). The authors reported the 2D-PAGE separation of proteins of potential interest from *Mycobacterium paratuberculosis* culture filtrates. The electrophoresis yielded 240 protein spots, 83% of which reacted with serum from *M. paratuberculosis*-infected cattle in immunoblots, however only 15% reacted with bovine serum that had been absorbed with *Mycobacterium phlei* antigens. Subsequently, 24 protein spots were selected for identification; 14 proteins were identified. These proteins were concluded to be strong candidates for use in an improved serodiagnostic test for bovine paratuberculosis (Cho *et al.*, 2006).

Another technique that has been used to identify diagnostic candidate proteins from bacterial pathogens is phage display, a technology first described by Smith in 1985 (Smith, 1985). The method displays recombinant peptides (or proteins) on the surface of a bacteriophage, and the peptides/proteins can then be selected for by allowing the phage to interact with immobilised ligands (Mullen *et al.*, 2006). Foreign DNA sequences are fused to the phage genome so that the foreign protein is expressed as a fusion with one of the viral coat proteins which can offer an efficient way of purifying and characterising proteins (Mullen *et al.*, 2006).

With regards to the current work, the most relevant report of phage display usage was by Liu *et al.* (2011). This group constructed a whole genomic surface protein phage display library of *M. tuberculosis* to allow the direct selection, identification, expression, purification and functional research of the surface proteins from this organism. The phage display library was preadsorbed with sera from BCG-vaccinated individuals to reduce cross-reactivity of antibodies and selective screening was then performed to enrich for clones specific to TB sera. Using this phage display approach, the authors identified six antigens, three of which had not previously been reported as diagnostic antigens (Liu *et al.*, 2011).

Despite the techniques described above relying on sera recognition of proteins, a protein does not necessarily need to be immunogenic to be useful as a diagnostic, such as in the case of a biomarker. A biomarker (or biological marker) is a general term for an indicator of a biological state and could, for example, be the presence or measure of a hormone, cytokine, cell or protein. A biomarker may be indicative of the type of infection present (e.g. bacterial or viral) or it may be used as a diagnostic to identify a causative organism of disease (Chapula *et al.,* 2011; Krishnamurthy *et al.,* 1996). The identification of biomarkers of several bacterial species by matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF MS) analysis has been described (Krishnamurthy *et al.,* 1996). In a study by Krishnamurthy *et al.* (1996), proteins were isolated from whole cells of *Brucella melitensis, Yersinia pestis* and several *Bacillus* isolates and analysed by MALDI-TOF MS. The mass spectrometric profiles generated showed specific biomarkers for individual bacteria that enabled detection of the pathogenic bacteria *Bacillus anthracis, Y. pestis* and *B. melitensis* and also distinguished the organisms from corresponding non-pathogenic species. Using multiple strains of the *Bacillus* species *B. anthracis, B. thurgiensis, B. cereus* and *B. subtilis,* both genus and strain specific biomarkers were derived from the molecular mass of the intact proteins (Krishnamurthy *et al.,* 1996).

Another study that identified specific protein biomarkers for differentiation of closely-related bacterial strains was reported using a top-down liquid chromatography/mass spectrometry (LC/MS) approach (Williams *et al.,* 2005). The expressed protein profiles of two closely related shiga-toxigenic *E. coli* serotype 0157:H7 strains and the nonpathogenic *E. coli* K-12 strain were compared. A number of potential biomarkers were identified, including a cytolethal distending toxin unique to one of the 0157:H7 strains. A specific PCR for the detection of this protein biomarker was developed using its sequence information to derive the genetic sequence, and the PCR allowed identification of the strains (Williams *et al.*, 2005). The authors also report the use of this technique to identify additional unique biomarkers which differentiated nearly identical 0157:H7 strains (Williams *et al.*, 2005).

The MS methods described above have been successful in identifying a biomarker for the identification of several bacterial pathogens, including Gram positive species. Applying this methodology to *Cp* could offer a good starting point to a further alternative approach in the development of a novel and/or improved CLA diagnostic tool.

In the current work no targets suitable for mass production and use in a diagnostic tool were found. All proteins studied were either not expressed as recombinants or failed to distinguish between animals with CLA and those naïve to *Cp*. Although unfortunately as it stands both of the recombinant proteins produced here are not suitable for use in a diagnostic, if the proteins could be purified without denaturing them, this may not be the case. Recombinant protein production proved complex and highly time consuming, had more time been available, more troubleshooting could have been performed and more of the targets successfully expressed and purified. Further work would then be necessary to determine whether any of the targets proposed here could have real diagnostic potential.

If resources allowed, it would also be beneficial to apply other methods to identifying proteins with potential diagnostic applications. Proteomics such as 2D analysis of the *Cp* proteome could be used in conjunction with the *Cp* genome to better approach the identification of proteins to investigate in a possible diagnostic tool.

Chapter Seven: General Discussion

In recent years, CLA has become endemic in the UK and is becoming of increasing concern due to the both the animal welfare issues but also the financial costs to the sheep-producing industry. Control of CLA has proven somewhat complex and antibiotic treatment of the disease is generally considered to be refractory, whilst vaccination options are limited, particularly in the UK. Furthermore understanding of the pathogenesis of *Cp*, remains insufficient and molecular characterisation of *Cp* prior to this project had been rudimentary; only three virulence factors, namely PLD, CP40 and mycolic acid have been described in the literature. However, since the genomic era there have been advances in the field of bacteriology and genomics has been utilised to aid understanding of bacteria including pathogens of humans and animals. NGS platforms have dramatically reduced the costs of sequencing and made the acquisition of a whole bacterial genome sequence greatly more attainable. The study of such a genome sequence can provide a wealth of information on the lifestyle of an organism and the genes involved in important attributes such as virulence, host adaptation and stress response. Whole genome sequences have been used in numerous studies to, for example, characterise bacteria and discover vaccine candidates leading to novel vaccine development (Pizza *et al.*, 2000).

Hence the initial aim in this thesis was to complete the genome sequence of the ovine *Cp* 3/99-5 isolate from the Scottish Borders which could (once thoroughly annotated) go on to represent the reference organism in the UK. The *Cp* 3/99-5 isolate was sequenced to a very high quality (**section 3.2.1**) with a read depth of greater than 40 and the sequence was almost completely closed as part of this project with only three gaps remaining. A thorough annotation of the genome was performed by compiling output generated from several web-based tools followed by manual curation in Artemis sequence viewer. Importantly this allowed for identification of proteins responsible for all aspects of the organism’s lifestyle, including potential virulence factors. In addition to those previously described, these include resuscitation factors *rpfA* and *rpfB*, the two subtilisin-like serine proteases *sprT* (which was expressed as a recombinant in Chapter Six,) and *clpB*, as well as *norB* which likely plays an important role in the organism’s survival within macrophages. Four sortases were also identified in *Cp* 3/99-5 and numerous sortase substrate proteins. Future studies to complement this work could include laboratory studies into those genes potentially involved in virulence, and it would be interesting to construct mutations in these genes and perform *in vitro* and *in vivo* assays in order to determine their involvement in the pathogenesis of *Cp.* This could be achieved using the allele replacement mutagenesis system developed in *Cp* by Walker (2009).

Additionally the study of *Cp* genomes in the current work has gained insight into genome content between strains and genes potentially involved in host adaptation. The two biochemically diverse biovars of *Cp* are distinct in their host species; biovar *ovis* infects sheep and goats and biovar *equi* is isolated from horses and cattle. Other than host specificity, the main difference described in the literature between the biovars is their ability to reduce nitrate; biovar *equi* is capable of reducing nitrates to nitrites, whereas the *ovis* biovar is not. Interestingly, Songer *et al.* (1988) found that host preference marked by nitrate reductase production may not exist in cattle like the other host species studied. Although the existence of two biovars has been well studied and is widely excepted, little is known about the genes responsible for such distinct host specificity.

In the current study, sequencing of an equine *Cp* isolate (1/06-A), *C. ulcerans* NCTC 12077 and a second ovine *Cp* isolate (42/02-A), was performed in addition to the *Cp* 3/99-5 isolate. Sequencing again was very efficient resulting in few gaps in the genomes and high coverage (**section 4.2.1**). As was shown in Chapter Four, the genome sequences were all highly homologous indicating conservation of the majority of genes within *Cp* and also with *C. ulcerans.* That said the comparison of genome content between the *Cp* 3/99-5, *C. ulcerans* NCTC 12077 and *C. diphtheriae* NCTC 13129 did show that approximately 25% of genes were unique to each genome. The vast quantity of information that the *C. ulcerans* and *Cp* genomes sequenced in this project brings to the field is substantial and will facilitate the understanding of the pathogenicity of *Cp* but also the control of CLA. Crucially the genomes sequenced here have yet to be fully exploited and, as discussed in this chapter, provide scope for a large amount of future work.

Comparative analyses of the *Cp* genomes sequenced in the current work has provided an indication of the overall level of gene conservation between *Cp* strains, of particular interest with regards to the two biovars. Indeed a number of genes have been identified as the ‘accessory’ genome belonging to each biovar and while validation of the accessory genome of each biovar requires further substantiation by inclusion of additional genomes it does give some indication as to the difference between the two biovars. Although it has not been within the scope or aims of this study to look in great detail at the difference between the biovars and host specificity, this would be a highly interesting area for future work to proceed in. These studies would also benefit from sequencing of at least one isolate from a bovine host such as 2/06-L, from which targets proteins were sequenced in the current work. From the comparative and phylogenetic analyses conducted within this thesis, a clear similarity of isolates originating from bovine and equine hosts, which belong to biovar *ovis* and are incapable of reducing nitrate to nitrite, can be observed. However, as discussed above the host preference in cattle may not be so well defined which would make the study of bovine isolates particularly interesting. Significantly, collaboration beyond this PhD thesis includes the incorporation of the three *Cp* strains sequenced here being utilised in a larger pan-genomic study of *Cp*. This will be a larger comparative genomics study investigating isolates from a range of hosts but also geographical locations, and will further investigate the host specificity of *Cp.*

A further aim of this thesis was to accomplish an updated phylogenetic analysis of the *Corynebacterium* genus using Mycobacteria as an outgroup to allow rooting of phylogenetic trees. The phylogenetic (and comparative) studies confirmed *C. ulcerans* to be a closer relative of *Cp* than *C. diphtheriae* is, and the high level of homology within the genus highlighted the importance of considering possible cross-reactivity, particularly with *C. ulcerans* and *C. diphtheriae,* when designing a diagnostic for *Cp*. That said, as previously discussed, there does not necessarily need to be a lack of a diagnostic target gene in all other bacteria. If a target is present in a non-pathogenic bacterium, there would be no implications on the reliability of a diagnostic tool developed using this target. Furthermore, the presence of a target in a pathogenic bacterium could also have no implication on the diagnostic if this pathogen does not infect a host species that can be infected by *Cp.*

Despite substantial research into immunisation against *Cp*, there is no commercially available vaccine for routine use in the UK. What is more, immunisation on emergency license employs vaccines which are not wholly effective in the UK, such as Glanvac™, Caseous D-T™ and Case-Bac™. It is therefore imperative that novel strategies for combating the disease are developed, and it appears logical that such strategies be focused upon the development of improved diagnostic reagents, to allow *Cp*-infected animals to be identified more efficiently and prevent the disease spreading, particularly into *Cp*-naïve flocks. With this in mind there has been a considerable body of work conducted in the area of diagnostics with a focus on serological tests. Although assessment of an ELISA in the UK has perhaps not been as extensive as in other countries, Binns *et al.* (2007) showed promising results from an ELISA developed to detect antibodies to *Cp* in serum. However, a low sensitivity had to be overcome by a large sample size per flock. Later the ELITEST CLA, the only commercially available CLA ELISA, was developed and uses recombinant PLD to detect anti-PLD IgG antibodies in sera. However, with 95% specificity and 85% sensitivity, the ELITEST CLA is only operated on a flock/group basis (Dr Michael Fontaine, unpublished data). Hence an objective of this work was to utilise the *Cp* genome sequence to identify potential diagnostic targets which could be used to improve upon the ELITEST CLA or replace it. In the current work seven potential targets were studied, although only two were successfully expressed as recombinants and neither differentiated between sera from CLA-free and CLA-infected animals.

An alternative to the approach used here to find potential diagnostic targets would be to use the more conventional approach of looking for those proteins which have immunogenic properties. It would be valuable to perform a Western blot with CLA positive sera in an attempt to identify the antigenic proteins. This could be performed on both the secretome and full proteome of *Cp*. Also the use of low iron conditions, such as that created by using a chelating agent in the culture medium, will gain further knowledge about protein expression under stress. Importantly, this would mimic the iron restriction that *Cp* will encounter in the host, and could determine whether this particular stress has an effect on the profile of immunogenic proteins *in vivo*. Once these proteins are identified, recombinant proteins could be created and subsequently the diagnostic potential of the recombinant proteins assessed in more detail using, for example, *in vivo* and *in vitro* virulence studies. Such experimental studies could be carried out in a mouse model (Batey, 1986a; Hamid, 1975) or ideally using the well-established sheep infection model that has been created at Moredun Research Institute (Fontaine *et al.*, 2006). This method would be much more time consuming, however it may be disputed that this would be a more successful method for finding diagnostic (and/or vaccine) targets. The proteins that would be found using this method could be compared to those described following the genome sequencing and bioinformatic analyses of the current study to give an indication as to whether this argument would hold true.

In the current study, some of the proteins identified as possible diagnostic targets were not successfully expressed as recombinants. This was despite several attempts using BL21(DE3) and BL21(DE3)pLysS expression strains, two plasmid vectors (pET-15b and Champion$^{\text{TM}}$ pET SUMO), as well as various expression induction conditions employing three different temperatures and two IPTG concentrations. It would be worth continuing work on these proteins, as perhaps they could be expressed successfully in a different expression host, or under different conditions. It would also be valuable to persevere with protein expression and purification of all targets in the native state. Indeed the recombinant CP40 and Cp3995_0570 assessed here were denatured, and it would have been more appropriate (and more representative to the *in vivo* seroreactivity) to perform the ELISA using these proteins in a native conformation.

Diagnosis is of key importance in the control of CLA as the disease is not generally apparent from the exterior of an infected animal, so prevention of spread by identifying infected animals prior to their contact with *Cp* naïve animals or introduction to *Cp* naïve flocks is fundamental. Although no target suitable for use in a diagnostic test was described, the work presented in this thesis provides substantial of information useful for aiding the process of designing a diagnostic tool for CLA. Further investigation of genes/proteins described here will certainly identify immunogenic proteins previously not described in *Cp,* and perseverance to produce the potential targets discussed as recombinant proteins in the native form could lead to the improvement or replacement of the single diagnostic ELISA currently available on the commercial market.

References

Abreu, de O. S. R., Mota, R. A., Rosinha, G. M. S., Forner, O., Pinheiro Júnior, J. W., Renata, R.B. Pereira, R. R. B., de Castro, R. S., Elisei, C., Soares, C. O., Araújo, F. R. & Madureira, R. C. (2008). Genotypic comparison between *Corynebacterium pseudotuberculosis* samples obtained from sheep and goats with caseous lymphadenitis, raised in the semi-arid region of Pernambuco, *Pesquisa Veterinária Brasileira*, **28** (10), pp. 481-487.

[Anon] (2005). SAC monitoriong scheme for caseous lymphadenitis, *Veterinary Record*, **157** (11), p. 303.

[Anon] (2012). Caseous Lymphadenitis (CLA). *SAC website* http://www.sruc.ac.uk/info/120113/premium_sheep_and_goat_health_schemes/511/diseases/6

Afonseca, V. D., Moraes, P. M., Dorella, F. A., Pacheco, L. G. C., Meyer, R., Portela, R. W., Miyoshi, A., & Azevedo, V. (2008). A description of genes of *Corynebacterium pseudotuberculosis* useful in diagnostics and vaccine applications, *Genetics and Molecular Research*, **7** (1), pp. 252-260.

al-Rawashdeh, O. F. & al-Qudah, K. M. (2000). Effect of shearing on the incidence of caseous lymphadenitis in Awassi sheep in Jordan, *Journal of Veterinary Medicine B Infectious diseases and veterinary public health*, **47** (4), pp. 287-293.

Aleman, M., Spier, S. J., Wilson, W. D., & Doherr, M. (1996). *Corynebacterium pseudotuberculosis* infection in horses: 538 cases (1982-1993), *Journal of the American Veterinary Medical Association*, **209** (4), pp. 804-809.

Alikhan, N. F., Petty, N.K., Zakour, N. L. B. & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons, *BMC Genomics*, **12**, pp. 402-411.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, **25** (17), pp. 3389-3402.

Anderson, D. E., Rings, D. M., & Kowalski, J. (2004). Infection with *Corynebacterium pseudotuberculosis* in five alpacas, *Javma-Journal of the American Veterinary Medical Association*, **225** (11), pp. 1743-1747.

Andersson, S. G., Zomorodipour, A., Winkler, H. H., & Kurland, C. G. (1995). Unusual organization of the rRNA genes in *Rickettsia prowazekii*, *Journal of Bacteriology*, **177** (14), pp. 4171-4175.

Andrews, S. C., Robinson, A. K., & Rodríguez-Quiñones, F. (2003). Bacterial iron homeostasis, *FEMS Microbiology Reviews*, **27** (2-3), pp. 215-237.

Anisimova, M. & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative, *Systematic Biology*, **55** (4), pp. 539-552.

Augustine, J. L. & Renshaw, H. W. (1986). Survival of *Corynebacterium pseudotuberculosis* in axenic purulent exudate on common barnyard fomites, *American Journal of Veterinary Research*, **47** (4), pp. 713-715.

Aziz, R. *et al.* (2008). The RAST Server: Rapid Annotations using Subsystems Technology, *BMC Genomics*, **9** (1), p. 75.

Baird, G. & Malone, F. (2001). Post-mortem examination of sheep from caseous lymphadenitis (CLA) infected flocks, *Proceedings of the 5th International Sheep Veterinary Conference, University of Stellenbosch , South Africa, 21-25 January, 2001* p. unpaginated.

Baird, G., Synge, B., Armstrong, D., & Dercksen, D. (2001). Caseous lymphadenitis in sheep in the United Kingdom, *Veterinary Record*, **149** p. 399.

Baird, G., Synge, B., & Dercksen, D. (2004). Survey of caseous lymphadenitis seroprevalence in British terminal sire sheep breeds, *Veterinary Record*, **154** (16), pp. 505-506.

Baird, G. (2006). Treatment of ovine caseous lymphadenitis, *Veterinary Record*, **159** (15), p. 500.

Baird, G. J. 2007. "Caseous lymphadenitis," in *Diseases of Sheep 4th Edition*, I. D. Aitken, ed., Blackwell, pp. 306-311.

Baird, G. J. & Fontaine, M. C. (2007). *Corynebacterium pseudotuberculosis* and its role in ovine caseous lymphadenitis, *Journal of Comparative Pathology*, **137** (4), pp. 179-210.

Baird, G. J. & Malone, F. E. (2010). Control of caseous lymphadenitis in six sheep flocks using clinical examination and regular ELISA testing, *Veterinary Record*, **166** (12), pp. 358-362.

Barksdale, L., Linder, R., Sulea, I. T., & Pollice, M. (1981). Phospholipase-D activity of *Corynebacterium pseudotuberculosis (Corynebacterium ovis)* and *Corynebacterium ulcerans*, a distinctive marker within the genus *Corynebacterium* , *Journal of Clinical Microbiology*, **13** pp. 335-343.

Bashir, N., Kounsar, F., Mukhopadhyay, S., & Hasnain, S. E. (2010). *Mycobacterium tuberculosis* conserved hypothetical protein rRv2626c modulates macrophage effector functions, *Immunology*, **130** (1), pp. 34-45.

Batey, R. G. (1986a). Pathogenesis of caseous lymphadenitis in sheep and goats, *Australian Veterinary Journal*, **63** (9), pp. 269-272.

Batey, R. G. (1986b). Aspects of pathogenesis in a mouse model of infection by *Corynebacterium pseudotuberculosis.*, *Australian Journal* of *Experimental Biology & Medical Science*, **64** (3), pp. 237-249.

