Genetic Characterisation of *Streptococcus pneumoniae* Serotype 1 Isolates in Relation to Invasiveness

Richard Manuel Harvey, B.Sc. (Biomedical Science) (Hons), AMusA

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy from the University of Adelaide

August 2010

Discipline of Microbiology and Immunology
School of Molecular and Biomedical Sciences
The University of Adelaide
Adelaide, S.A., Australia

Abstract

Streptococcus pneumoniae (the pneumococcus) is one of the most significant causes of human mortality and morbidity, and is a leading cause of diseases such as pneumonia, invasive disease (including bacteraemia and meningitis [IPD]) and otitis media. However, the pneumococcus is more commonly carried asymptomatically within the nasopharynx. The likelihood of the pneumococcus progressing from asymptomatic carriage to IPD varies between strains, and is associated with certain serotypes and clones. In particular, serotype 1 strains have a high-attack rate as they readily progress from a state of transient carriage to IPD. Recently, a closely-related group of hypervirulent serotype 1 clones have been responsible for epidemics of IPD with unusually high mortality rates. In contrast, epidemic asymptomatic carriage of serotype 1 clones has been found in a number of remote indigenous communities in the Northern Territory of Australia. Such isolates of serotype 1 from asymptomatic carriage are unusual and provided a rare opportunity to perform genomic comparisons with invasive serotype 1 isolates in order to identify serotype-independent factors that contribute to differences in the invasive potential of the pneumococcus.

Preliminary work using the non-invasive serotype 1 isolates from the Northern Territory and a collection of invasive human isolates of both indigenous and non-indigenous origin identified three virulence profiles that were non-invasive, intermediately virulent, or highly virulent in mice. Subsequently, phenomic analyses did not identify differences in the amount of capsule or differences in the apparent molecular weight or relative expression of a selection of well-characterised protein virulence factors that correlated with a virulence phenotype. However, in preliminary genomic comparisons the chromosomal toxin-antitoxin (TA) system of the PPI-1 variable region (PezAT) was identified in only highly virulent serotype 1 isolates, but absent from intermediately virulent and non-invasive serotype 1 isolates.

Therefore, the broad objectives of this study where to determine the clonal relatedness of isolates representing all three virulence phenotypes, characterise the potential role of the PPI-1 variable region in IPD and identify additional variable regions of the pneumococcal genome that were associated with heightened virulence.

Interestingly, it was shown that the highly virulent strain 1861 was a one-locus variant of the sequence type 217 clone of lineage B, responsible for severe IPD in parts of Africa. Therefore, the highly virulent nature of strain 1861 (and strain 4496) in mice is likely to also be reflected in humans. In contrast, the non-invasive and intermediately virulent strains were of lineage A, which includes the most frequently detected clones in Europe and the United States. In addition, different organisations of the PPI-1 variable region correlated with certain lineages of serotype 1. For example, the lineage A isolates lacked *pezAT* and instead contained a transcriptionally active immunity system against the bacteriocin, mersacidin. Interestingly, following a survey of a variety of *S. pneumoniae* strains representing a broad array of serotypes, the mersacidin immunity system was identified as the most common feature of the PPI-1 variable region, and is also present in the pandemic carriage Spanish^{23F} ST81 clone. In contrast, the highly virulent isolates of lineages B and C encoded *pezAT* and a number of genes predicted to encode enzymes that catalyse the rate-limiting steps of pathways involved in the degradation and biosynthesis of some amino acids and the biosynthesis and conversion of UDP-sugars. Interestingly, key components of this region exhibited preferential expression in the lungs and blood when compared to the nasopharynx of infected mice. Subsequently, it was shown using replacement mutants of the PPI-1 variable region in a D39 background that the region from the highly virulent strains promotes greater competitive fitness within the blood, lungs and nasopharyngeal tissue, compared to the equivalent region from the intermediately virulent and non-invasive strains in co-infected mice. Whilst the mechanism by which the PPI-1 variable region contributes to survival *in vivo* is not clear, a possibility is that centralised regulation of a number of metabolic pathways may enhance the survival of the pneumococcus in the lungs and blood.

Whilst the PPI-1 variable region was important for the competitive fitness of D39 during disease, it was not clear whether this region was solely responsible for the differences observed in invasive potential between the highly virulent, intermediately virulent and non-invasive serotype 1 isolates. Therefore, comparative genomic hybridisation (CGH) and next generation genome sequencing were used to identify additional regions of the genome that are associated with the highly virulent isolates. It was found that genes homologous to the platelet-binding protein B (PblB) and a *Streptococcus mitis* lysogenic phage endolysin were present in the genome of only the highly virulent strains, and not in either the intermediately virulent and non-invasive strains. In addition, regions encoding a putative ABC transporter and enzymes predicted to be involved in the degradation of sialic acid, ZmpD, and a 64-kb Tn*5253*-like conjugative transposon that included a TA system that is highly homologous to *pezAT*, were found in only the highly virulent strains and not in the intermediately virulent or non-invasive isolates. Subsequent *in vivo* gene expression comparisons revealed that the phage-associated endolysin exhibited significantly greater expression in the lungs and blood of infected mice than the nasopharynx, which highlighted a potential mechanism for increased surface display of PblB in the lungs and blood. Whilst yet to be proven experimentally, it is thought that greater surface display of PblB could contribute to the rapid invasion of the blood that is characteristic of the highly virulent serotype 1 strains. In addition to PblB, greater expression of the sialic acid-associated ABC transporter was observed in the blood when compared to the lungs and nasopharynx of infected mice.