Batycka, M., Inglis, N. F., Cook, K., Adam, A., Fraser-Pitt, D., Smith, D. G. E., Main, L., Lubben, A., & Kessler, B. M. (2006). Ultra-fast tandem mass spectrometry scanning combined with monolithic column liquid chromatography increases throughput in proteomic analysis, *Rapid Communications in Mass Spectrometry*, **20** (14), pp. 2074-2080.

Baum, D. (2008). Reading a phylogenetic tree: The meaning of monophyletic groups, *Nature Education*, **1** (1).

Bek-Pederson, S. (1997). An outbreak of Morel's disease among Lacoune sheep imported to Denmark., *Proceedings of the Sheep Veterinary Society*, **21** pp. 143-144.

Bendtsen, J. D., Nielsen, H., von Heijne, G., & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0, *Journal of Molecular Biology*, **340** (4), pp. 783-795.

Bercovier, H., Kafri, O., & Sela, S. (1986). Mycobacteria possess a surprisingly small number of ribosomal RNA genes in relation to the size of their genome, *Biochemical and Biophysical Research Communications*, **136** (3), pp. 1136-1141.

Bernheimer, A. W., Linder, R., & Avigad, L. S. (1980). Stepwise degradation of membrane sphingomyelin by corynebacterial phospholipases, *Infection and Immunity*, **29** (1), pp. 123-131.

Biberstein, E. L., Knight, H. D., & Jang, S. (1971). Two biotypes of *Corynebacterium pseudotuberculosis*, *Veterinary Record*, **89** (26), pp. 691-692.

Billington, S. J., Esmay, P. A., Songer, J. G., & Jost, B. H. (2002). Identification and role in virulence of putative iron acquisition genes from *Corynebacterium pseudotuberculosis*, *FEMS Microbiology Letters*, **208** (1), pp. 41-45.

Binns, S. H., Green, L. E., & Bailey, M. (2007). Development and validation of an ELISA to detect antibodies to *Corynebacterium pseudotuberculosis* in ovine sera, *Veterinary Microbiology*, **123** (1-3), pp. 169-179.

Birnboim, H. C. & Doly, J. (1979). Rapid alkaline extraction procedure for screening recombinant plasmid DNA, *Nucleic Acids Research*, **7** (6), pp. 1513-1523.

Björkroth, J., Korkeala, H. & Funke, G. (1999). rRNA gene RFLP as an identification tool for *Corynebacterium* species, *International Journal of Systematic and Evolutionary Microbiology*, **49**: pp.983-989.

Bolt, F. (2009). The population structure of the *Corynebacterium diphtheriae* group. PhD thesis, University of Warwick.

Bolt, F., Cassiday, P., Tondella, M. L., DeZoysa, A., Efstratiou, A., Sing, A., Zasada, A., Bernard, K., Guiso, N., Badell, E., Rosso, M. L., Baldwin, A. and Dowson, C. (2010). Multilocus sequence typing identifies evidence for recombination and two distinct lineages of *Corynebacterium diphtheriae, Journal of Clinical Microbiology*, **48** (11), pp. 4177-4185.

Bonifait, L., de la Cruz Dominguez-Punaro, Vaillancourt, K., Bart, C., Slater, J., Frenette, M., Gottschalk, M., & Grenier, D. (2010). The cell envelope subtilisin-like proteinase is a virulence determinant for *Streptococcus suis*, *BMC Microbiology*, **10** p. 42.

Boyle, J. S., Brady, J. L., & Lew, A. M. (1998). Enhanced responses to a DNA vaccine encoding a fusion antigen that is directed to sites of immune induction, *Nature*, **392** (6674), pp. 408-411.

Braga, W. U., Chavera, A., & Gonzalez, A. (2006). *Corynebacterium pseudotuberculosis* infection in highland alpacas (Lama pacos) in Peru, *Veterinary Record*, **159** (1), pp. 23-24.

Brazas, M. D. & Hancock, R. E. W. (2005). Using microarray gene signatures to elucidate mechanisms of antibiotic action and resistance, *Drug Discovery Today*, **10** (18), pp. 1245-1252.

Brodin, P., Rosenkrands, I., Andersen, P., Cole, S. T., and Brosch, R. (2004). ESAT-6 proteins: protective antigens and virulence factors? *Trends in Microbiology* **12** (11), 500-508

Brogden, K. A. & Ester, M. (1990). Alterations in the phospholipid composition and morphology of ovine erythrocytes after intravenous inoculation of *Corynebacterium pseudotuberculosis, American Journal of Veterinary Research*, **51** (6), pp. 874-877.

Brogden, K. A., Glenn, J. S., East, N., & Audibert, F. (1996). A *Cornynebacterium pseudotuberculosis* bacterin with muramyl dipeptide induces antibody liters, increases the time of onset, and decreases naturally occurring external abscesses in sheep and goats, *Small Ruminant Research*, **19** (2), pp. 161-168.

Brown, C. C. & Olander, H. J. (1987). Caseous lymphadenitis of goats and sheep: a review, *Veterinary Bulletin*, **57** (1), pp. 1-12.

Brown, J. S. & Holden, D. W. (2002). Iron acquisition by Gram-positive bacterial pathogens, *Microbes and Infection*, **4** (11), pp. 1149-1156.

Brown, T. A. (2007). "Molecular Phylogenetics," in *Genomes 3*, 3 ed, Garland Science Publishing, Taylor & Francis Group, NY, USA, pp. 595-626.

Burrell, D. H. (1980). Simplified double immunodiffusion technique for detection of *Corynebacterium ovis* antitoxin, *Research in Veterinary Science*, **28** (2), pp. 234-237.

Burrell, D. H. (1980). Hemolysis Inhibition Test for detection of antibody to *Corynebacterium ovis* exotoxin, *Research in Veterinary Science*, **28** (2), pp. 190-194.

Burts, M. L., Williams, W. A., DeBord, K., & Missiakas, D. M. (2005). EsxA and EsxB are secreted by an ESAT-6-like system that is required for the pathogenesis of *Staphylococcus aureus* infections, *Proceedings of the National Academy of Sciences of the United States of America*, **102** (4), pp. 1169-1174.

Cabrera, G., Xiong, A., Uebel, M., Singh, V. K., & Jayaswal, R. K. (2001). Molecular characterization of the iron-hydroxamate uptake system in *Staphylococcus aureus*, *Applied and Environmental Microbiology*, **67** (2), pp. 1001-1003.

Cameron, C. M., Minnaar, J. L., Engelbrecht, M. M., & Purdom, M. R. (1972). Immune response of merino sheep to inactivated *Corynebacterium pseudotuberculosis* vaccine, *Journal of Veterinary Research*, **39** (1), pp. 11-24.

Carne, H. R., Wickham, N., & Kater, J. C. (1956). A toxic lipid from the surface of *Corynebacterium ovis*, *Nature*, **178** pp. 701-702.

Carne, H. R. & Onon, E. O. (1978). Action of *Corynebacterium ovis* exotoxin on endothelial cells of blood-vessels, *Nature*, **271** (5642), pp. 246-248.

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G., & Parkhill, J. (2005). ACT: the Artemis comparison tool, *Bioinformatics*, **21** (16), pp. 3422-3423.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis, *Molecular Biology and Evolution*, **17** (4), pp. 540-552.

Cerdeira, L. T. *et al.* (2011a). Whole-genome sequence of *Corynebacterium pseudotuberculosis* PAT10 strain isolated from sheep in Patagonia, Argentina, *Journal of Bacteriology*, **193** (22), pp. 6420-6421.

Cerdeira, L. T. *et al.* (2011b). Complete genome sequence of *Corynebacterium pseudotuberculosis* strain CIP 52.97, isolated from a horse in Kenya, *Journal of Bacteriology*, **193** (24), pp. 7025-7026.

Cerdeno-Tarraga, A. M. *et al.* (2003). The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129, *Nucleic Acids Research*, **31** (22), pp. 6516-6523.

Cetinkaya, B., Karahan, M., Atil, E., Kalin, R., De Baere, T., & Vaneechoutte, M. (2002). Identification of *Corynebacterium pseudotuberculosis* isolates from sheep and goats by PCR, *Veterinary Microbiology*, **88** (1), pp. 75-83.

Chalupa, P., Beran, O., Herwald, H., Kaspříková, N. & Holub, M. (2011). Evaluation of potential biomarkers for the discrimination of bacterial and viral infections, *Infection*, **39**, pp. 411-417.

Chaplin, P. J., De Rose, R., Boyle, J. S., McWaters, P., Kelly, J., Tennent, J., Lew, A. M., & Scheerlinck, J. P. Y. (1999). Targeting improves the efficacy of a DNA vaccine against *Corynebacterium pseudotuberculosis* in sheep, *Infection and Immunity*, **67** (12), pp. 6434-6438.

Chastanet, A., Derre, I., Nair, S., & Msadek, T. (2004). ClpB, a novel member of the *Listeria monocytogenes* CtsR regulon, is involved in virulence but not in general stress tolerance, *Journal of Bacteriology*, **186** (4), pp. 1165-1174.

Chaudhuri, R. R., Loman, N. J., Snyder, L. A. S., Bailey, C. M., Stekel, D. J., & Pallen, M. J. (2008). xBASE2: a comprehensive resource for comparative bacterial genomics, *Nucleic Acids Research*, **36** (suppl 1), p. D543-D546.

Chen, H., Huang, N., & Sun, Z. (2006). SubLoc: a server/client suite for protein subcellular location based on SOAP, *Bioinformatics*, **22** (3), pp. 376-377.

Cho, D., Sung, N. & Collins, M. T. (2006). Identification of proteins of potential diagnostic value for bovine paratuberculosis, *Proteomics*, **6**, pp. 5785-5794.

Chou, K. C. & Shen, H. B. (2008). Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms, *Nature Protocols*, **3** (2), pp. 153-162.

Cloonan, N. *et al.* (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing, *Nature Methods*, **5** (7), pp. 613-619.

Collins, M. D., Goodfellow, M., & Minnikin, D. E. (1982). A survey of the structures of mycolic acids in *Corynebacterium* and related taxa, *Journal of General Microbiology*, **128** pp. 129-149.

Condon, C., Liveris, D., Squires, C., Schwartz, I., & Squires, C. L. (1995). rRNA operon multiplicity in *Escherichia coli* and the physiological implications of *rrn* inactivation, *Journal of Bacteriology*, **177** (14), pp. 4152-4156.

Connor, K. M., Quirie, M. M., Baird, G., & Donachie, W. (2000). Characterization of United Kingdom isolates of *Corynebacterium pseudotuberculosis* using pulsed-field gel electrophoresis, *Journal of Clinical Microbiology*, **38** (7), pp. 2633-2637.

Connor, K. M., Fontaine, M. C., Rudge, K., Baird, G. J., & Donachie, W. (2007). Molecular genotyping of multinational ovine and caprine *Corynebacterium pseudotuberculosis* isolates using pulsed-field gel electrophoresis, *Veterinary Research*, **38** (4), pp. 613-623.

Dautle, M. P., Ulrich, R. L. & Hughes, T. A. (2002). Typing and subtyping of 83 clinical isolates purified from surgically implanted feeding tubes by random amplified polymorphic DNA amplification, *Journal of Clinical Microbiology*, **40** (2), pp. 414-421.

De Rose, R., Tennent, J., McWaters, P., Chaplin, P. J., Wood, P. R., Kimpton, W., Cahill, R., & Scheerlinck, J. P. Y. (2002). Efficacy of DNA vaccination by different routes of immunisation in sheep, *Veterinary Immunology and Immunopathology*, **90** (1-2), pp. 55-63.

De Zoysa, A., Hawkey, P. M., Engler, K., George, R. t., Mann, G., Reilly, W., Taylor, D., & Efstratiou, A. (2005). Characterization of toxigenic *Corynebacterium ulcerans* strains isolated from humans and domestic cats in the United Kingdom, *Journal of Clinical Microbiology*, **43** (9), pp. 4377-4381.

Delcher, A. L., Bratke, K. A., Powers, E. C., & Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer, *Bioinformatics*, **23** (6), pp. 673-679.

DelVicchio, V. G., Connolly, J. P., Alefantis, T. G., Walz, A., Quan, M. A., Patra, G., Ashton, J. M., Whittington, J. T., Chafin, R. D., Liang, X., Grewal, P., Khan, A. S. & Mujer, C. V. (2006). Proteomic profiling and identification of immunodominant spore antigens of *Bacillus anthracis, Bacillus cereus* and *Bacillus thuringiensis*, *Applied Environmental Microbiology,* **72** (9), pp. 6355-6363.

Dercksen, D. P., Terlaak, E. A., & Schreuder, B. E. C. (1996). Eradication programme for caseous lymphadenitis in goats in the Netherlands, *Veterinary Record*, **138** (10), p. 237.

Dercksen, D. P., Brinkhof, J. M. A., Dekker-Nooren, T., van Maanen, K., Bode, C. F., Baird, G., & Kamp, E. M. (2000). A comparison of four serological tests for the diagnosis of caseous lymphadenitis in sheep and goats, *Veterinary Microbiology*, **75** (2), pp. 167-175.

Dorella, F. A., Pacheco, L. G. C., Oliveira, S. C., Miyoshi, A., & Azevedo, V. (2006). *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence, *Veterinary Research*, **37** (2), pp. 201-218.

Dorneles, E. M., Santana, J. A., Andrade, G. I., Santos, E. L., Guimaraes, A. S., Mota, R. A., Santos, A. S., Miyoshi, A., Azevedo, V., Gouveia, A. M., Lage, A. P. And Heinemann, M. B. (2012). Molecular characterisation of *Corynebacterium pseudotuberculosis* isolated from goats using ERIC-PCR, *Genetic and Molecular Research,* **11** (3), pp. 2051-2059.

Dubos, R. J. & Middlebrook, G. (1948). The effect of wetting agents on the growth of tubercle bacilli, *The Journal of Experimental Medicine*, **88** (1), pp. 81-88.

Einhauer, A. & Jungbauer, A. (2001). The FLAG™ peptide, a versatile fusion tag for the purification of recombinant proteins, *Journal of Biochemical & Biophysical Methods*, **49** (1-3), pp. 455-465.

Egen, N. B., Cuevas, W. A., McNamara, P. J., Sammons, D. W., Humphreys, R., & Songer, J. G. (1989). Purification of the phospholipase D of *Corynebacterium pseudotuberculosis* by recycling isoelectric focusing, *American Journal of Veterinary Research,* **50** (8), pp. 1319-1322.

Eggleton, D. G., Doidge, C. V., Middleton, H. D., & Minty, D. W. (1991a). Immunization against ovine caseous lymphadenitis-efficacy of monocomponent *Corynebacterium pseudotuberculosis* toxoid vaccine and combined Clostridial-Corynebacterial vaccines, *Australian Veterinary Journal,* **68** (10), pp. 320-321.

Eggleton, D. G., Haynes, J. A., Middleton, H. D., & Cox, J. C. (1991b), Immunization against ovine caseous lymphadenitis - correlation between *Corynebacterium pseudotuberculosis* toxoid content and protective efficacy in combined Clostridial-Corynebacterial vaccines, *Australian Veterinary Journal,* **68** (10), pp. 322-325.

Eggleton, D. G., Middleton, H. D., Doidge, C. V., & Minty, D. W. (1991c). Immunization against ovine caseous lymphadenitis - comparison of *Corynebacterium pseudotuberculosis* vaccines with and without bacterial cells, *Australian Veterinary Journal*, **68** (10), pp. 317-319.

Ellwood, M. & Nomura, M. (1980). Deletion of a ribosomal ribonucleic acid operon in *Escherichia coli*, *Journal of Bacteriology*, **143** (2), pp. 1077-1080.

Euzeby, J. P. (2005). List of bacterial names with standing in nomenclature. *Society for Systemic and Veterinary Bacteriology*

Farfour, E., Badell, E., Zasada, A., Hotzel, H., Tomaso, H., Guillot, S. & Guiso, N. (2012). Characterization and comparison of invasive *Corynebacterium diphtheriae* isolates from France and Poland, *Journal of Clinical Microbiology*, **50** (1), pp. 173-175.

Fleischmann, R. D. *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, *Science*, **269** (5223), pp. 496-512.

Foley, J. E., Spier, S. J., Mihalyi, J., Drazenovich, N. & Leutenegger, C. M. (2004). Molecular epidemiologic features of *Corynebacterium pseudotuberculosis* isolated from horses. *American Journal of Veterinary Research*, **65** (12), pp. 1734-1737

Fontaine, M. C., Baird, G., Connor, K. M., Rudge, K., Sales, J., & Donachie, W. (2006). Vaccination confers significant protection of sheep against infection with a virulent United Kingdom strain of *Corynebacterium pseudotuberculosis*, *Vaccine*, **24** (33-34), pp. 5986-5996.

Fontaine, M. C. & Baird, G. J. (2008). Caseous lymphadenitis, *Small Ruminant Research*, **76** (1-2), pp. 42-48.

Frangioni, J. V. & Neel, B. G. (1993). Solubilization and purification of enzymatically active glutathione S-transferase (pGEX) fusion proteins, *Analytical biochemistry*, **210** (1), pp. 179-187.

Fraser, G. (1961). Haemolytic activity of *Corynebacterium ovis*, *Nature*, **189** pp. 246-247.

Freney, J., Duperron, M. T., Courtier, C., Hansen, W., Allard, F., Boeufgras, J. M., Monget, D., & Fleurette, J. (1991). Evaluation of Api Coryne in comparison with conventional methods for identifying coryneform bacteria, *Journal of Clinical Microbiology*, **29** (1), pp. 38-41.

Funke, G., vonGraevenitz, A., Clarridge, J. E., & Bernard, K. A. (1997). Clinical microbiology of coryneform bacteria, *Clinical Microbiology Reviews*, **10** (1), p. 125.

Gande, R., Dover, L. G., Krumbach, K., Besra, G. S., Sahm, H., Oikawa, T., & Eggeling, L. (2007). The two carboxylases of *Corynebacterium glutamicum* essential for fatty acid and mycolic acid synthesis, *Journal of Bacteriology*, **189** (14), pp. 5257-5264.

Gardy, J. L., Laird, M. R., Chen, F., Rey, S., Walsh, C. J., Ester, M., & Brinkman, F. S. L. (2005). PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis, *Bioinformatics*, **21** (5), pp. 617-623.

Ghannoum, M. A. (2000). Potential role of phospholipases in virulence and fungal pathogenesis, *Clinical Microbiology Reviews* , **13** (1), p. 122-+.

Gilbert, W. & Maxam, A. (1973). The nucleotide sequence of the lac operator, *Proceedings of the National Academy of Sciences* , **70** (12), pp. 3581-3584.

Giuliani, M. M. *et al.* (2006). A universal vaccine for serogroup B meningococcus, *Proceedings of the National Academy of Sciences,* **103** (29), pp. 10834-10839.

Goetz, R. M. & Fuglsang, A. (2005). Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli, Biochemical and Biophysical Research Communications,* **327** (1), pp. 4-7.

Golaz, A., Hardy, I. R., Strebel, P., Bisgard, K. M., Vitek, C., Popovic, T., & Wharton, M. (2000). Epidemic diphtheria in the newly independent states of the former Soviet Union: Implications for diphtheria control in the United States, *Journal of Infectious Diseases,* **181** (Supplement 1), pp. S237-S243.

Gouy, M. & Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity, *Nucleic Acid Research*, **25** (10), pp. 7055-7074.

Groman, N., Schiller, J., & Russell, J. (1984). *Corynebacterium ulcerans* and *Corynebacterium pseudotuberculosis* responses to DNA probes derived from corynephage beta and *Corynebacterium diphtheriae*, *Infection and Immunity*, **45** (2), pp. 511-517.

Groth, A. C. & Calos, M. P. (2004). Phage Integrases: Biology and applications, *Journal of Molecular Biology,* **335** (3), pp. 667-678.

Guimaraes, A. *et al.* (2011). High sero-prevalence of caseous lymphadenitis identified in slaughterhouse samples as a consequence of deficiencies in sheep farm management in the state of Minas Gerais, Brazil, *BMC Veterinary Research,* **7** (1), p. 68.

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by Maximum Likelihood, *Systematic Biology,* **52** (5), pp. 696-704.

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R., & Goebel, W. (1990). Deletions of chromosomal regions coding for fimbriae and hemolysins occur *in vitro* and *in vivo* in various extra intestinal *Escherichia coli* isolates, *Microbial Pathogenesis,* **8** (3), pp. 213-225.

Hadfield, T. L., McEvoy, P., Polotsky, Y., Tzinserling, V. A., & Yakovlev, A. A. (2000). The pathology of diphtheria, *Journal of Infectious Diseases*, **181** (Supplement 1), pp. S116-S120.

Hall, K., McCluskey, B. J., & Cunningham, W. (2001). *Corynebacterium pseudotuberculosis* infections (pigeon fever) in horses in western Colorado: An epidemiological investigation, *Journal of Equine Veterinary Science*, **21** (6), pp. 284-286.

Hamid, Y. & Zaki, M. (1973). Immune response of goats artificially infected with *C. ovis, Journal of the Egyptian Veterinary Medical Association*, **33** (3/4), pp. 137-140.

Hamid, Y. M. (1975). The use of mouse protection tests in the diagnosis of caseous lymphadenitis in sheep, *Research in Veterinary Science*, **18** (2), pp. 223-224.

Henk, J. M. A., Karim, S. B., Xavier, N., & Axel, C. (2001). Molecular tools for the characterisation of antibiotic-resistant bacteria, *Veterinary Research*, **32** (3-4), pp. 363-380.

Hershberg, R., Altuvia, S. & Margalit, H. (2003). A survey of small RNA-encoding genes in *Escherichia coli, Nucleic Acids Research,* **31** (7), pp. 1813-1820.

Hiraoka, Y., Kawamata, K., Haraguchi, T. & Chikashige, Y. (2009). Codon usage bias is correlated with gene expression levels in the fission yeast *Schizosaccharomyces pombe, Genes to Cells*, **14** (4), pp. 499-509.

Hodgson, A. L., Bird, P., & Nisbet, I. T. (1990) Cloning, nucleotide sequence, and expression in *Escherichia coli* of the Phospholipase D gene from *Corynebacterium pseudotuberculosis, Journal of Bacteriology,* **172** (3), pp. 1256-1261.

Hodgson, A. L., Tachedjian, M., Corner, L. A., & Radford, A. J. (1994). Protection of sheep against caseous lymphadenitis by use of a single oral dose of a live recombinant *Corynebacterium pseudotuberculosis, Infection and Immunity,* **62** (12), pp. 5275-5280.

Hodgson, A. L. M., Krywult, J., Corner, L. A., Rothel, J. S., & Radford, A. J. (1992). Rational attenuation of *Corynebacterium pseudotuberculosis* - potential cheesy gland vaccine and live delivery vehicle, *Infection and Immunity*, **60** (7), pp. 2900-2905.

Hogg, J. S., Hu, F. Z., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, J. C., & Ehrlich, G. D. (2007). Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains, *Genome Biology*, **8** (6), pp. 1-18.

Holmes, R. K. (2000). Biology and molecular epidemiology of diphtheria toxin and the *tox* gene., *Journal of Infectious Diseases,* **181** (Supplement 1), p. S156-S167.

Hommez, J., Devriese, L. A., Vaneechoutte, M., Riegel, P., Butaye, P., & Haesebrouck, F. (1999). Identification of nonlipophilic Corynebacteria isolated from dairy cows with mastitis, *Journal of Clinical Microbiology,* **37** (4), pp. 954-957.

Huelsenbeck, J. P. & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees, *Bioinformatics,* **17** (8), pp. 754-755.

Hughes, J. P. & Biberstein, E. (1959). Chronic equine abscesses associated with *Corynebacterium pseudotuberculosis, Journal of the American Veterinary Medical Association*, **135** (11), pp. 559-562.

Huson, D. H. & Scornavacca, C. (2012). Dendroscope 3: an interactive viewer for rooted phylogenetic trees and networks, *Systematic Biology,* **61** (6), pp. 1061-1067.

Hyatt, D., Chen, G. L., LoCascio, P., Land, M., Larimer, F., & Hauser, L. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification, *BMC Bioinformatics*, **11** (1), p. 119.

Janda, J. M. & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls, *Journal of Clinical Microbiology*, **45** (9), pp. 2761-2764.

Join-Lambert, O. F., Ouache, M., Canioni, D., Beretti, J. L., Blanche, S., Berche, P., & Kayal, S. (2006). *Corynebacterium pseudotuberculosis* necrotizing lymphadenitis in a twelve-year-old patient, *The Pediatric Infectious Disease Journal*, **25** (9), pp. 848-851.

Jolly, R. D. (1965). Pathogenic action of exotoxin of *Corynebacterium ovis*, *Journal of Comparative Pathology and Therapeutics*, **75** (4), p. 417.

Josefsson, E., Hartford, O., O'Brien, L., Patti, J. M., & Foster, T. (2001). Protection against experimental *Staphylococcus aureus* arthritis by vaccination with clumping factor A, a novel virulence determinant, *Journal of Infectious Diseases*, **184** (12), pp. 1572-1580.

Kana, B. D. *et al.* (2008). The resuscitation-promoting factors of *Mycobacterium tuberculosis* are required for virulence and resuscitation from dormancy but are collectively dispensable for growth i*n vitro*, *Molecular Microbiology*, **67** (3), pp. 672-684.

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Research*, **30** (14), pp. 3059-3066.

Kayser, F. H. (2003). Safety aspects of enterococci from the medical point of view, *International Journal of Food Microbiology*, **88** (2-3), pp. 255-262.

Khamis, A., Raoult, D., & La Scola, B. (2004). *rpoB* gene sequencing for identification of *Corynebacterium* species, *Journal of Clinical Microbiology*, **42** (9), pp. 3925-3931.

Khamis, A., Raoult, D., & La Scola, B. (2005). Comparison between *rpoB* and 16S rRNA gene sequencing for molecular identification of 168 clinical isolates of *Corynebacterium*, *Journal of Clinical Microbiology*, **43** (4), pp. 1934-1936.

Klappenbach, J. A., Dunbar, J. M., & Schmidt, T. M. (2000). rRNA operon copy number reflects ecological strategies of bacteria, *Applied and Environmental Microbiology*, **66** (4), pp. 1328-1333.

Krishnamurthy, T., Ross, P. L. & Rajamani, U. (1996). Detection of pathogenic and non-pathogenic bacteria by matrix assisted laser desorption/ionisation time-of-flight mass spectrometry, *Rapid Communications in Mass Spectrometry*, **10**, pp. 883-888.

Kurtz, S., Phillippy, A., Delcher, A., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. (2004). Versatile and open software for comparing large genomes, *Genome Biology*, **5** (2), p. R12.

Lagesen, K., Hallin, P. F., Roland, E., Staerfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer:consistent annotation of rRNA genes in genomic sequences, *Nucleic Acids Research*, **35** (9), pp. 3100-3108.

Langille, M., Hsiao, W., & Brinkman, F. (2008). Evaluation of genomic island predictors using a comparative genomics approach, *BMC Bioinformatics*, **9** (1), p. 329.

Langille, M. & Brinkman, F. (2009). IslandViewer: an integrated interface for computational identification and visualization of genomic islands, *Bioinformatics*, **25** (5), pp. 664-665.

Langille, M. G. I., Hsiao, W. W. L., & Brinkman, F. S. L. (2010). Detecting genomic islands using bioinformatics approaches, *Nature Reviews Microbiology*, **8** (5), pp. 373-382.

Larsen, M. V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R. L., Jelsbak, L., Sicheritz-Ponten, T., Ussery, D. W., Aarestrup, F. M. & Lund, O. (2012). Multilocus sequence typing of total-genome-sequenced bacteria, *Journal of Clinical Microbiology*, **50** (4), pp. 1355-1361.

Leekitcharoenphon, P., Lukjancenko, O., Friis, C., Aaerestrup, F. M. & Ussery, D. W. (2012). Genomic variation in *Salmonella enterica* core genes for epidemiological typing, *BMC Genomics*, **13**, pp. 88-99.

Lenz, D. H., Mok, K. C., Lilley, B. N., Kulkarni, R. V., Wingreen, N. S. & Bassler, B. L. (2004). The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in *Vibrio harveyi* and *Vibrio cholera, Cell,* **118** (1), pp. 69-82.

Levin, B. R. & Edén, C. S. (1990). Selection and evolution of virulence in bacteria: an ecumenical excursion and modest suggestion, *Parasitology*, **100**, pp. S103-S115.

Li, L., Huang, D., Cheung, m. K., Nong, W., Huang, Q. & Kwan, H. S. (2012). BSRD: a repository for bacterial small regulatory RNA, *Nucleic Acids Research*, **41** (D1), D233-D238.

Liu, G., Wu, J., Yang, H., & Bao, Q. (2010). Codon usage patterns in *Corynebacterium glutamicum*: mutational bias, natural selection and amino acid conservation, *Comparative and Functional Genomics*, **2010** p. Article ID 343569.

Liu, S., Han, W., Sun, C., Lei, L., Feng, X., Yan, S., Diao, Y., Gao, Y., Zhao, H. L., Liu, Q., Yao, C., Li, M. (2011). Subtractive screening with the *Mycobacterium tuberculosis* surface protein phage display library, *Tuberculosis*, **91** (6), pp. 579-586.

Lloyd, S., Lindsay, H. J., Slater, J. D., & Jackson, P. G. G. (1990). Caseous lymphadenitis in goats in England, *Veterinary Record*, **127** (19), p. 478.

Loughney, K., Lund, E., & Dahlberg, J. E. (1983). Deletion of an rRNA gene set in *Bacillus subtilis, Journal of Bacteriology,* **154** (1), pp. 529-532.

Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Research*, **25** (5), pp. 955-964.

Löytynoja, A. & Milinkovitch, M. C. (2001). SOAP, cleaning multiple alignments from unstable blocks, *Bioinformatics*, **17** (6), pp. 573-574.

Maione, D. *et al.* (2005). Identification of a universal Group B Streptococcus vaccine by multiple genome screen, *Science*, **309** (5731), pp. 148-150.

Makinoshima, H. & Glickman, M. S. (2006). Site-2 proteases in prokaryotes: regulated intramembrane proteolysis expands to microbial pathogenesis, *Microbes and Infection*, **8** (7), pp. 1882-1888.

Markowitz, V. M. *et al.* (2010). The integrated microbial genomes system: an expanding comparative analysis resource, *Nucleic Acids Research*, **38** (suppl 1), p. D382-D390.

Maximesc, P., Pop, A., Oprisan, A., & Potorac, E. (1974). Diphtheria tox gene expressed in *Corynebacterium* species other than *C. diphtheriae*, *Journal of Hygiene Epidemiology Microbiology and Immunology*, **18** (3), p. 324-&.

Maximesc, P., Oprisan, A., Pop, A., & Potorac, E. (1974). Further studies on *Corynebacterium* species capable of producing Diphtheria-Toxin (*C. diphtheriae, C. ulcerans, C. ovis), Journal of General Microbiology*, **82** pp. 49-56.

McGinley, K. J., Labows, J. N., Zechman, J. M., Nordstrom, K. M., Webster, G. F., & Leyden, J. J. (1985). Analysis of cellular components, biochemical reactions, and habitat of human cutaneous lipophilic diphtheroids, *Journal of Investigative Dermatology*, **85** (4), pp. 374-377.

McGinley, K. J., Labows, J. N., Zechman, J. M., Nordstrom, K. M., Webster, G. F., & Leyden, J. J. (1985). Pathogenic JK group Corynebacteria and their similarity to human cutaneous lipophilic diphtheroids, *Journal of Infectious Diseases*, **152** (4), pp. 801-806.

Mckean, S. C., Davies, J. K., & Moore, R. J. (2007). Expression of phospholipase D, the major virulence factor of *Corynebacterium pseudotuberculosis*, is regulated by multiple environmental factors and plays a role in macrophage death, *Microbiology-SGM*, **153** pp. 2203-2211.

McLennan, A. G. (2006). The Nudix hydrolase superfamily, *Cellular and Molecular Life Sciences*, **63** (2), pp. 123-143.

McNamara, P. J., Bradley, G. A., & Songer, J. G. (1994). Targeted mutagenesis of the phospholipase-D gene results in decreased virulence of *Corynebacterium pseudotuberculosis*, *Molecular Microbiology*, **12** (6), pp. 921-930.

McNamara, P. J., Cuevas, W. A., & Songer, J. G. (1995). Toxic phospholipases D of *Corynebacterium pseudotuberculosis*, *C. ulcerans* and *Arcanobacterium haemolyticum*: cloning and sequence homology, *Gene*, **156** pp. 113-118.

McNeil, L. K., & Aziz, R. K. (2009). *In silico* reconstruction of the metabolic and pathogenic potential of bacterial genomes using subsystems, in *Microbial Pathogenomics*, volume 6, H.de Reuse & S. Bereswill, eds., Genome Dyn. Basel, Karger, pp. 21-34.

Menzies, P. I., Muckle, C. A., Brogden, K. A., & Robinson, L. (1991). A field trial to evaluate a whole cell vaccine for the prevention of caseous lymphadenitis in sheep and goat flocks, *Canadian Journal of Veterinary Research-Revue Canadienne de Recherche Veterinaire*, **55** (4), pp. 362-366.

Menzies, P. I., Muckle, C. A., Hwang, Y. T., & Songer, J. G. (1994). Evaluation of an Enzyme-Linked-Immunosorbent-Assay using an *Escherichia coli* recombinant phospholipase-D antigen for the diagnosis of *Corynebacterium pseudotuberculosis* infection, *Small Ruminant Research*, **13** (2), pp. 193-198.

Menzies, P. I., Hwang, Y. T., & Prescott, J. F. (2004). Comparison of an interferon-gamma to a phospholipase D enzyme-linked immunosorbent assay for diagnosis of *Corynebacterium pseudotuberculosis* infection in experimentally infected goats, *Veterinary Microbiology*, **100** (1-2), pp. 129-137.

Metzker, M. L. (2005). Emerging technologies in DNA sequencing, *Genome Research*, **15** (12), pp. 1767-1776.

Metzker, M. L. (2010). Sequencing technologies-the next generation, *Nature Reviews Genetics*, **11** (1), pp. 31-46.

Mills, A. E., Mitchell, R. D., & Lim, E. K. (1997). *Corynebacterium pseudotubrculosis* is a cause of human necrotising granulomatous lymphadenitis, *Pathology*, **29** pp. 231-233.

Milne, I., Lindner, D., Bayer, M., Husmeier, D., McGuire, G., Marshall, D. F., & Wright, F. (2009). TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops, *Bioinformatics*, **25** (1), pp. 126-127.

Mika, F., Hengge, R. (2013). Small Regulatory RNAs in the control of motility and biofilm formation in *E. coli* and *Salmonella*. *International Journal of Molecular Sciences*, **14** (3), pp. 4560-4579.

Mizuguchi, H., Nakatsuji, M., Fujiwara, S., Takagi, M., & Imanaka, T. (1999). Characterization and application to hot start PCR of neutralizing monoclonal antibodies against KOD DNA polymerase, *Journal of Biochemistry*, **126** (4), pp. 762-768.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server, *Nucleic Acids Research*, **35** (suppl 2), p. W182-W185.

Moura-Costa, L. F. *et al.* (2008). Evaluation of the humoral and cellular immune response to different antigens of *Corynebacterium pseudotuberculosis* in Caninde goats and their potential protection against caseous lymphadenitis, *Veterinary Immunology and Immunopathology*, **126** (1-2), pp. 131-141.

Muckle, C. A. & Gyles, C. L. (1983). Relation of lipid content and exotoxin production to virulence of *Corynebacterium pseudotuberculosis* in mice, *American Journal of Veterinary Research*, **44** (6), pp. 1149-1153.

Muckle, C. A. & Gyles, C. L. (1986). Exotoxic activities of *Corynebacterium pseudotuberculosis*, *Current Microbiology*, **13** (2), pp. 57-60.

Mullen, L. M., Nair, S. P., Ward, J. M., Rycroft, A. N. & Henderson, B. (2006). Phage display in the study of infectious diseases, *Trends in Microbiology*, **14** (3), pp. 141-147.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing, *Science*, **320** (5881), pp. 1344-1349.

Nairn, M. E. & Robertson, J. P. (1974). *Corynebacterium pseudotuberculosis* infection of sheep: role of skin lesions and dipping fluids, *Australian Veterinary Journal*, **50** (12), pp. 537-542.

Nakamura, Y., Nishio, Y., Ikeo, K., & Gojobori, T. (2003). The genome stability in *Corynebacterium* species due to lack of the recombinational repair system, *Gene*, **317** (0), pp. 149-155.

Nan, J., Brostromer, E., Liu, X. Y., Kristensen, O., & Su, X. D. (2009). Bioinformatics and structural characterization of a hypothetical protein from *Streptococcus mutans*: Implication of antibiotic resistance, *PLoS ONE*, **4** (10), p. e7245.

Nathan, C. & Shiloh, M. U. (2000). Reactive oxygen and nitrogen intermediates in the relationship between mammalian hosts and microbial pathogens, *Proceedings of the National Academy of Sciences,* **97** (16), pp. 8841-8848.

Nielsen, H., Engelbrecht, J., Brunak, S., & vonHeijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites, *Protein Engineering*, **10** (1), pp. 1-6.

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment, *Journal of Molecular Biology,* **302** (1), pp. 205-217.

Odds, F. C. (2005). Genomics, molecular targets and the discovery of antifungal drugs., *Revista Iberoamericana de Micología.*, **22** (229), p. 237.

Oliveira, L., Madureira, P., Andrade, E. B., Bouaboud, A., Morello, E., Ferreira, P., Poyart, C., Trieu-Cuot, P. & Dramsi, S. (2012). Group B Streptococcus GAPDH is released upon cell lysis, associates with bacterial surface, and induces apoptosis in murine macrophages, *PLos One*, **7**(1), pp. e29963

Olson, M. E., Ceri, H., Morck, D. W., Buret, A. G., & Read, R. R. (2002). Biofilm bacteria: formation and comparative susceptibility to antibiotics, *Canadian Journal of Veterinary Research,* **66** (2), pp. 86-92.

Pacheco, L. G., Pena, R. R., Castro, T. L., Dorella, F. A., Bahia, R. C., Carminati, R., Frota, M. N., Oliveira, S. C., Meyer, R., Alves, F. S., Miyoshi, A. & Azevedo, V. (2007). Multiplex PCR assay for identification of *Corynebacterium pseudotuberculosis* from pure cultures and for rapid detection of this pathogen in clinical samples. *Journal of Medical Microbiology,* **56** (4), pp. 480-486.

Pacheco, L. *et al.* (2011). A combined approach for comparative exoproteome analysis of *Corynebacterium pseudotuberculosis*, *BMC Microbiology*, **11** (1), p. 12.

Pagliero, E., Chesnel, L., Hopkins, J., Croizé, J., Dideberg, O., Vernet, T. De Guilmi, A. M. (2004). Biochemical characterization of *Streptococcus pneumoniae* peniciilin-binding protein 2b and its implication in β-lactam resistance, *Antimicrobial Agents % Chemotherapy*, **48** (5), pp. 1848-1855.

Pallen, M. J. (2002). The ESAT-6/WXG100 superfamily––and a new Gram-positive secretion system?, *Trends in Microbiology*, **10** (5), pp. 209-212.

Pallen, M. J., Lam, A. C., Antonio, M., & Dunbar, K. (2001). An embarrassment of sortases - a richness of substrates?, *Trends in Microbiology*, **9** (3), pp. 97-101.

Papaioannou, N., Zavlaris, M., Giadinis, N. D., Petridou, E. J., & Psychas, V. (2010). A case of kidney infection by *Corynebacterium pseudotuberculosis* in sheep, *Journal of the Hellenic Veterinary Medical Society*, **61** (1), pp. 29-35.

Pascual, C., Lawson, P. A., Farrow, J. A., Gimenez, M. N. & Collins, M. D. (1995). Phylogenetic analysis of the genus *Corynebacterium* based on 16S rRNA gene sequences, *International Journal of Systemic Bacteriology*, **45** (4), pp. 724-728.

Paton, M., Rose, I., Hart, R., Sutherland, S., Mercy, A., & Ellis, T. (1996). Post-shearing management affects the seroincidence of *Corynebacterium pseudotuberculosis* infection in sheep flocks, *Preventive Veterinary Medicine*, **26** (3-4), pp. 275-284.

Paton, M. W., Mercy, A. R., Wilkinson, F. C., Gardner, J. J., Sutherland, S. S., & Ellis, T. M. (1988). The effects of caseous lymphadenitis on wool production and bodyweight in young sheep, *Australian Veterinary Journal*, **65** (4), pp. 117-119.

Paton, M. W., Rose, I. R., Hart, R. A., Sutherland, S. S., Mercy, A. R., Ellis, T. M., & Dhaliwal, J. A. (1994). New infection with *Corynebacterium pseudotuberculosis* reduces wool production, *Australian Veterinary Journal*, **71** (2), pp. 47-49.

Paton, M. W., Buller, N. B., Rose, I. R., & Ellis, T. M. (2002). Effect of the interval between shearing and dipping on the spread of *Corynebacterium pseudotuberculosis* infection in sheep, *Australian Veterinary Journal*, **80** (8), pp. 494-496.

Paton, M. W., Walker, S. B., Rose, I. R., & Watt, G. F. (2003). Prevalence of caseous lymphadenitis and usage of caseous lymphadenitis vaccines in sheep flocks, *Australian Veterinary Journal*, **81** (1-2), pp. 91-95.

Paton, M. W., Collett, M. G., Pepin, M., & Bath, G. F. 2005, "*Corynebacterium pseudotuberculosis* infections.," in *Infectious Diseases of Livestock*, 3rd edn, A. W. Coetzer & R. C. Tustin, eds., Oxford University Press Southern Africa, CapeTown, pp. 1917-1930.

Pavan, M. E., Robles, C., Cairó, F. M., Marcellino, R. & Pettinari, M. J. (2012). Identification of *Corynebacterium pseudotuberculosis* from sheep by PCR-restriction analysis using the RNA polymerase β-subunit gene (*rpoB*), **92** (2), pp.202-206.

Pearson, W. R. (1991), Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms, *Genomics*, **11** (3), pp. 635-650.

Peden, J. (1999). Analysis of codon usage. PhD thesis, University of Nottingham.

Peel, M. M., Palmer, G. G., Stacpoole, A. M., & Kerr, T. G. (1997). Human lymphadenitis due to *Corynebacterium pseudotuberculosis*: report of ten cases from Australia and review, *Clinical Infectious Diseases*, **24** (2), pp. 185-191.

Pepin, M., Fontaine, J. J., Pardon, P., Marly, J., & Parodi, A. L. (1991). Histopathology of the early phase during experimental *Corynebacterium pseudotuberculosis* infection in lambs, *Veterinary Microbiology*, **29** (2), pp. 123-134.

Pepin, M., Pardon, P., Marly, J., Lantier, F., & Arrigo, J. L. (1993). Acquired immunity after primary caseous lymphadenitis in sheep, *American Journal of Veterinary Research*, **54** (6), pp. 873-877.

Perkins, D. N., Pappin, D. J. C., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data, *Electrophoresis*, **20** (18), pp. 3551-3567.

Pethick, F. E. *et al.* (2012a). Complete genome sequence of *Corynebacterium pseudotuberculosis* strain 1/06-A, isolated from a horse in North America, *Journal of Bacteriology*, **194** (16), p. 4476.

Pethick, F. E. *et al.* (2012b). Complete genome sequences of *Corynebacterium pseudotuberculosis* strains 3/99-5 and 42/02-A, isolated from sheep in Scotland and Australia, respectively, *Journal of Bacteriology*, **194** (17), pp. 4736-4737.

Piontkowski, M. D. & Shivvers, D. W. (1998). Evaluation of a commercially available vaccine against *Corynebacterium pseudotuberculosis* for use in sheep, *Journal of the American Veterinary Medical Association*, **212** (11), pp. 1765-1768.

Pischimarov, J., Kuenne, C., Billion, A., Hemberger, J., Cemič, F., Chakraborty, T. & Hain, T. (2012). sRNAdb: a small non-coding RNA database for gram-positive bacteria, *BMC Genomics*, **13**, pp. 384-391.

Pizza, M. *et al.* (2000). Identification of vaccine candidates against Serogroup B Meningococcus by whole-genome sequencing, *Science*, **287** (5459), pp. 1816-1820.

Popp, M. W. & Ploegh, H. L. (2011). Making and breaking peptide bonds: protein engineering using sortase, *Angewandte Chemie International Edition*, **50** (22), pp. 5024-5032.

Power, E. G. (1996). RAPD typing in microbiology-a technical review, *Journal of Hospital Infection*, **34** (4), pp. 247-265.

Prescott, J. F., Menzies, P. I., & Hwang, Y. T. (2002). An interferon-gamma assay for diagnosis of *Corynebacterium pseudotuberculosis* infection in adult sheep from a research flock, *Veterinary Microbiology*, **88** (3), pp. 287-297.

Quinn, P. J., Carter, M. E., Markey, B., & Carter, G. R. (2009). "*Corynebacterium* species and *Rhodococcus equi*," in *Clinical veterinary microbiology*, pp. 137-143.

Radostits, O. M., Gay, C. C., Blood, D. C., & Hinchcliff, K. W. (2000). *Veterinary Medicine 9th Edition* Saunders, London.

Rainey, F. A., Ward-Rainey, N. L., Janssen, P. H., Hippe, H., & Stackebrandt, E. (1996). *Clostridium paradoxum* DSM 7308T contains multiple 16S rRNA genes with heterogeneous intervening sequences, *Microbiology*, **142** (8), pp. 2087-2095.

Rannala, B. & Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference, *Journal of Molecular Evolution*, **43** (3), pp. 304-311.

Rappuoli, R. (2000). Reverse vaccinology, *Current Opinion in Microbiology*, **3** (5), pp. 445-450.

Ratledge, C. & Dover, L. G. (2000). Iron metabolism in pathogenic bacteria, *Annual Review of Microbiology*, **54** (1), pp. 881-941.

Rebouças, M. F. *et al.* (2011). *Corynebacterium pseudotuberculosis* secreted antigen-induced specific gamma-interferon production by peripheral blood leukocytes: potential diagnostic marker for caseous lymphadenitis in sheep and goats, *Journal of Veterinary Diagnostic Investigation*, **23** (2), pp. 213-220.

Renshaw, H. W., Graff, V. P., & Gates, N. L. (1979). Visceral caseous lymphadenitis in thin ewe syndrome-isolation of *Corynebacterium*, *Staphylococcus*, and *Moraxella spp* from internal abscesses in emaciated ewes, *American Journal of Veterinary Research*, **40** (8), pp. 1110-1114.

Restrepo-Montoya, D., Vizcaino, C., Nino, L. F., Ocampo, M., Patarroyo, M. E., & Patarroyo, M. A. (2009). Validating subcellular localization prediction tools with mycobacterial proteins, *BMC Bioinformatics*, **10**, pp 134-142.

Rhem, M. N., Lech, E. M., Patti, J. M., McDevitt, D., Hook, M., Jones, D. B., & Wilhelmus, K. R. (2000). The collagen-binding adhesin is a virulence factor in *Staphylococcus aureus* keratitis, *Infection and Immunity*, **68** (6), pp. 3776-3779.

Rivera, J., Vannakambadi, G., Hook, M., & Speziale, P. (2007). Fibrinogen-binding proteins of Gram-positive bacteria, *Thrombosis and Haemostasis*, **98** (3), pp. 503-511.

Rocha, E. P. C., Cornet, E., & Michel, B. (2005). Comparative and evolutionary analysis of the bacterial homologous recombination systems, *PLoS Genetics*, **1** (2), p. e15.

Rothberg, J. M. & Leamon, J. H. (2008). The development and impact of 454 sequencing, *Nature Biotechnology*, **26** (10), pp. 1117-1124.

Rothel, J. S, Jones, S. L., Corner, L.A., Cox, J.C. & Wood, P. R. (1990). A sandwich immunoassay for bovine interferon-gamma and its use for the detection of tuberculosis in cattle, *Australian Veterinary Journal*, **67** (4), pp. 134-137.

Ruimy, R., Riegel, P., Boiron, P., Monteil, H., & Christen, R. (1995). Phylogeny of the genus *Corynebacterium* deduced from analyses of small-subunit ribosomal DNA sequences, *International Journal of Systematic Bacteriology*, **45** (4), pp. 740-746.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., & Barrell, B. (2000). Artemis: sequence visualization and annotation, *Bioinformatics*, **16** (10), pp. 944-945.

Sadarangani, M. & Pollard, A. J. (2010). Serogroup B meningococcal vaccines–an unfinished story, *The Lancet Infectious Diseases*, **10** (2), pp. 112-124.

Sanciu, G., Marogna, G., Paglietti, B., Cappuccinelli, P., Leori, G., & Rappelli, P. (2012). Outbreak of mastitis in sheep caused by multi-drug resistant *Enterococcus faecalis* in Sardinia, Italy, *Epidemiology & Infection*, **FirstView** pp. 1-3.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors, *Proceedings of the National Academy of Sciences*, **74** (12), pp. 5463-5467.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage [phi]X174 DNA, *Nature*, **265** (5596), pp. 687-695.

Sauvage, E., Kerff, F., Terrak, M., Ayala, J. A. & Charlier, P. (2008). The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis, *FEMS Microbiology Reviews*, **32** (2), pp. 234-258.

Schmiel, D. H. & Miller, V. L. (1999). Bacterial phospholipases and pathogenesis, *Microbes and Infection*, **1** (13), pp. 1103-1112.

Schmitt, M. P. (1997). Utilization of host iron sources by *Corynebacterium diphtheriae*: identification of a gene whose product is homologous to eukaryotic heme oxygenases and is required for acquisition of iron from heme and hemoglobin, *Journal of Bacteriology*, **179** (3), pp. 838-845.

Schreuder, B. E. C., Terlaak, E. A., & Dercksen, D. P. (1994). Eradication of caseous lymphadenitis in sheep with the help of a newly developed ELISA technique, *Veterinary Record*, **135** (8), pp. 174-176.

Selim, S. A. (2001). Oedematous skin disease of buffalo in Egypt, *Journal of Veterinary Medicine Series B-Infectious Diseases and Veterinary Public Health*, **48** (4), pp. 241-258.

Senturk, S. & Temizel, M. (2006). Clinical efficacy of rifamycin SV combined with oxytetracycline in the treatment of caseous lymphadenitis in sheep, *Veterinary Record*, **159** (7), pp. 216-217.

Shapiro, B., Rambaut, A., & Drummond, A. J. (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences, *Molecular Biology and Evolution*, **23** (1), pp. 7-9.

Shen, H. B. & Chou, K. C. (2007). Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins, *Protein Engineering Design & Selection*, **20** (1), pp. 39-46.

Shigidi, M. T. A. (1978). Indirect hemagglutination test for sero-diagnosis of *C. ovis* infection in sheep, *Research in Veterinary Science*, **24** (1), pp. 57-60.

Shiloh, M. U., MacMicking, J. D., Nicholson, S., Brause, J. E., Potter, S., Marino, M., Fang, F., Dinauer, M., & Nathan, C. (1999). Phenotype of mice and macrophages deficient in both phagocyte oxidase and inducible nitric oxide synthase, *Immunity*, **10** (1), pp. 29-38.

Shpigel, N. Y., Elad, D., Yeruham, I., Winkler, M., & Saran, A. (1993). An outbreak of *Corynebacterium pseudotuberculosis* infection in an Israeli dairy herd, *Veterinary Record*, **133** (4), pp. 89-94.

Silva, A. *et al.* (2011). Complete genome sequence of *Corynebacterium pseudotuberculosis* I19, a strain isolated from a cow in Israel with bovine mastitis, *Journal of Bacteriology*, **193** (1), pp. 323-324.

Simmons, C. P., Hodgson, A. L. M., & Strugnell, R. A. (1997). Attenuation and vaccine potential of aroQ mutants of *Corynebacterium pseudotuberculosis*, *Infection and Immunity*, **65** (8), pp. 3048-3056.

Smith, G.P. (1985). Filamentous fusion phage: novel expression vectors that displayed cloned antigens on the virion surface, *Science*, **228** (4705), pp. 1315-1317.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences, *Journal of Molecular Biology*, **147** (1), pp. 195-197.

Sohling, B. & Gottschalk, G. (1996). Molecular analysis of the anaerobic succinate degradation pathway in *Clostridium kluyveri*, *Journal of Bacteriology*, **178** (3), pp. 871-880.

Sommer, D., Delcher, A., Salzberg, S., & Pop, M. (2007). Minimus: a fast, lightweight genome assembler, *BMC Bioinformatics*, **8** (1), p. 64.

Songer, J. G., Beckenbach, K., Marshall, M. M., Olson, G. B., & Kelley, L. (1988), Biochemical and genetic characterization of *Corynebacterium pseudotuberculosis*, *American Journal of Veterinary Research*, **49** (2), pp. 223-226.

Soriano, F., Zapardiel, J., & Nieto, E. (1995). Antimicrobial susceptibilities of *Corynebacterium* species and other non-spore-forming gram-positive bacilli to 18 antimicrobial agents, *Antimicrobial Agents and Chemotherapy*, **39** (1), pp. 208-214.

Souckova, A. & Soucek, A. (1972). Inhibition of the hemolytic action of α and β lysins of *Staphylococcal pyogenes* by *Corynebacterium hemolytica, C. ovis* and *C. ulcerans, Toxicon*, **10** (5), pp. 501-504.

Spier, S. J., Leutenegger, C. M., Carroll, S. P., Loye, J. E., Pusterla, J. B., Carpenter, T. E., Mihalyi, J. E., & Madigan, J. E. (2004). Use of a real-time polymerase chain reaction-based fluorogenic 5 ' nuclease assay to evaluate insect vectors of *Corynebacterium pseudotuberculosis* infections in horses, *American Journal of Veterinary Research*, **65** (6), pp. 829-834.

St.Geme, J. W. & Falkow, S. (1994). A *Haemophilus influenzae* IgA protease-like protein promotes intimate interaction with human epithelial cells, *Molecular Microbiology*, **14** (2), pp. 217-233.

Stevenson, B. S. & Schmidt, T. M. (2004). Life history implications of rRNA gene copy number in *Escherichia coli*, *Applied and Environmental Microbiology*, **70** (11), pp. 6670-6677.

Sting, R., Steng, G., & Spengler, D. (1998). Serological studies on *Corynebacterium pseudotuberculosis* infections in goats using enzyme-linked immunosorbent assay, *Journal of Veterinary Medicine Series B-Infectious Diseases and Veterinary Public Health*, **45** (4), pp. 209-216.

Stoops, S. G., Renshaw, H. W., & Thilsted, J. P. (1984), Ovine caseous lymphadenitis: Disease prevalence, lesion distribution, and thoracic manifestations in a population of mature culled sheep from western United States, *American Journal of Veterinary Research*, **45** (3), pp. 557-561.

Strätz, M., Mau, M., & Timmis, K. N. (1996). System to study horizontal gene exchange among microorganisms without cultivation of recipients, *Molecular Microbiology*, **22** (2), pp. 207-215.

Sunil, V., Menzies, P. I., Shewen, P. E., & Prescott, J. F. (2008). Performance of a whole blood interferon-gamma assay for detection and eradication of caseous lymphadenitis in sheep, *Veterinary Microbiology*, **128** (3-4), pp. 288-297.

Sutherland, S. S., Ellis, T. M., Mercy, A. R., Paton, M., & Middleton, H. (1987). Evaluation of an Enzyme-Linked-Immunosorbent-Assay for the detection of *Corynebacterium pseudotuberculosis* infection in sheep, *Australian Veterinary Journal*, **64** (9), pp. 263-266.

Tabor, C. W. & Tabor, H. (1985). Polyamines in microorganisms, *FEMS Microbiology Reviews*, **49** (1), pp. 81-99.

Talavera, G. & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments, *Systematic Biology*, **56** (4), pp. 564-577.

Tatusov, R. *et al.* (2003). The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, **4** (1), p. 41.

Tauch, A., Bischoff, N., Brune, I., & Kalinowskia, J. (2003). Insights into the genetic organization of the *Corynebacterium diphtheriae* erythromycin resistance plasmid pNG2 deduced from its complete nucleotide sequence, *Plasmid*, **49** (1), pp. 63-74.

Tenorio, E., Saeki, T., Fujita, K., Kitakawa, M., Baba, T., Mori, H., & Isono, K. (2003). Systematic characterization of *Escherichia coli* genes/ORFs affecting biofilm formation, *FEMS Microbiology Letters*, **225** (1), pp. 107-114.

Ter Laak, E. A., Bosch, J., Bijl, G. C., & Schreuder, B. E. C. (1992). Double-antibody sandwich Enzyme-Linked Immunosorbent Assay and immunoblot analysis used for control of caseous lymphadenitis in goats and sheep, *American Journal of Veterinary Research*, **53** (7), pp. 1125-1132.

Tettelin, H. *et al.* (2000). Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58, *Science*, **287** (5459), pp. 1809-1815.

Tettelin, H. *et al.* (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome', *Proceedings of the National Academy of Sciences of the United States of America*, **102** (39), pp. 13950-13955.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Research*, **22** (22), pp. 4673-4680.

Thorpe, C., Edwards, L., Snelgrove, R., Finco, O., Rae, A., Grandi, G., Guilio, R., & Hussell, T. (2007). Discovery of a vaccine antigen that protects mice from *Chlamydia pneumoniae* infection, *Vaccine*, **25** (12), pp. 2252-2260.

Ton-That, H. & Schneewind, O. (2003). Assembly of pili on the surface of *Corynebacterium diphtheriae*, *Molecular Microbiology*, **50** (4), pp. 1429-1438.

Trost, E. *et al.* (2010). The complete genome sequence of *Corynebacterium pseudotuberculosis* FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence, *BMC Genomics*, **11** (1), p. 728.

Trost, E. *et al.* (2011). Comparative analysis of two complete *Corynebacterium ulcerans* genomes and detection of candidate virulence factors, *BMC Genomics*, **12** (1), p. 383.

Upadhye, V., Majumdar, A., Gomashe, A., Joshi, D., Gangane, N., Thamke, D., Mendiratta, D., & Harinath, B. C. (2009). Inhibition of *Mycobacterium tuberculosis* secretory serine protease blocks bacterial multiplication both in axenic culture and in human macrophages., *Scandinavian Journal of Infectious Diseases*, **41** (8), pp. 569-576.

van Nimwegen, E., Zavolan, M., Rajewsky, N., & Siggia, E. D. (2002). Probabilistic clustering of sequences: Inferring new bacterial regulons by comparative genomics, *Proceedings of the National Academy of Sciences*, **99** (11), pp. 7323-7328.

Vandamme, A. M. (2003). "Basic concepts of molecular evolution," in *The Phylogenetic Handbook: A practical approach to DNA and protein phylogeny*, M. Salemi & A. M. Vandamme, eds., Cambridge University Press, UK, pp. 1-23.

Viguetti, S. Z., Pacheco, L. G. C., Santos, L. S., soares, S. C., Bolt, F., Baldwin, A., Dowson, C. G., Rosso, M. L., Guiso, N., Miyoshi, A., Hirata, R., Mattos-Guaraldi, A. L. & Azvedo, V. (2012). Multilocus sequence types of invasive *Corynebacterium diphtheriae* isolated in the Rio de Janeiro urban area, Brazil, *Epidemiology and Infection*, **140** pp. 617-620.

Villesen, P. (2007). FaBox: an online toolbox for fasta sequences, *Molecular Ecology Notes*, **7** (6), pp. 965-968.

Vitreschak, A. G., Mironov, A. A., Lyubetsky, V. A., & Gelfand, M. S. (2008). Comparative genomic analysis of T-box regulatory systems in bacteria, *RNA*, **14** (4), pp. 717-735.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W., Surovcik, K., Meinicke, P., & Merkl, R. (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models, *BMC Bioinformatics*, **7** (1), p. 142.

Wagner, K. S., White, J. M., Crowcroft, N. S., De Martin, S., Mann, G., & Efstratiou, A. (2010). Diphtheria in the United Kingdom, 1986-2008: the increasing role of *Corynebacterium ulcerans.*, *Epidemiology and Infection*, **138** (11), pp. 1519-1530.

Walker, C. A. (2009). Iron-dependent regulation of gene expression in *Corynebacterium pseudotuberculosis,* PhD thesis, University of Glasgow.

Walker, J., Jackson, H. J., Eggleton, D. G., Meeusen, E. N. T., Wilson, M. J., & Brandon, M. R. (1994). Identification of a novel antigen from *Corynebacterium pseudotuberculosis* that protects sheep against caseous lymphadenitis, *Infection and Immunity*, **62** (6), pp. 2562-2567.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics, *Nature Reviews Genetics*, **10** (1), pp. 57-63.

Wankhede, G., Majumdar, A., Kamble, P. D., & Harinath, B. C. (2011). Mycobacterial secretory SEVA TB ES-31 antigen, a chymotrypsin-like serine protease with lipase activity and drug target potential, *Biomedial Research*, **22** (1), pp. 45-48.

Wassarman, K. M. (2002). Small RNAs in bacteria: diverse regulators of gene expression in response to environmental changes, *Cell,* **109** (2) pp. 141-144.

Waters, L. S. & Storz, G. (2009). Regulatory RNAs in bacteria, *Cell,* **136** (4), pp. 615-628.

Watts, J. L., Lowery, D. E., Teel, J. F., & Rossbach, S. (2000). Identification of *Corynebacterium bovis* and other coryneforms isolated from bovine mammary glands, *Journal of Dairy Science,* **83** (10), pp. 2373-2379.

Widom, R. L., Jarvis, E. D., LaFauci, G., & Rudner, R. (1988). Instability of rRNA operons in *Bacillus subtilis, Journal of Bacteriology,* **170** (2), pp. 605-610.

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J., & Bahler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution, *Nature,* **453** (7199), pp. 1239-1243.

Williams, T. L., Monday, S. R., Feng, P. C. H. & Musser, S. M. (2005). Identifying new PCR targets for pathogenic bacteria using top-down LC/MS protein discovery, *Journal of Biomedical Techniques*, **16** (2), pp. 134-142.

Williamson, L. H. (2001). Caseous lymphadenitis in small ruminants, *Veterinary Clinics of North America: Food Animal Practice,* **17** (2), pp. 359-371.

Wilson, M. J., Brandon, M. R., & Walker, J. (1995). Molecular and biochemical characterization of a protective 40-kilodalton antigen from *Corynebacterium pseudotuberculosis*, *Infection and Immunity*, **63** (1), pp. 206-211.

Wizemann, T. M. *et al.* (2001). Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection, *Infection and Immunity,* **69** (3), pp. 1593-1598.

Wood, P. R., Corner, L. A., Rothel, J. S., Ripper, J. L., Fifis, T., McCormick, B. S., Francis, B., Melville, L., Small, K., De Witte, K., Tolson, J., Ryan, T.J., de Lisle, G.W., Cox, J.C. & Jones, S.L. (1992). A field evaluation of serological and cellular diagnostic tests for bovine tuberculosis. *Veterinary Microbiology*, **31**(1), pp. 71-79.

Wright, F. (1990). The effective number of codons used in a gene, *Gene,* **87** (1), pp. 23-29.

Wu, M. & Eisen, J. (2008). A simple, fast, and accurate method of phylogenomic inference, *Genome Biology,* **9** (10), pp. 1-11.

Xu, W. L., Shen, J. Y., Dunn, C. A., Desai, S., & Bessman, M. J. (2001). The Nudix hydrolases of *Deinococcus radiodurans, Molecular Microbiology,* **39** (2), pp. 286-290.

Yang, J. Y., Brooks, S., Meyer, J. A., Blakesley, J. A., Zelazny, A. M., Segre, J. A. & Snitkin, E. S. (2013). Pan-PCR, a computational method for designing bacterial-typing assays based on whole genome sequence data, *Journal of Clinical Microbiology,* **51**(3), pp. 752-758.

Yeruham, I., Elad, D., VanHam, M., Shpigel, N. Y., & Perl, S. (1997). *Corynebacterium pseudotuberculosis* infection in Israeli cattle: Clinical and epidemiological studies, *Veterinary Record*, **140** (16), pp. 423-427.

Yeruham, I., Elad, D., Friedman, S., & Perl, S. (2003). *Corynebacterium pseudotuberculosis* infection in Israeli dairy cattle, *Epidemiology and Infection*, **131** (2), pp. 947-955.

Yeruham, I., Elad, D., Perl, S., & Ram, A. (2003). Necrotic-ulcerative dermatitis on the heels of heifers in a dairy herd infected with *Corynebacterium pseudotuberculosis*, *Veterinary Record*, **152** (19), pp. 598-600.

Yozwiak, M. L. & Songer, J. G. (1993). Effect of *Corynebacterium pseudotuberculosis* phospholipase D on viability and chemotactic resonses of ovine neutrophils, *American Journal of Veterinary Research*, **54** (3), pp. 392-397.

Yu, C. S., Chen, Y. C., Lu, C. H., & Hwang, J. K. (2006). Prediction of protein subcellular localization, *Proteins-Structure Function and Bioinformatics*, **64** (3), pp. 643-651.

Yukawa, H. *et al.* (2007). Comparative analysis of the *Corynebacterium glutamicum* group and complete genome sequence of strain R, *Microbiology*, **153** (4), pp. 1042-1058.

Zaki, M. M. (1976). Relation between the toxogenicity and pyogenicity of *Corynebacterium ovis* in experimentally infected mice, *Research in Veterinary Science*, **20** (2), pp. 197-200.

Zankari, E., Hasman, H., Kaas, R. S., Seyfarth, A. M., Agers, Y., Lund, O., Larsen, M. V. & Aarestrup, F. M. (2013). Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing, *Journal of Antimicrobial Chemotherapy*, **68**, pp. 771-777.

Zavoshti, F., Khoojine, A., Helan, J., Hassanzadeh, B., & Heydari, A. (2011). Frequency of caseous lymphadenitis (CLA) in sheep slaughtered in an abattoir in Tabriz: comparison of bacterial culture and pathological study, *Comparative Clinical Pathology* pp. 1-5.

Zdobnov, E. M. & Apweiler, R. (2001). InterProScan-an integration platform for the signature-recognition methods in InterPro, *Bioinformatics*, **17** (9), pp. 847-848.

Zerbino, D. R. & Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs, *Genome Research*, **18** pp. 821-829.

Zhao, H. K., Yonekawa, K., Takahashi, T., Kikuchi, N., Hiramune, T., & Yanagawa, R. (1993). Isolation of *Corynebacterium pseudotuberculosis* from the cervical canal of clinically normal sows, *Research in Veterinary Science*, **55** (3), pp. 356-359.

Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y., & Yu, J. (2010). The next-generation sequencing technology and application, *Protein & Cell*, **1** (6), pp. 520-536.

Appendices

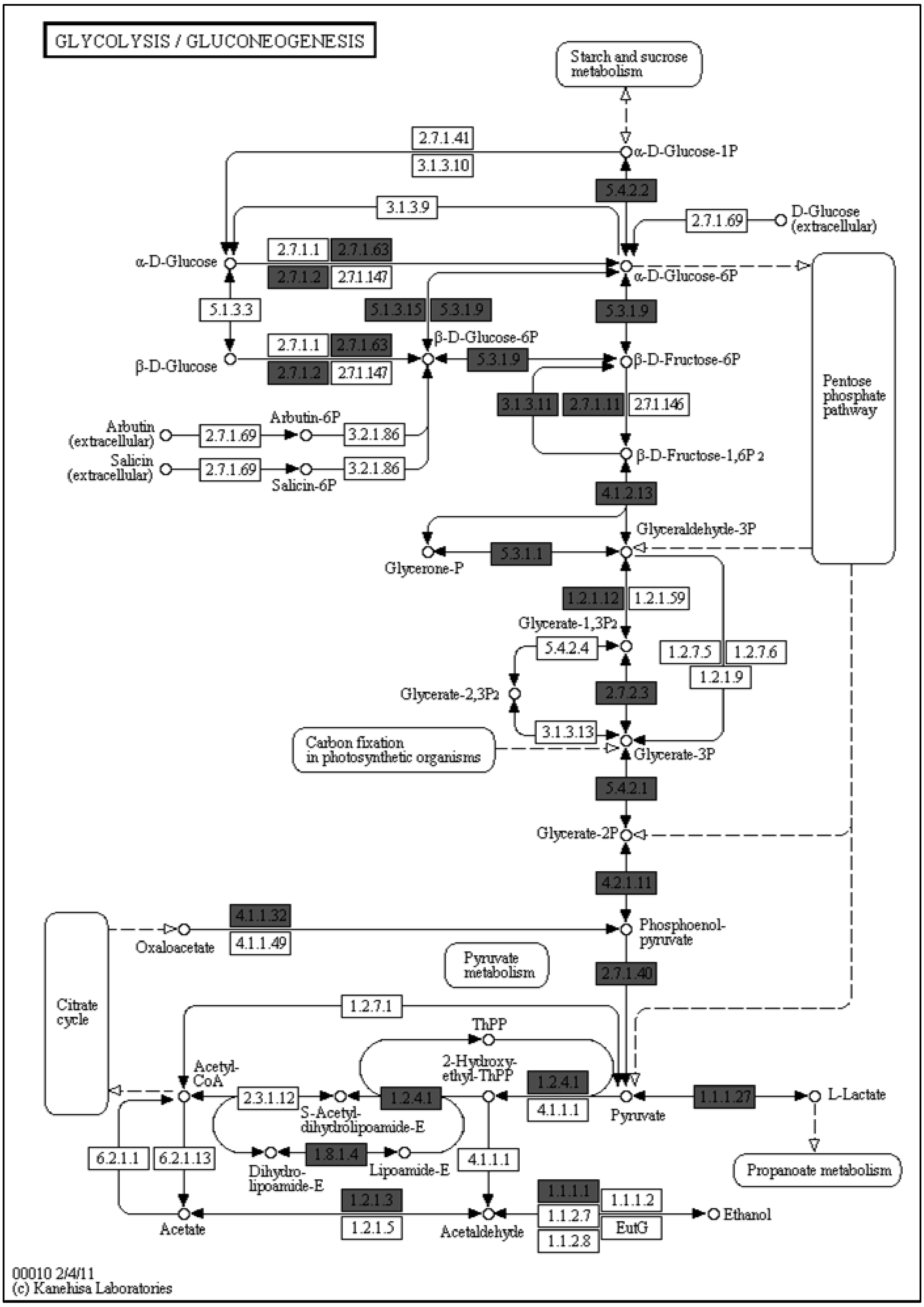
Appendix One: Common Buffers and reagents

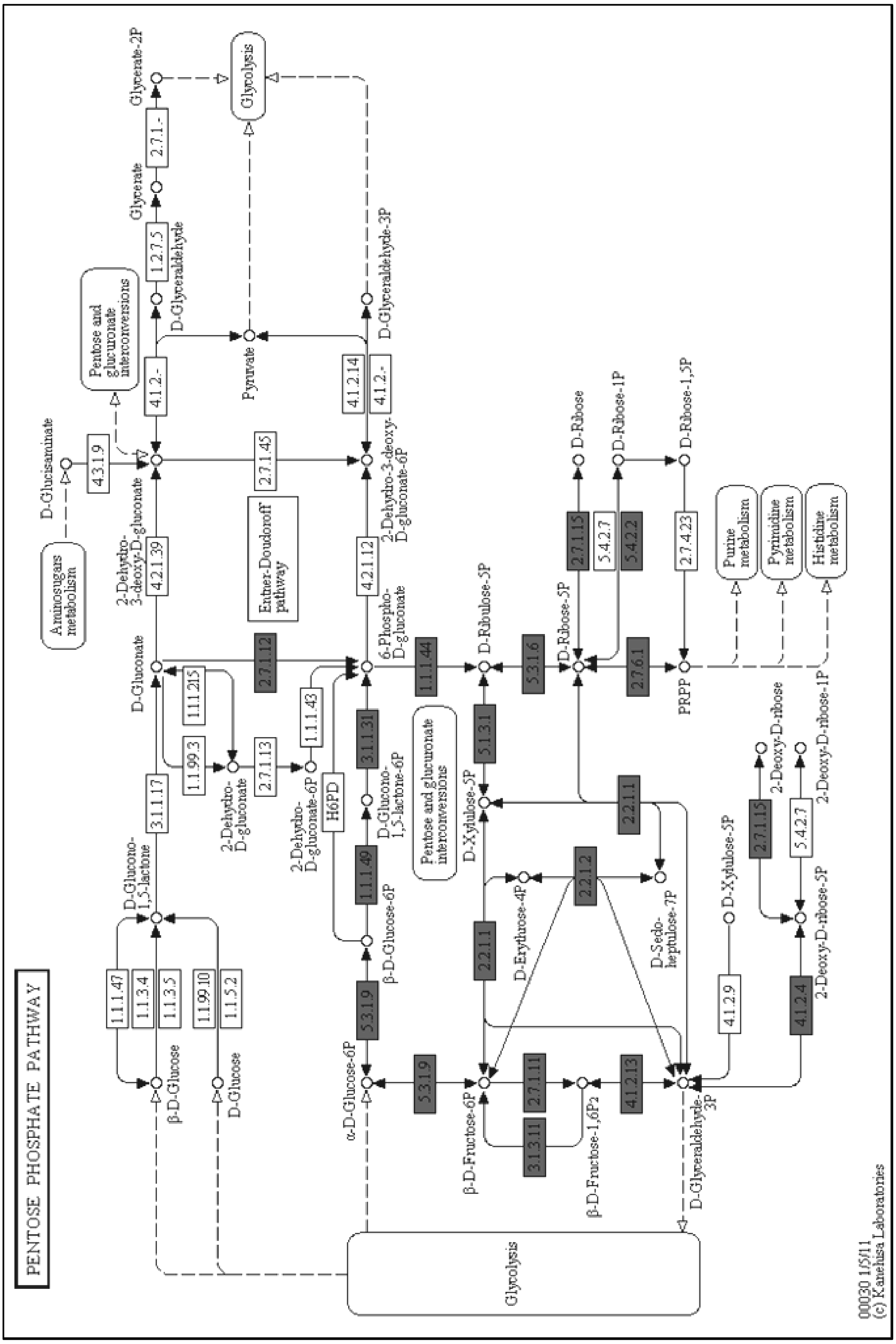
Reagent	Recipe
Binding Buffer	50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 10 mM imidazole, 1 × Protease Inhibitor Cocktail (Roche)
Blocking Buffer	1% (w/v) casein in Tris-buffered saline, (pH 7.4)
Buffer P1	50 mM Tris-HCl, 10 mM EDTA (pH 8.0)
Coating Buffer	50 mM carbonate-bicarbonate (pH 9.6) containing 1 µg/ml of antigen
Corynebacterium Chemically Defined Medium (CCDM)	10.38 g/L RPMI-1640 (Sigma, R 8755), 86 mM L-glutamic acid and 10 % (w/v) glucose, made in ddH2O and adjusted to pH 4, before addition of 0.2g/L NaHCO3 and pH adjustment to 7.1. Filter sterilised, addition of 1 × Amino acid solution (Sigma R7131).
Denaturing Binding Buffer	8 M Urea, 20 mM Na_2HPO_4 (pH 7.8), 500 mM NaCl
Denaturing Elution Buffer	8 M Urea, 20 mM Na_2HPO_4 (pH 4.0), 500 mM NaCl
Denaturing Wash Buffer	8 M Urea, 20 mM Na_2HPO_4 (pH 6.0), 500 mM NaCl
Dilution Buffer	1% Bovine serum albumin in Tris-buffered saline, (pH 7.4)
Elution Buffer	50 mM Tris-HCl, 300mM NaCl, 250 mM imidazole, I × Protease Inhibitor Cocktail (Roche)
Guanidinium Lysis Buffer	6 M Guanidine Hydrochloride, 20 mM Na_2HPO_4 (pH 7.8), 500 mM NaCl
Luria- Bertani (LB) Agar	1 % w/v Bacto-tryptone; 0.5 % w/v Bacto

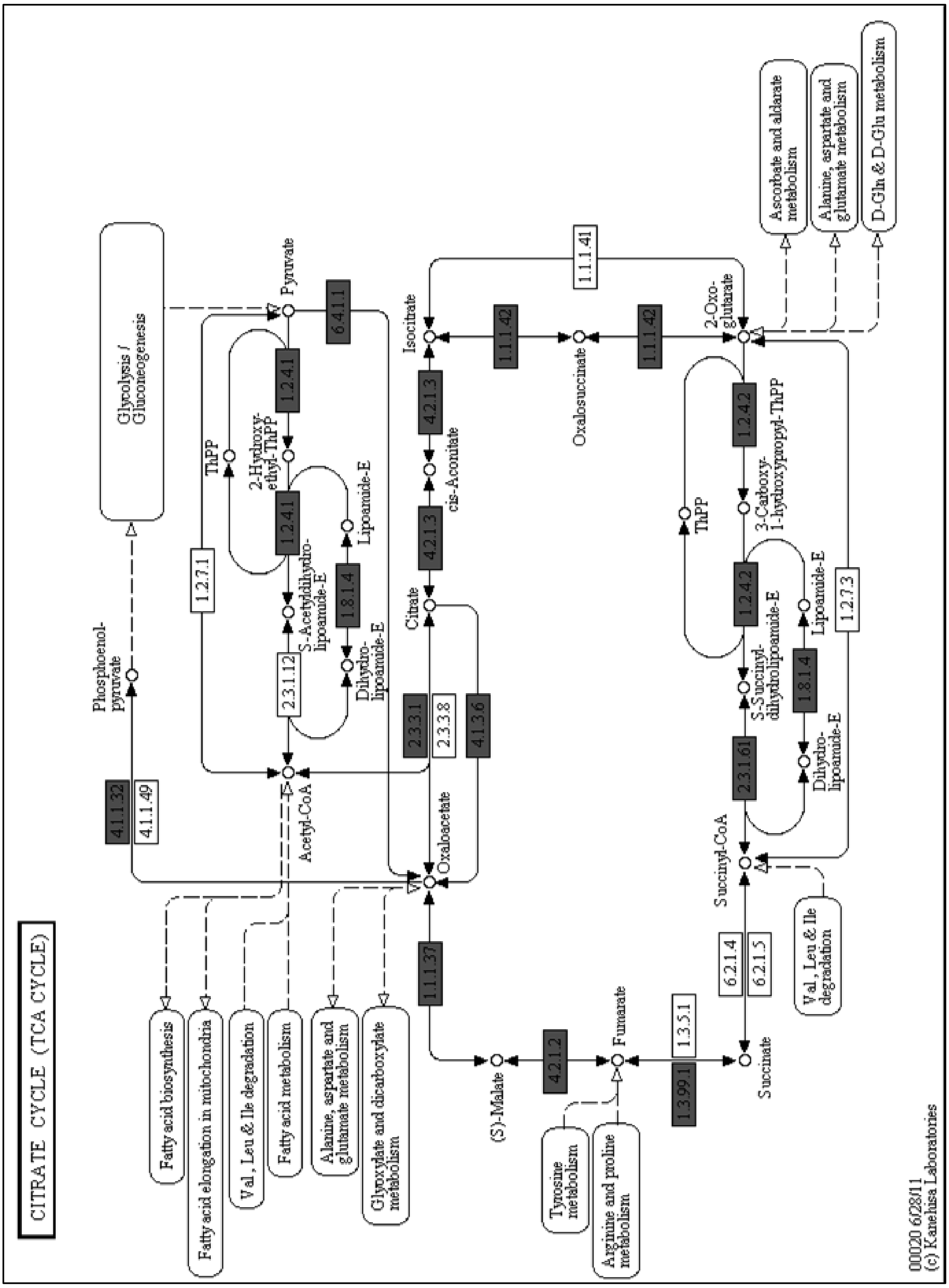
	yeast extract; 1 % w/v NaCl; 1.5 % agar
Luria- Bertani (LB) Broth	1 % w/v Bacto-tryptone; 0.5 % w/v Bacto yeast extract; 1 % w/v NaCl
Lysis Buffer	Binding buffer with the addition of 1 × BugBuster® solution (Novagen)
SDS-PAGE Running buffer (10 x)	25 mM Tris, 192mM glycine and 0.1 % SDS in 1 l ddH_2O
SOC medium	2% w/v tryptone, 0.5% w/v yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_{2,}$ 10 mM $MgSO_{4,}$ 20 mM glucose
TAE buffer	40 mM Tris-acetate, pH 8.0; 1 mM EDTA in ddH_2O
TE buffer	10mM Tris-HCl, pH 8.0; 1 mM EDTA
Tris-buffered saline (10x)	0.5 M Tris-HCl, pH 8.8; 1.5 M NaCl adjusted to pH 7.4 with HCl in ddH_2O
TBST	1 × Tris-buffered saline with 0.05% Tween-20
Tris-glycine buffer (transfer buffer)	25 mM Tris, 192 mM glycine, 20 % methanol
Wash Buffer	50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 20 mM imidazole, 0.1% Tween 20

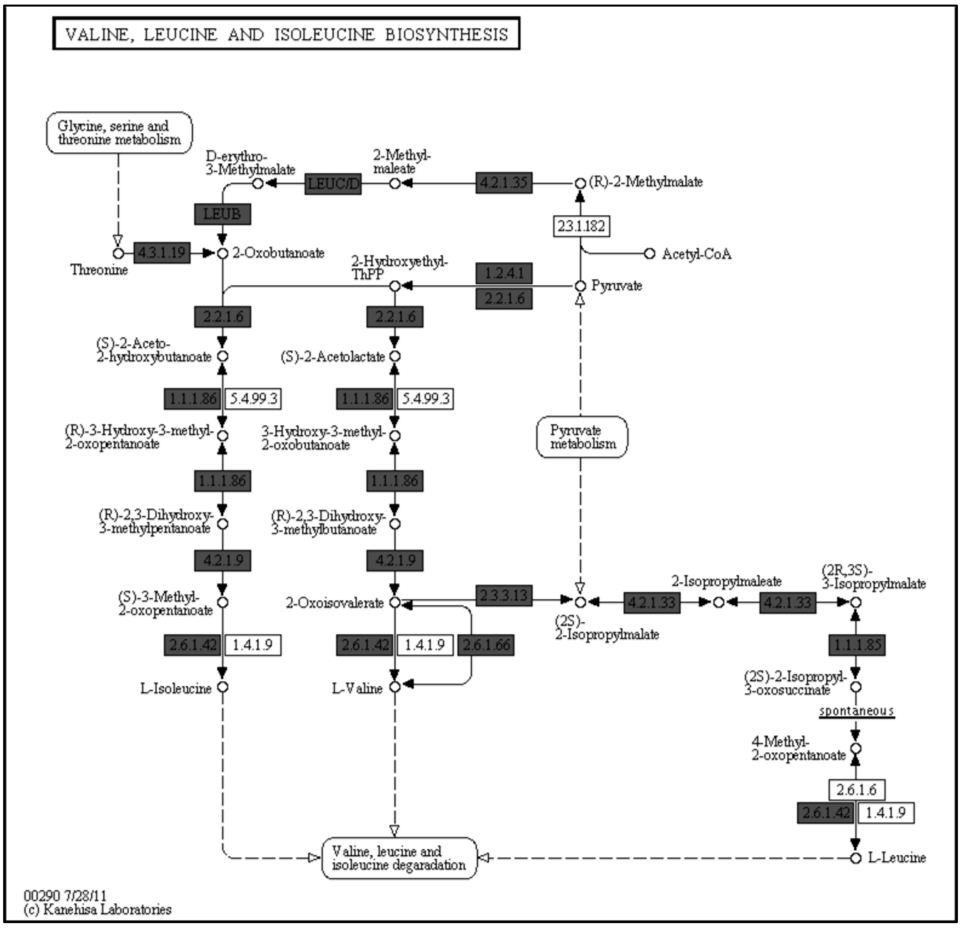
Appendix Two: Metabolic pathways of *Cp* 3/99-5

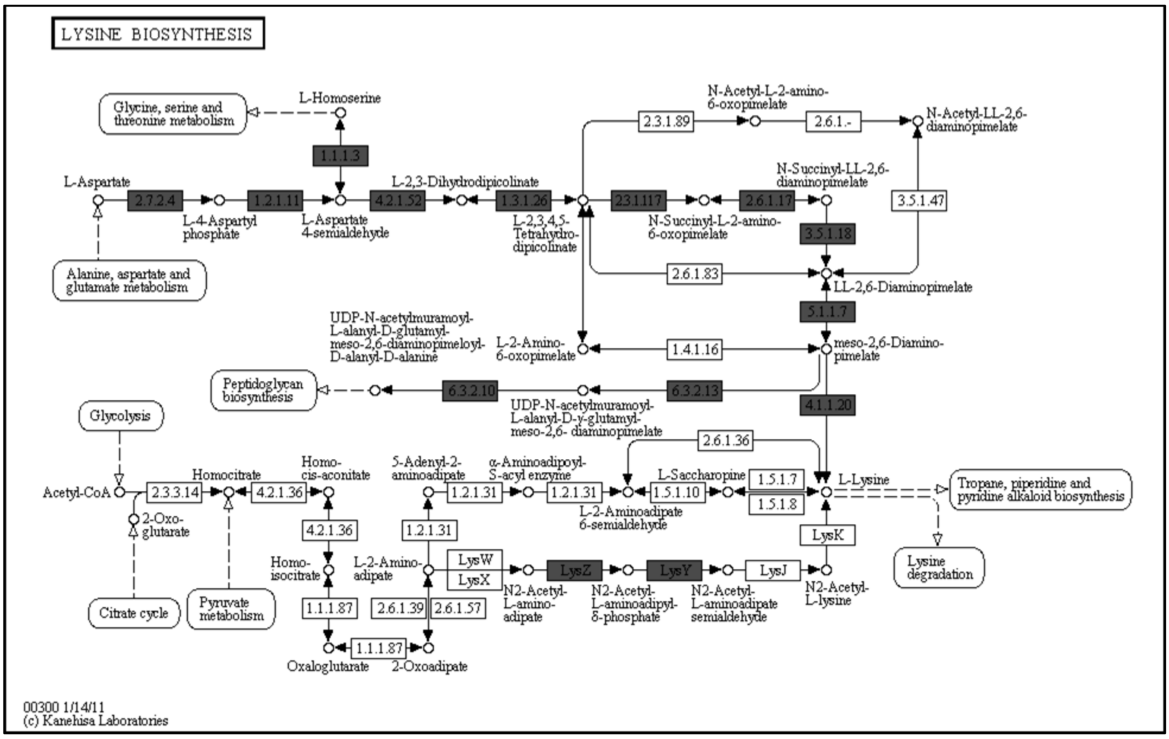
Metabolic pathways are reproduced from KEGG; the presence of genes in the *Cp* 3/99-5 genome is shown by shaded boxes, indicating the completeness of the pathway.

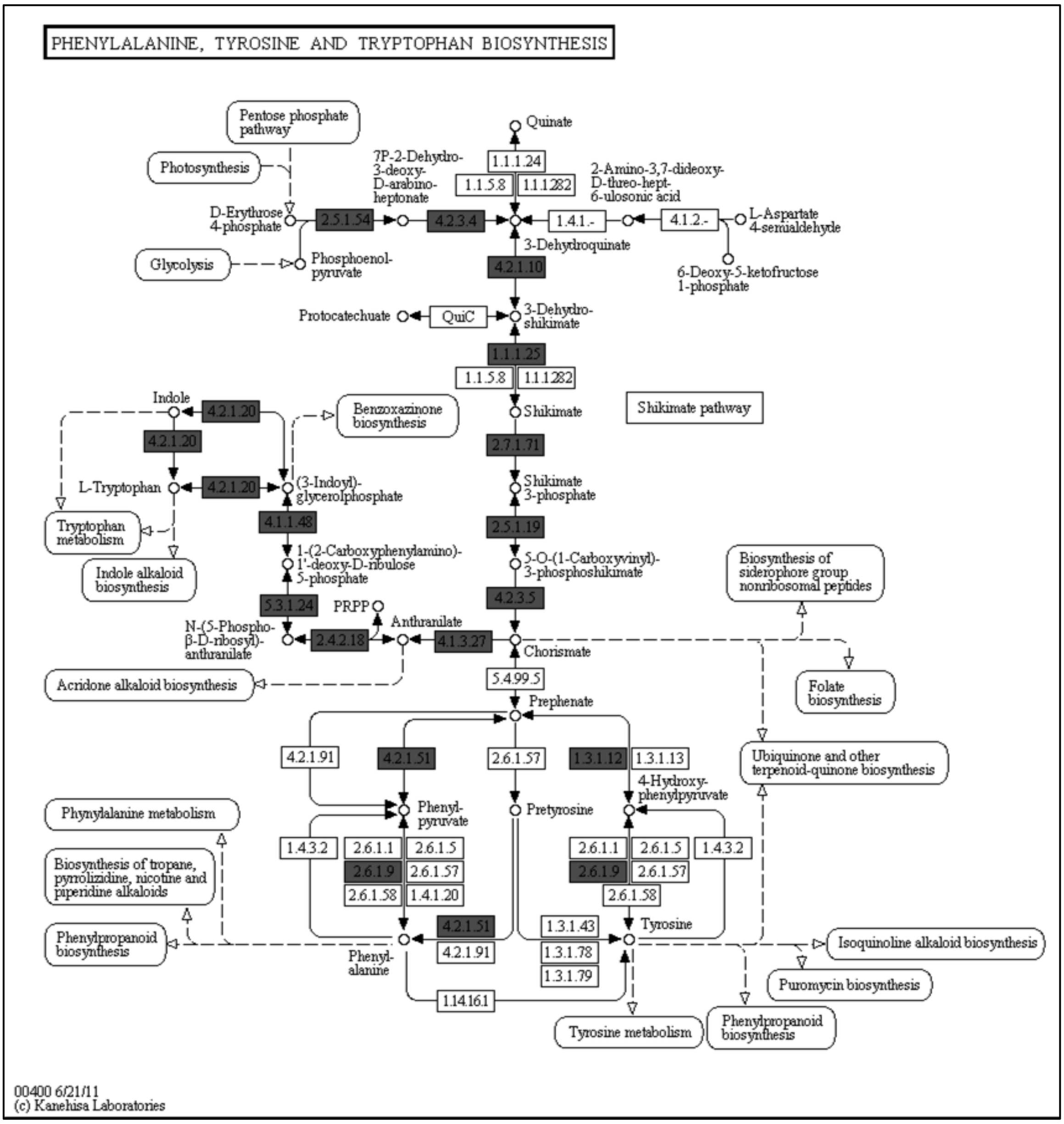


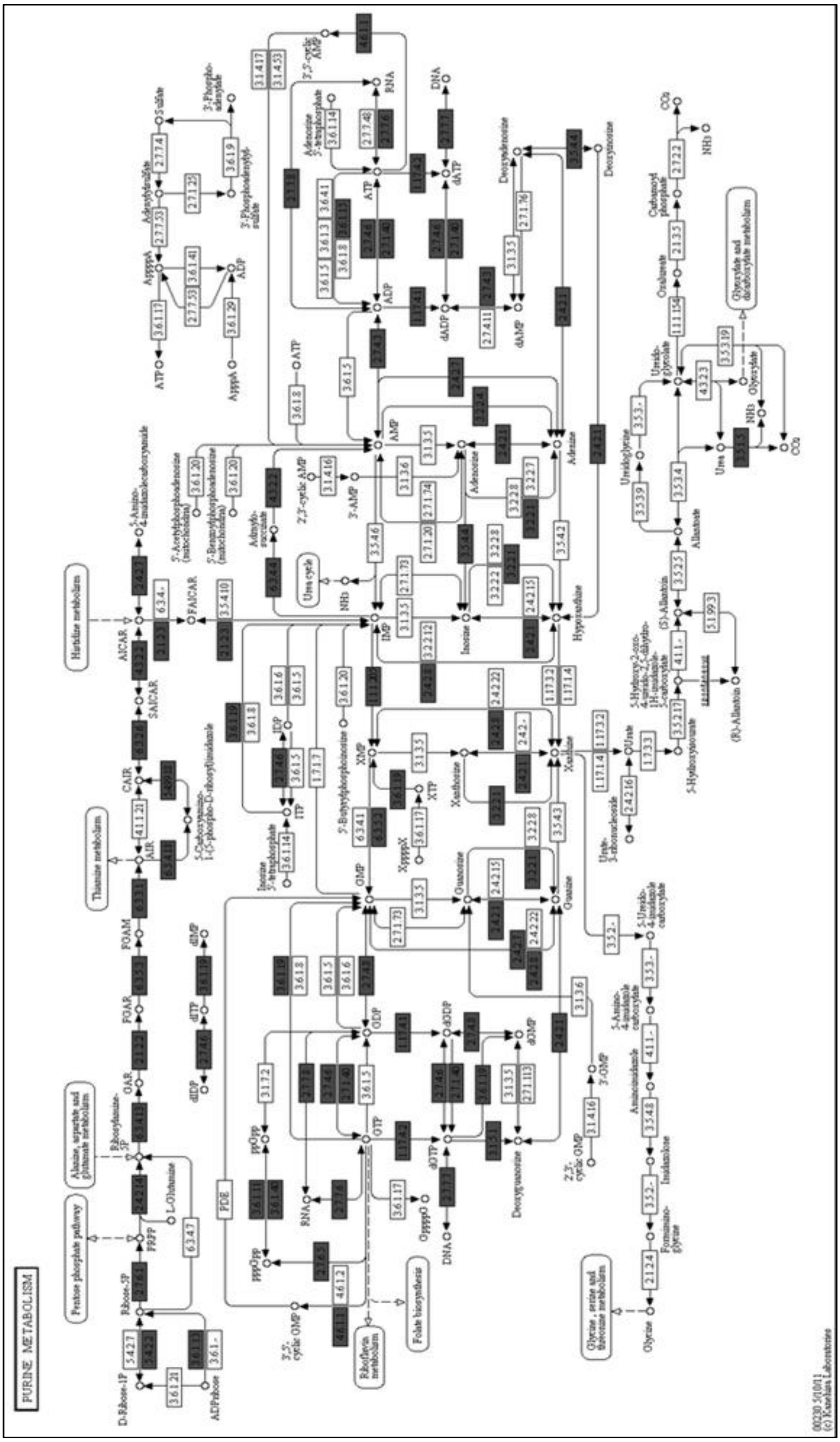

PENTOSE PHOSPHATE PATHWAY
Aminosugars metabolism
D-Glucosaminate
4.3.1.9
Pentose and glucuronate interconversions
D-Glyceraldehyde
1.2.7.5
Glycerate
2.7.1.-
Glycerate-2P
Glycolysis
4.1.2.-
Pyruvate
2.7.1.45
4.1.2.14
4.1.2.-
D-Glyceraldehyde-3P
2-Dehydro-3-deoxy-D-gluconate
4.2.1.39
Entner-Doudoroff pathway
2-Dehydro-3-deoxy-D-gluconate-6P
4.2.1.12
6-Phospho-D-gluconate
D-Gluconate
2.7.1.12
1.1.1.215
1.1.99.3
2.7.1.13
2-Dehydro-D-gluconate
1.1.1.43
2-Dehydro-D-gluconate-6P
H6PD
3.1.1.17
D-Glucono-1,5-lactone
1.1.1.47
1.1.3.4
1.1.3.5
1.1.99.10
1.1.5.2
β-D-Glucose
D-Glucose
3.1.1.31
D-Glucono-1,5-lactone-6P
1.1.1.49
β-D-Glucose-6P
5.3.1.9
α-D-Glucose-6P
5.3.1.9
1.1.1.44
D-Ribulose-5P
5.3.1.6
D-Ribose-5P
2.7.1.15
5.4.2.7
5.4.2.2
D-Ribose
D-Ribose-1P
D-Ribose-1,5P
2.7.4.23
2.7.6.1
PRPP
Purine metabolism
Pyrimidine metabolism
Histidine metabolism
5.1.3.1
Pentose and glucuronate interconversions
D-Xylulose-5P
2.2.1.1
2.2.1.2
D-Erythrose-4P
2.2.1.1
D-Sedo-heptulose-7P
2-Deoxy-D-ribose
2.7.1.15
5.4.2.7
2-Deoxy-D-ribose-1P
D-Xylulose-5P
2-Deoxy-D-ribose-5P
4.1.2.9
4.1.2.4
2.7.1.11
3.1.3.11
β-D-Fructose-6P
β-D-Fructose-1,6P2
4.1.2.13
D-Glyceraldehyde-3P
Glycolysis
00030 1/5/11
(c) Kanehisa Laboratories

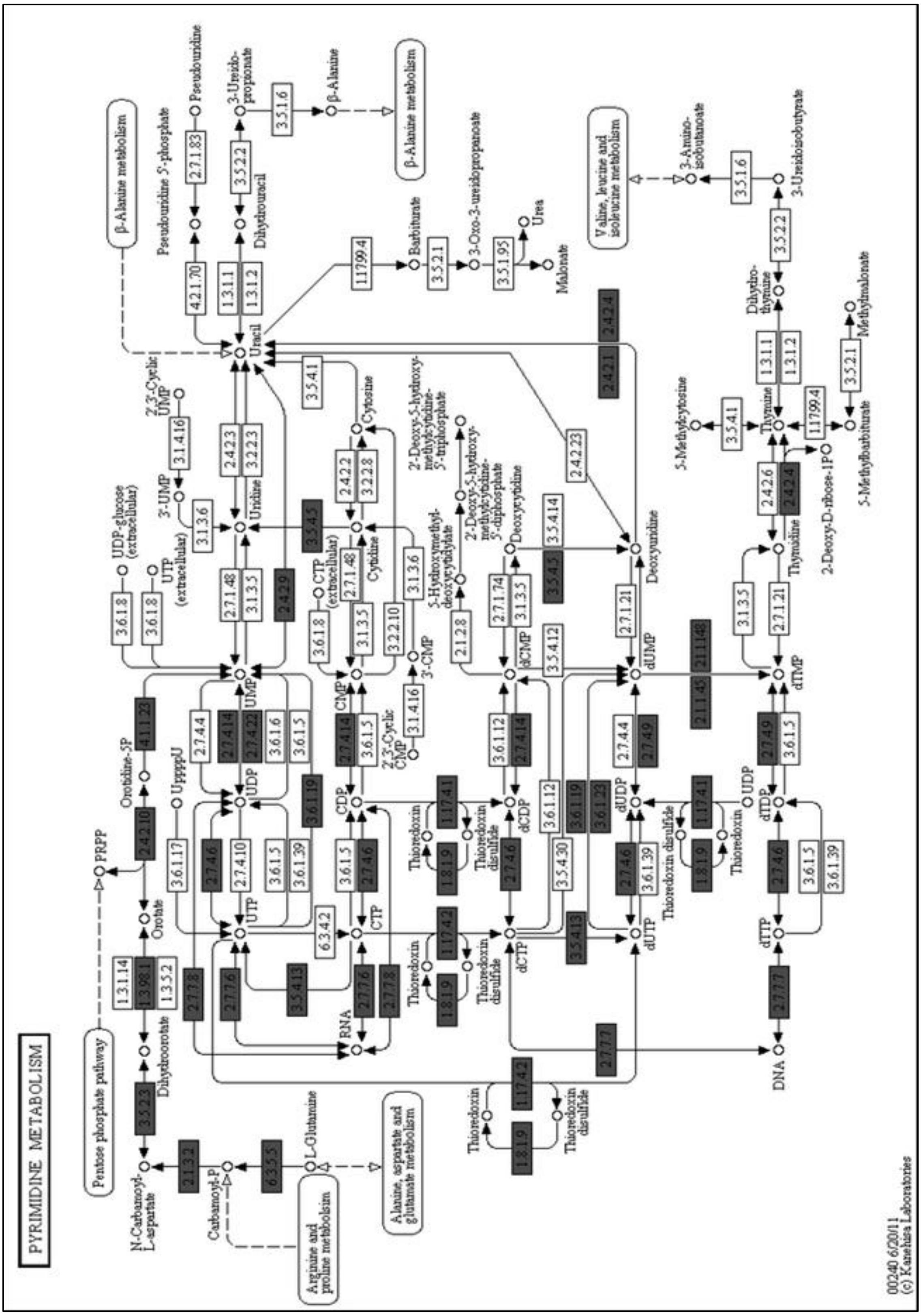

CITRATE CYCLE (TCA CYCLE)
Glycolysis / Gluconeogenesis
Phosphoenol-pyruvate
Pyruvate
4.1.1.32
4.1.1.49
1.2.7.1
6.4.1.1
1.2.4.1
ThPP
2-Hydroxy-ethyl-ThPP
Lipoamide-E
S-Acetyldihydro-lipoamide-E
1.8.1.4
2.3.1.12
Dihydro-lipoamide-E
Acetyl-CoA
Oxaloacetate
2.3.3.1
2.3.3.8
4.1.3.6
Citrate
4.2.1.3
cis-Aconitate
Isocitrate
1.1.1.42
Oxalosuccinate
1.1.1.41
2-Oxo-glutarate
1.2.4.2
3-Carboxy-1-hydroxypropyl-ThPP
S-Succinyl-dihydrolipoamide-E
2.3.1.61
1.2.7.3
Succinyl-CoA
6.2.1.4
6.2.1.5
Succinate
1.3.5.1
1.3.99.1
Fumarate
4.2.1.2
(S)-Malate
1.1.1.37
Fatty acid biosynthesis
Fatty acid elongation in mitochondria
Val, Leu & Ile degradation
Fatty acid metabolism
Alanine, aspartate and glutamate metabolism
Glyoxylate and dicarboxylate metabolism
Tyrosine metabolism
Arginine and proline metabolism
Val, Leu & Ile degradation
Ascorbate and aldarate metabolism
Alanine, aspartate and glutamate metabolism
D-Gln & D-Glu metabolism
00020 6/28/11
(c) Kanehisa Laboratories

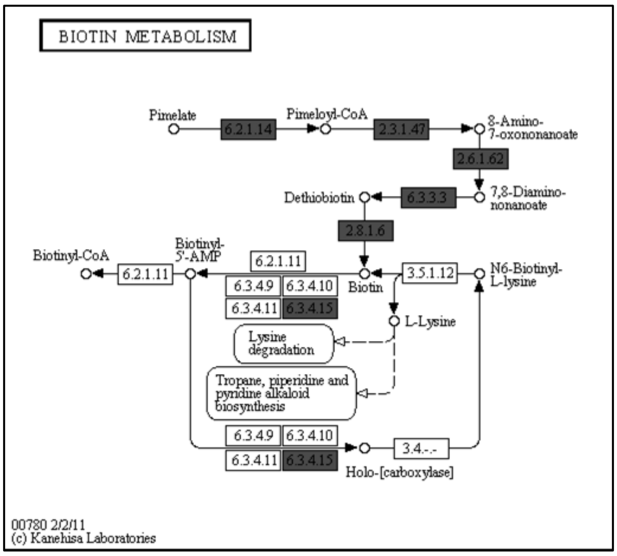

VALINE, LEUCINE AND ISOLEUCINE BIOSYNTHESIS
Glycine, serine and threonine metabolism
D-erythro-3-Methylmalate
LEUC/D
2-Methyl-maleate
4.2.1.35
(R)-2-Methylmalate
2.3.1.182
Acetyl-CoA
LEUB
Threonine
4.3.1.19
2-Oxobutanoate
2-Hydroxyethyl-ThPP
1.2.4.1
2.2.1.6
Pyruvate
2.2.1.6
2.2.1.6
(S)-2-Aceto-2-hydroxybutanoate
(S)-2-Acetolactate
1.1.1.86
5.4.99.3
1.1.1.86
5.4.99.3
(R)-3-Hydroxy-3-methyl-2-oxopentanoate
3-Hydroxy-3-methyl-2-oxobutanoate
Pyruvate metabolism
1.1.1.86
1.1.1.86
(R)-2,3-Dihydroxy-3-methylpentanoate
(R)-2,3-Dihydroxy-3-methylbutanoate
4.2.1.9
4.2.1.9
(S)-3-Methyl-2-oxopentanoate
2-Oxoisovalerate
2.3.3.13
(2S)-2-Isopropylmalate
4.2.1.33
2-Isopropylmaleate
4.2.1.33
(2R,3S)-3-Isopropylmalate
2.6.1.42
1.4.1.9
2.6.1.42
1.4.1.9
2.6.1.66
1.1.1.85
L-Isoleucine
L-Valine
(2S)-2-Isopropyl-3-oxosuccinate
spontaneous
4-Methyl-2-oxopentanoate
2.6.1.6
2.6.1.42
1.4.1.9
Valine, leucine and isoleucine degradation
L-Leucine
00290 7/28/11
(c) Kanehisa Laboratories


LYSINE BIOSYNTHESIS
Glycine, serine and threonine metabolism
L-Homoserine
1.1.1.3
N-Acetyl-L-2-amino-6-oxopimelate
2.3.1.89
2.6.1.-
N-Acetyl-LL-2,6-diaminopimelate
L-Aspartate
2.7.2.4
1.2.1.11
L-2,3-Dihydrodipicolinate
4.2.1.52
1.3.1.26
2.3.1.117
2.6.1.17
N-Succinyl-LL-2,6-diaminopimelate
L-4-Aspartyl phosphate
L-Aspartate 4-semialdehyde
L-2,3,4,5-Tetrahydro-dipicolinate
N-Succinyl-L-2-amino-6-oxopimelate
3.5.1.47
3.5.1.18
Alanine, aspartate and glutamate metabolism
2.6.1.83
LL-2,6-Diaminopimelate
5.1.1.7
UDP-N-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminopimeloyl-D-alanyl-D-alanine
L-2-Amino-6-oxopimelate
1.4.1.16
meso-2,6-Diamino-pimelate
Peptidoglycan biosynthesis
6.3.2.10
6.3.2.13
UDP-N-acetylmuramoyl-L-alanyl-D-γ-glutamyl-meso-2,6- diaminopimelate
4.1.1.20
Glycolysis
2.6.1.36
Homocitrate
Homo-cis-aconitate
5-Adenyl-2-aminoadipate
α-Aminoadipoyl-S-acyl enzyme
L-Saccharopine
1.5.1.7
L-Lysine
Tropane, piperidine and pyridine alkaloid biosynthesis
Acetyl-CoA
2.3.3.14
4.2.1.36
1.2.1.31
1.2.1.31
1.5.1.10
1.5.1.8
2-Oxo-glutarate
L-2-Aminoadipate 6-semialdehyde
4.2.1.36
1.2.1.31
LysK
Homo-isocitrate
L-2-Amino-adipate
LysW
LysX
LysZ
LysY
LysJ
Lysine degradation
N2-Acetyl-L-amino-adipate
N2-Acetyl-L-aminoadipyl-δ-phosphate
N2-Acetyl-L-aminoadipate semialdehyde
N2-Acetyl-L-lysine
Citrate cycle
Pyruvate metabolism
1.1.1.87
2.6.1.39
2.6.1.57
1.1.1.87
Oxaloglutarate
2-Oxoadipate
00300 1/14/11
(c) Kanehisa Laboratories

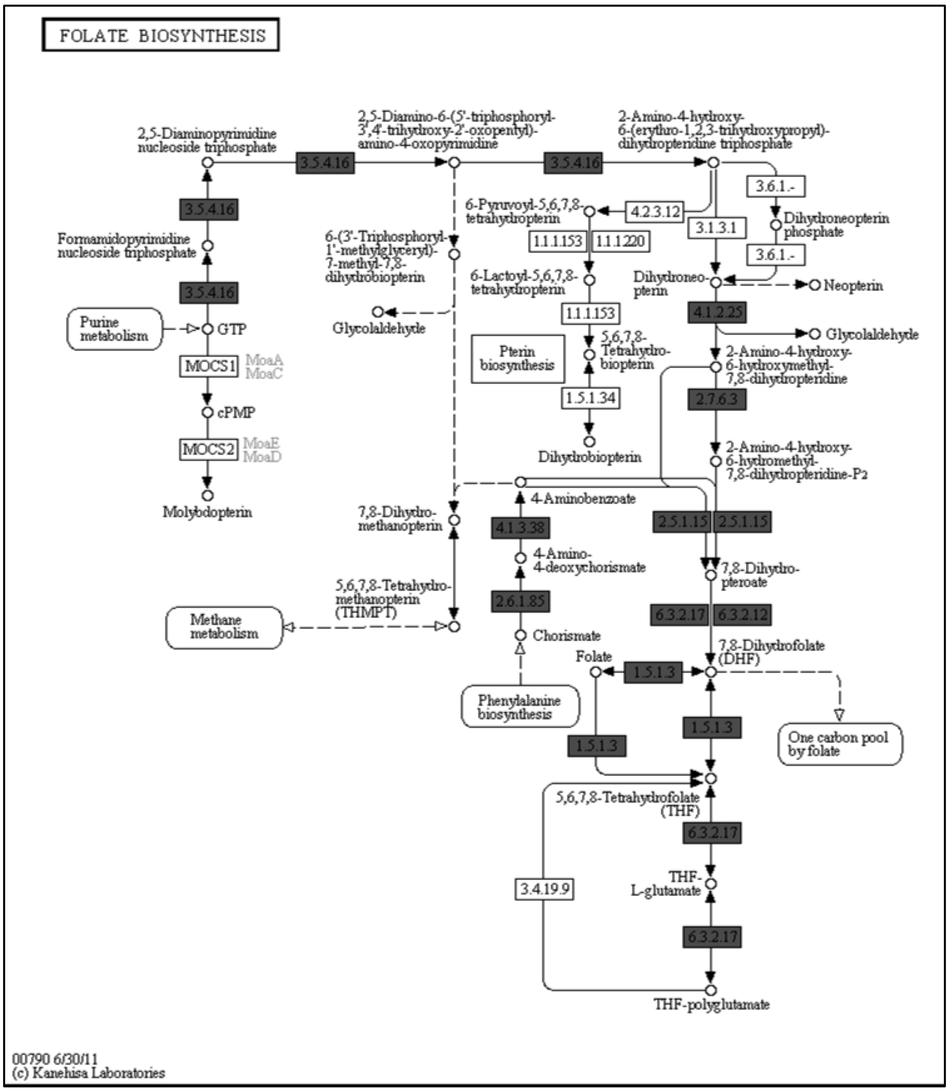

PHENYLALANINE, TYROSINE AND TRYPTOPHAN BIOSYNTHESIS
Pentose phosphate pathway
Photosynthesis
Glycolysis
D-Erythrose 4-phosphate
Phosphoenol-pyruvate
2.5.1.54
7P-2-Dehydro-3-deoxy-D-arabino-heptonate
4.2.3.4
Quinate
1.1.1.24
1.1.5.8
1.1.1282
2-Amino-3,7-dideoxy-D-threo-hept-6-ulosonic acid
1.4.1.-
4.1.2.-
L-Aspartate 4-semialdehyde
6-Deoxy-5-ketofructose 1-phosphate
3-Dehydroquinate
4.2.1.10
Protocatechuate
QuiC
3-Dehydro-shikimate
1.1.1.25
1.1.5.8
1.1.1282
Shikimate
Shikimate pathway
2.7.1.71
Shikimate 3-phosphate
2.5.1.19
5-O-(1-Carboxyvinyl)-3-phosphoshikimate
4.2.3.5
Chorismate
Biosynthesis of siderophore group nonribosomal peptides
Folate biosynthesis
Indole
4.2.1.20
4.2.1.20
L-Tryptophan
4.2.1.20
(3-Indoyl)-glycerolphosphate
Benzoxazinone biosynthesis
Tryptophan metabolism
Indole alkaloid biosynthesis
4.1.1.48
1-(2-Carboxyphenylamino)-1'-deoxy-D-ribulose 5-phosphate
5.3.1.24
PRPP
Anthranilate
N-(5-Phospho-β-D-ribosyl)-anthranilate
2.4.2.18
4.1.3.27
Acridone alkaloid biosynthesis
5.4.99.5
Prephenate
4.2.1.91
4.2.1.51
2.6.1.57
1.3.1.12
1.3.1.13
4-Hydroxy-phenylpyruvate
Ubiquinone and other terpenoid-quinone biosynthesis
Phynylalanine metabolism
Phenyl-pyruvate
Pretyrosine
1.4.3.2
2.6.1.1
2.6.1.5
2.6.1.9
2.6.1.57
2.6.1.58
1.4.1.20
2.6.1.1
2.6.1.5
1.4.3.2
2.6.1.9
2.6.1.57
2.6.1.58
Biosynthesis of tropane, pyrrolizidine, nicotine and piperidine alkaloids
Phenylpropanoid biosynthesis
4.2.1.51
4.2.1.91
Phenyl-alanine
1.3.1.43
1.3.1.78
1.3.1.79
Tyrosine
Isoquinoline alkaloid biosynthesis
Puromycin biosynthesis
1.14.16.1
Tyrosine metabolism
Phenylpropanoid biosynthesis
00400 6/21/11
(c) Kanehisa Laboratories


PURINE METABOLISM
Pentose phosphate pathway
Alanine, aspartate and glutamate metabolism
Thiamine metabolism
Histidine metabolism
Urea cycle
Glycine, serine and threonine metabolism
Glyoxylate and dicarboxylate metabolism
Riboflavin metabolism
Folate biosynthesis
D-Ribose-1P
Ribose-5P
ADPribose
PRPP
L-Glutamine
Ribosylamine-5P
GAR
FGAR
FGAM
AIR
CAIR
SAICAR
AICAR
FAICAR
5-Amino-4-imidazolecarboxyamide
5-Carboxyamino-1-(5-phospho-D-ribosyl)imidazole
ATP
ADP
AMP
ApppA
AppppA
Adenylylsulfate
3'-Phosphoadenylyl-sulfate
3'-Phosphoadenylate
Sulfate
3',5'-cyclic AMP
2',3'-cyclic AMP
3'-AMP
Adenosine
Adenosine 5'-tetraphosphate
Adenylo-succinate
RNA
DNA
dATP
dADP
dAMP
Deoxyadenosine
Deoxyinosine
Adenine
Hypoxanthine
Inosine
IMP
ITP
IDP
Inosine 5'-tetraphosphate
dITP
dIDP
dIMP
XMP
XTP
XppppX
Xanthosine
Xanthine
5'-Benzylphosphoinosine
GMP
GDP
GTP
dGTP
dGDP
dGMP
Guanosine
Guanine
Deoxyguanosine
GppppG
pppGpp
ppGpp
3',5'-cyclic GMP
2',3'-cyclic GMP
3'-GMP
PDE
NH3
Urate
Urate-3-ribonucleoside
5-Hydroxyisourate
5-Hydroxy-2-oxo-4-ureido-2,5-dihydro-1H-imidazole-5-carboxylate
(S)-Allantoin
(R)-Allantoin
Allantoate
Ureidoglycine
Ureido-glycolate
Glyoxylate
Urea
CO2
Oxaluric acid
Oxalurate
Carbamoyl phosphate
5-Ureido-4-imidazole carboxylate
5-Amino-4-imidazole carboxylate
Aminoimidazole
Imidazolone
Formimino-glycine
Glycine
00230 5/10/11
(c) Kanehisa Laboratories

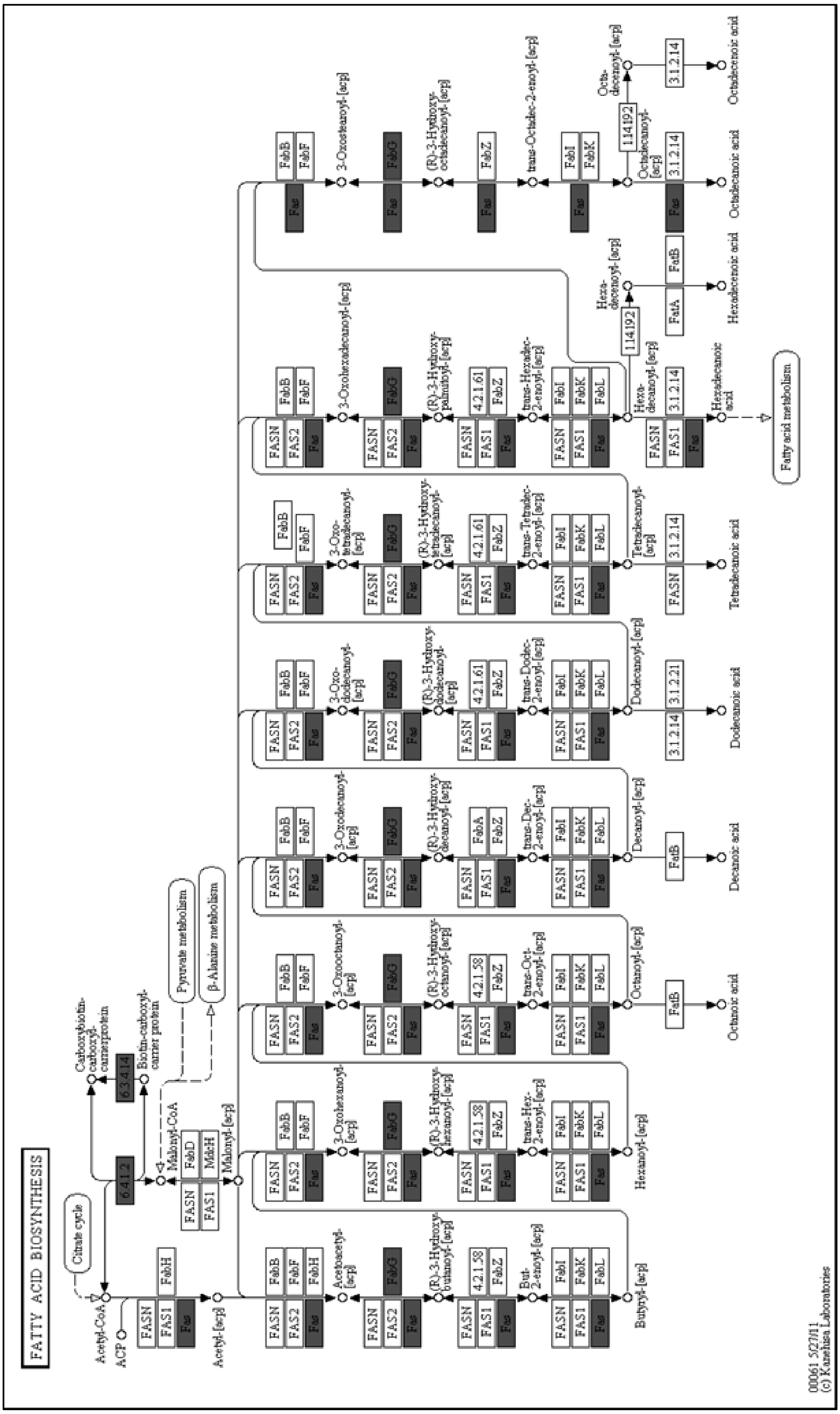

PYRIMIDINE METABOLISM
Pentose phosphate pathway
PRPP
N-Carbamoyl-L-aspartate
Dihydroorotate
Orotate
Orotidine-5P
Carbamoyl-P
L-Glutamine
Arginine and proline metabolism
Alanine, aspartate and glutamate metabolism
β-Alanine metabolism
UDP-glucose (extracellular)
UTP (extracellular)
CTP (extracellular)
UpppU
UMP
UDP
UTP
CMP
CDP
CTP
RNA
DNA
Uridine
Uracil
Cytidine
Cytosine
3'-UMP
2',3'-Cyclic UMP
3'-CMP
2',3'-Cyclic CMP
Pseudouridine
Pseudouridine 5'-phosphate
Dihydrouracil
3-Ureido-propionate
β-Alanine
Barbiturate
3-Oxo-3-ureidopropanoate
Urea
Malonate
dCMP
dCDP
dCTP
dUMP
dUDP
dUTP
dTMP
dTDP
dTTP
Deoxycytidine
Deoxyuridine
Thymidine
Thymine
Dihydro-thymine
3-Ureidoisobutyrate
3-Amino-isobutanoate
Valine, leucine and isoleucine metabolism
5-Methylcytosine
5-Methylbarbiturate
Methylmalonate
2-Deoxy-D-ribose-1P
5-Hydroxymethyl-deoxycytidylate
2'-Deoxy-5-hydroxy-methylcytidine-5'-diphosphate
2'-Deoxy-5-hydroxy-methylcytidine-5'-triphosphate
Thioredoxin
Thioredoxin disulfide
6.3.5.5
2.1.3.2
3.5.2.3
1.3.1.14
1.3.98.1
1.3.5.2
2.4.2.10
4.1.1.23
3.6.1.17
3.6.1.8
2.7.4.4
2.7.4.14
2.7.4.22
3.6.1.6
3.6.1.5
2.7.4.6
2.7.4.10
3.6.1.39
3.6.1.19
6.3.4.2
3.5.4.13
2.7.7.6
2.7.7.8
2.7.1.48
3.1.3.5
2.4.2.9
2.4.2.3
3.2.2.3
3.1.3.6
3.1.4.16
4.2.1.70
2.7.1.83
1.3.1.1
1.3.1.2
3.5.2.2
3.5.1.6
1.17.99.4
3.5.2.1
3.5.1.95
3.5.4.1
2.4.2.2
3.2.2.8
3.5.4.5
3.2.2.10
2.1.2.8
2.7.1.74
3.5.4.14
3.5.4.12
3.6.1.12
3.5.4.30
3.6.1.23
1.17.4.1
1.17.4.2
1.8.1.9
2.7.1.21
2.1.1.45
2.1.1.148
2.4.2.23
2.4.2.1
2.4.2.4
2.7.4.9
2.7.7.7
3.1.3.5
2.4.2.6
00240 6/20/11
(c) Kanehisa Laboratories


BIOTIN METABOLISM

00780 2/2/11
(c) Kanehisa Laboratories


FOLATE BIOSYNTHESIS

00790 6/30/11
(c) Kanehisa Laboratories


FATTY ACID BIOSYNTHESIS
Citrate cycle
Acetyl-CoA
ACP
FASN
FAS1
Fas
FabH
Acetyl-[acp]
6.4.1.2
6.3.4.14
Carboxybiotin-carboxyl-carrier protein
Biotin-carboxyl-carrier protein
Malonyl-CoA
FabD
MdcH
Malonyl-[acp]
Pyruvate metabolism
β-Alanine metabolism
FabB
FabF
FAS2
FabG
4.2.1.58
4.2.1.61
FabA
FabZ
FabI
FabK
FabL
Acetoacetyl-[acp]
(R)-3-Hydroxybutanoyl-[acp]
But-2-enoyl-[acp]
Butyryl-[acp]
3-Oxohexanoyl-[acp]
(R)-3-Hydroxyhexanoyl-[acp]
trans-Hex-2-enoyl-[acp]
Hexanoyl-[acp]
3-Oxooctanoyl-[acp]
(R)-3-Hydroxyoctanoyl-[acp]
trans-Oct-2-enoyl-[acp]
Octanoyl-[acp]
FatB
Octanoic acid
3-Oxodecanoyl-[acp]
(R)-3-Hydroxydecanoyl-[acp]
trans-Dec-2-enoyl-[acp]
Decanoyl-[acp]
Decanoic acid
3-Oxo-dodecanoyl-[acp]
(R)-3-Hydroxydodecanoyl-[acp]
trans-Dodec-2-enoyl-[acp]
Dodecanoyl-[acp]
3.1.2.14
3.1.2.21
Dodecanoic acid
3-Oxo-tetradecanoyl-[acp]
(R)-3-Hydroxytetradecanoyl-[acp]
trans-Tetradec-2-enoyl-[acp]
Tetradecanoyl-[acp]
Tetradecanoic acid
3-Oxohexadecanoyl-[acp]
(R)-3-Hydroxypalmitoyl-[acp]
trans-Hexadec-2-enoyl-[acp]
Hexadecanoyl-[acp]
1.14.19.2
Hexadecenoyl-[acp]
FatA
Hexadecenoic acid
Hexadecanoic acid
Fatty acid metabolism
3-Oxostearoyl-[acp]
(R)-3-Hydroxyoctadecanoyl-[acp]
trans-Octadec-2-enoyl-[acp]
Octadecanoyl-[acp]
Octadecenoyl-[acp]
Octadecanoic acid
Octadecenoic acid
00061 5/27/11
(c) Kanehisa Laboratories

Appendix Three: *Cp* 3/99-5 proteins detected by LC-MS^2

CDS	Gene	Product
Cp3995_2123		Hypothetical protein
Cp3995_0012		Hypothetical protein
Cp3995_0034	*pbpA*	Penicillin-binding protein A
Cp3995_0082		Hypothetical protein
Cp3995_0129		Hypothetical protein
Cp3995_0196		Hypothetical protein
Cp3995_0202	*pbpB*	Penicillin binding protein transpeptidase
Cp3995_0240	*slpA*	Surface layer protein A
Cp3995_0392		L,D-transpeptidase catalytic domain, region YkuD
Cp3995_0458	*htaA*	Cell surface hemin receptor
Cp3995_0462	*htaC*	Hypothetical protein
Cp3995_0508	*lytR*	Transcriptional regulator lytR
Cp3995_0581		Hypothetical protein
Cp3995_0603	*rpfA*	Resuscitation-promoting factor
Cp3995_0625		Hypothetical protein
Cp3995_0653		Uncharacterized metalloprotease
Cp3995_0673	*pepD*	Serine protease
Cp3995_0918	*lpqC*	Poly(3-hydroxybutyrate) depolymerase
Cp3995_0920		Hypothetical protein
Cp3995_1191		Hypothetical protein
Cp3995_1510		Uncharacterized protein HtaC
Cp3995_1511		Cell-surface hemin receptor
Cp3995_1716	*ftn*	Ferritin-like protein
Cp3995_1810	*lpqE*	Lipoprotein LpqE
Cp3995_1848	*lipY*	Exported lipase
Cp3995_1890		Hypothetical protein
Cp3995_1946		Membrane protein
Cp3995_1947		Serine protease CP40
Cp3995_2010	*cmtC*	Trehalose corynomycolyl transferase C
Cp3995_2012	*cmtB*	Trehalose corynomycolyl transferase B
Cp3995_2029		Peptidoglycan recognition protein

Cp3995_2093	*rplI*	50S ribosomal protein L9
Cp3995_2120		Hypothetical protein
Cp3995_2135		Hypothetical protein
Cp3995_0026	*pld*	Phospholipase D
Cp3995_0067		Lysozyme M1
Cp3995_0174		Surface antigen
Cp3995_0234		Hydrolase domain-containing protein
Cp3995_0321		Hypothetical protein
Cp3995_0322		Hypothetical protein
Cp3995_0338	*tuf*	Elongation factor Tu
Cp3995_0372		PAP2 superfamily protein
Cp3995_0379		Hypothetical protein
Cp3995_0391	*nanH*	Neuraminidase (sialidase)
Cp3995_0446		Hypothetical protein
Cp3995_0459	*hmuT*	Hemin-binding periplasmic protein hmuT
Cp3995_0543		Hydrolase domain-containing protein
Cp3995_0544		Hypothetical protein
Cp3995_0590	*cynT*	Carbonic anhydrase
Cp3995_0591	*fecB*	Periplasmic binding protein
Cp3995_0634	*oppA2*	Oligopeptide-binding protein oppA
Cp3995_0692	*rpfB*	Resuscitation-promoting factor RpfB
Cp3995_0725	*eno*	Enolase
Cp3995_0775	*sprX*	Trypsin-like serine protease
Cp3995_0778		Hypothetical protein
Cp3995_0898		Hypothetical protein
Cp3995_1002	*ciuA*	Iron ABC transporter substrate-binding protein
Cp3995_1099		Invasion-associated protein p60
Cp3995_1250	*ahpD*	Alkyl hydroperoxide reductase AhpD
Cp3995_1401		Hypothetical protein
Cp3995_1457		Cell-wall peptidase. NlpC/P60 protein
Cp3995_1465	*ctaC*	Cytochrome c oxidase subunit II
Cp3995_1733	*dsbG*	DsbG protein
Cp3995_1793		Corynomycolyl transferase
Cp3995_1812		Hypothetical protein
Cp3995_1851		Hypothetical protein

Cp3995_1853	*porH*	Cell wall channel
Cp3995_1854		Cell wall channel
Cp3995_1891		Hypothetical protein
Cp3995_1892		Hypothetical protein
Cp3995_1902		Hypothetical protein
Cp3995_1930		Hypothetical protein
Cp3995_1940		Secretion protein HlyD
Cp3995_2042		Hypothetical protein
Cp3995_2043		Hypothetical protein

Complete Genome Sequences of *Corynebacterium pseudotuberculosis* Strains 3/99-5 and 42/02-A, Isolated from Sheep in Scotland and Australia, Respectively

Florence E. Pethick,[a,b] Alex F. Lainson,[a] Raja Yaga,[a] Allen Flockhart,[a] David G. E. Smith,[a,b] Willie Donachie,[a] Louise T. Cerdeira,[c] Artur Silva,[c] Erik Bol,[c] Thiago S. Lopes,[c] Maria S. Barbosa,[c] Anne C. Pinto,[d] Anderson R. dos Santos,[d] Siomar C. Soares,[d] Sintia S. Almeida,[d] Luis C. Guimaraes,[d] Flavia F. Aburjaile,[d] Vinicius A. C. Abreu,[d] Dayana Ribeiro,[d] Karina K. Fiaux,[d] Carlos A. A. Diniz,[d] Eudes G. V. Barbosa,[d] Ulisses P. Pereira,[e] Syed S. Hassan,[d] Amjad Ali,[d] Syeda M. Bakhtiar,[d] Fernanda A. Dorella,[d] Adriana R. Carneiro,[c] Rommel T. J. Ramos,[c] Flavia S. Rocha,[d] Maria P. C. Schneider,[c] Anderson Miyoshi,[d] Vasco Azevedo,[d] and Michael C. Fontaine[a]

Moredun Research Institute, Pentlands Science Park, Bush Loan, Edinburgh, Midlothian, United Kingdom[a]; Institute of Infection, Immunity and Inflammation. College of Medical, Veterinary and Life Sciences, University of Glasgow Garscube Estate, Glasgow, United Kingdom[b]; Instituto de Ciências Biológicas, Univeridade Federal do Pará, Belém, Brazil[c]; Instituto de Ciências Biológicas, Univeridade Federal de Minas Gerais, Belo Horizonte, Brazil[d]; and Departmento de Medicina Veterinária, Universidade Federal de Lavras, Lavras, Brazil[e]

Here, we report the whole-genome sequences of two ovine-pathogenic *Corynebacterium pseudotuberculosis* isolates: strain 3/99-5, which represents the first *C. pseudotuberculosis* genome originating from the United Kingdom, and 42/02-A, the second from Australia. These genome sequences will contribute to the objective of determining the global pan-genome of this bacterium.

Corynebacterium pseudotuberculosis is responsible for several diseases in multiple host species, the most notable being caseous lymphadenitis (CLA), a chronic pyogenic disease of small ruminants (1). Previous work has revealed that ovine *C. pseudotuberculosis* strains represent a largely clonal population but that these are distinct from other strains associated with equine disease (2). To identify the core genome and other variable accessory genes contributing to host specificity, work is under way to sequence representative *C. pseudotuberculosis* strains from multiple geographic locations and host species. To facilitate this study, we have sequenced the complete genomes of two ovine *C. pseudotuberculosis* strains, isolated from natural outbreaks of CLA: 42/02-A, an isolate from Perth, Australia, and 3/99-5, from the Scottish Borders, United Kingdom.

The *C. pseudotuberculosis* 3/99-5 and 42/02-A genomes were sequenced using 454 GS-FLX and Solexa 50-bp paired-end sequencing. Reads were assembled using Velvet (8) and CABOG (Celera Assembler with the Best Overlap Graph) (5), and gaps were closed using unmapped 454 and Illumina reads. Structural annotation was achieved using the following software: FgenesB (a gene predictor), RNAmmer (an rRNA predictor) (3), tRNA-scan-SE (a tRNA predictor) (4), and Tandem Repeat Finder (to predict repeatDNAregions) (http://tandem.bu.edu/trf/trf.html). Functional annotation was performed using InterProScan (7) analysis and homology analyses using public databases. Manual annotation was then completed using Artemis software (6).

The presence of pseudogenes within the genomes was determined using CLCBio Workbench 4.02 software. Manual analysis was also conducted based on Phred quality of each base and analysis of coverage depth at the frameshift region, allowing identification of false-positive pseudogene results.

The *C. pseudotuberculosis* 3/99-5 genome consists of a single 2,337,938-bp circular chromosome with an average G_C content of 52.18%. The genome was predicted to contain 2,142 coding sequences (CDS), four rRNA operons, and 49 tRNAs. In addition, 36 pseudogenes were also identified.

The highly similar *C. pseudotuberculosis* 42/02-A genome consists of a single 2,337,606-bp circular chromosome with an average G_C content of 52.19%. This genome was predicted to contain 2,051 coding sequences (CDS), four rRNA operons, and 49 tRNAs. In addition, 52 pseudogenes were also identified.

The sequencing of these isolates will aid comparison of genomes deriving from multiple geographical locations and host species in a wider pan-genome project. Widespread comparisons should offer insights into the organism's pathogenicity and host specificity as well as evolutionary relationships between strains originating from different geographical locations.

Nucleotide sequence accession numbers. The *C. pseudotuberculosis* 3/99-5 and 42/02-A genome sequences described in this study have been deposited in the GenBank database under accession numbers CP003152.1 and CP003062, respectively. The *C. pseudotuberculosis* 3/99-5 genome has also been deposited in the RefSeq database under accession number NC_016781.1.

ACKNOWLEDGMENTS

We acknowledge Quality Meat Scotland (QMS) and the Rural and Environment Science and Analytical Services Division (RESAS) of the Scottish Government for funding this work. This work was part of the Rede Paraense de Genômica e Proteômica, supported by FAPESPA, Núcleo Amazônico de Excelência em Genômica de Microorganismos—Pronex/ CNPq/FAPESPA, CAPES, and FAPEMIG.

REFERENCES

1. **Baird GJ, Fontaine MC.** 2007. *Corynebacterium pseudotuberculosis* and its role in ovine caseous lymphadenitis. J. Comp. Pathol. **137**:179 –210.

Received 25 May 2012 Accepted 5 June 2012
Address correspondence to Michael C. Fontaine, michael.fontaine@moredun.ac.uk.

doi:10.1128/JB.00918-12

2. **Conner KM, Fontaine MC, Rudge K, Baird GJ, Donachie W.** 2007. Molecular genotyping of multinational ovine and caprine *Corynebacterium pseudotuberculosis* isolates using pulsed-field gel electrophoresis. Vet. Res. **38**:613– 623.
3. **Lagesen K, et al.** 2007. RNAmmer:consistent annotation of rRNA genes in genomic sequences. Nucleic Acids Res. **35**:3100 –3108.
4. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**: 955–964.
5. **Miller JR, et al.** 2008. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics **24**:2818 –2824
6. **Rutherford K, et al.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16**:944 –945.
7. **Zdobnov EM, Apweiler R.** 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17**:847– 848.
8. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**:821– 829.

Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Strain 1/06-A, Isolated from a Horse in North America

Florence E. Pethick,a,b Alex F. Lainson,a Raja Yaga,a Allen Flockhart,a David G. E. Smith,a,b Willie Donachie,a Louise T. Cerdeira,c Artur Silva,c Erik Bol,c Thiago S. Lopes,c Maria S. Barbosa,c Anne C. Pinto,d Anderson R. dos Santos,d Siomar C. Soares,d Sintia S. Almeida,d Luis C. Guimaraes,d Flavia F. Aburjaile,d Vinicius A. C. Abreu,d Dayana Ribeiro,d Karina K. Fiaux,d Carlos A. A. Diniz,d Eudes G. V. Barbosa,d Ulisses P. Pereira,e Syed S. Hassan,d Amjad Ali,d Syeda M. Bakhtiar,d Fernanda A. Dorella,d Adriana R. Carneiro,c Rommel T. J. Ramos,c Flavia S. Rocha,d Maria P. C. Schneider,c Anderson Miyoshi,d Vasco Azevedo,d and Michael C. Fontainea

Moredun Research Institute, Pentlands Science Park, Bush Loan, Edinburgh, Midlothian, United Kingdoma; Institute of Infection, Immunity and Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Garscube Estate, Glasgow, United Kingdomb; Instituto de Ciências Biológicas, Univeridade Federal do Pará, Belém, Brazilc; Instituto de Ciências Biológicas, Univeridade Federal de Minas Gerais, Belo Horizonte, Brazild; and Departmento de Medicina Veterinária, Universidade Federal de Lavras, Lavras, Brazile

Corynebacterium pseudotuberculosis **causes disease in several animal species, although distinct biovars exist that appear to be restricted to specific hosts. In order to facilitate a better understanding of the differences between biovars, we report here the complete genome sequence of the equine pathogen *Corynebacterium pseudotuberculosis* strain 1/06-A.**

Corynebacterium pseudotuberculosis is a cause of disease in several host species. Two distinct biovars have been described (1); *C. pseudotuberculosis* bv. ovis, which predominantly affects small ruminant species (primarily sheep and goats), and *C. pseudotuberculosis* bv. equi, which predominantly affects horses. Significantly, strains belonging to *C. pseudotuberculosis* bv. equi are unable to cause infection in small ruminants. Previous studies have revealed that *C. pseudotuberculosis* represents a largely clonal population (2). Clearly, however, each biovar has adapted to a particular host, although the nature of these adaptations is currently unclear.

Most previously sequenced strains of *C. pseudotuberculosis* belong to biovar ovis, and here we report the genome sequence of *C. pseudotuberculosis* 1/06-A, an equine isolate originating from North America.

The *C. pseudotuberculosis* 1/06-A genome was sequenced using 454 GS-FLX and Solexa 50-bp paired-end sequencing. Reads were assembled using Velvet (8) and CABOG (Celera assembler with the best overlap graph) (5), and gaps were closed using unmapped 454 and Illumina reads.

Structural annotation was achieved using the following software: FgenesB (a gene predictor); RNAmmer (an rRNA predictor) (3); tRNA-scan-SE (a tRNA predictor) (4); Tandem Repeat Finder (to predict repeat DNA regions; http://tandem.bu.edu/trf/trf.html). Functional annotation was performed using InterProScan (7) analysis and homology analyses using public databases. Manual annotation was then completed using Artemis software (6).

The presence of pseudogenes within the genome was determined using CLCBio Workbench 4.02 software. Manual analysis was also conducted based on the Phred quality of each base, and with analysis of coverage depth at the frameshift region allowed identification of false-positive pseudogene results.

The *C. pseudotuberculosis* 1/06-A genome consists of a single 2,279,118-bp circular chromosome with an average G_C content of 52.20%. The genome was predicted to contain 1,963 coding sequences (CDSs), four rRNA operons, and 49 tRNAs. In addition, 103 pseudogenes were identified.

The sequencing of this isolate will allow comparison of genomes derived from the two distinct biovars and should offer insights into the organism's host specificity.

Nucleotide sequence accession number. The genome sequence described in this study has been deposited in the GenBank database under the accession number CP003082.

ACKNOWLEDGMENTS

We acknowledge Richard Walker for providing us with the isolate. This work was funded by the Rural and Environment Science and Analytical Services Division of the Scottish Government. This work was part of the Rede Paraense de Genômica e Proteômica, supported by FAPESPA, Núcleo Amazônico de Excelência em Genômica de Microorganismos— Pronex/CNPq/FAPESPA, CAPES, and FAPEMIG.

REFERENCES

1. **Baird GJ, Fontaine MC.** 2007. *Corynebacterium pseudotuberculosis* and its role in ovine caseous lymphadenitis. J. Comp. Pathol. **137**:179 –210.

2. **Connor KM, Fontaine MC, Rudge K, Baird GJ, Donachie W.** 2007. Molecular genotyping of multinational ovine and caprine *Corynebacterium pseudotuberculosis* isolates using pulsed-field gel electrophoresis. Vet. Res. **38**:613– 623.

3. **Lagesen K, et al.** 2007. RNAmmer: consistent annotation of rRNA genes in genomic sequences. Nucleic Acids Res. **35**:3100 –3108.

4. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**: 955–964.

5. **Miller JR, et al.** 2008. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics **24**:2818 –2824.

6. **Rutherford K, et al.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16**:944 –945.

7. **Zdobnov EM, Apweiler R.** 2001. InterProScan: an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17**:847–848.

8. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**:821– 829.

Received 25 May 2012 Accepted 5 June 2012
Address correspondence to Michael C. Fontaine, michael.fontaine@moredun.ac.uk.

doi:10.1128/JB.00922-12