Therefore, whilst the role of the region remains to be determined, it might be possible that the region enables the utilisation of host-derived sialic acids as an energy source in the blood, thus promoting survival and growth.

However, a significant roadblock encountered in this study was the inability to genetically manipulate the highly virulent serotype 1 isolates. In order to confirm the importance of genes such as that in the PPI-1 variable region and *pblB* in virulence, mutagenesis of these regions was attempted. However, despite numerous attempts to optimise the transformation protocol, it is possible that some defect in the competence system that is linked to the over-expression of *comW* might be responsible for the inability to transform strains 1861 and 4496.

In this study a number of genomic regions were identified that via putative roles in metabolism, sugar acquisition and degradation and adherence to human platelets and their patterns of expression *in vivo* promote the invasion and survival of the pneumococcus in the blood and lungs. Such findings broaden the understanding of the progression to IPD from asymptomatic carriage and highlight strain-specific differences that could make some strains more virulent than others.

Declaration

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web via the University's digital research repository, the Library catalogue, the Australasian Digital Theses Program (ADTP) and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

………………………...

Richard Manuel Harvey

Acknowledgements

Firstly I would like to thank my principal supervisor, Professor James Paton, for the time and effort that he has committed and the opportunities that he has provided to me over the time that I have been a student in his lab. Thank you also to my Co-supervisor, Dr. Uwe Stroeher, for his commitment both when he was in the lab and since he left. I am sure that both James and Uwe's guidance will have a lasting impact on my career. Thank you to Dr. David Ogyunniyi and Dr. Judy Morona for their ideas and help with various aspects of my project. Thank you to Dr. Lauren McAllister for helping me brainstorm ideas and for great company throughout the time of my project. I would also like to acknowledge other past and present members of the Paton lab, such as Dr. Adrienne Paton, Dr. Tony Focareta, Dr. Adam Potter, Dr. Claudia Trappetti, Dr. Sylvia Herold, James Byrne and Dr. Chris McDevitt for their assistance, which has made an important contribution to my project. Thank you also to Jan Cook and more recently Stephanie Philp for their effort in keeping the lab running smoothly. Thanks must especially go to Jan for helping to organise and perform the large-scale mouse experiments. I would also like to thank Dr. Heidi Smith-Vaughan and Dr. Amanda Leach from the Menzie's School of Health Research for providing indigenous isolates, and Andrew Lawrence for the isolates from the Women's and Children's Hospital in Adelaide. Thank you also to Dr. Andre Rickers and Rob King from Geneworks in Adelaide for their assistance with the genomic sequencing. I would also like to thank Professor Tim Mitchell from the University of Glasgow for his suggestions for my project and for his hospitality during my week in Glasgow. Thank you also to Charlie Plumptre, Nadine Verhoeven, Daphne Mermans, Dr. Kerrie Grabowicz, Dr. Trisha Rogers, Ursula Talbot, Dr. Hui Wang, Dr. Alistair Standish, Dr. Kim LeMessurier, Dr. Damien Chong and Melissa Chai for making the Paton lab such a great place to work. A

special thank you must also go to Brock Herdman for his continued company and entertainment since honours. Finally, I must thank my family. In particular, I would like to thank my parents for providing me with ongoing opportunities, support and encouragement. Thank you also to my daughter Lilly for spewing on some of the drafts of my thesis and for those beautiful smiles when I get home in the evening – Daddy loves you!! I would like to especially thank my wife Katie for not only her assistance in editing and formatting this thesis, or for doing the bulk of the housework while I've been writing up, but for her constant love and support, which makes life so enjoyable.

Abbreviations

Abbreviations acceptable to the American Society for Microbiology are used without definition in this thesis. Additional abbreviations are defined when first used in the text, and are listed below.

3HIBDH	3-hydroxyisobutyrate dehydrogenase
ACT	Artemis comparison tool
aorE	Shikimate dehydrogenase
ARs	Accessory regions
BA	Blood agar
BCAAs	Branched-chain amino acids
BgaA	β-galactosidase
BHI	Brain heart infusion broth
BSA	Bovine serum albumin
cCAT	complete-CAT medium
CcpA	Catabolite control protein A
CCR	Carbon catabolite repression
CD	Conserved domain
CGH	Comparative genomic hybridisation
ChoP	Phosphorylcholine
CI	Competitive index
Cml	Chloramphenicol
CSOM	Chronic suppurative otitis media
CSP	Competence stimulating peptide
CTM	cCAT medium supplemented with BSA

DC	Dendritic cell
Ddl	D-alanine-D-alanine dehydrogenase
Dgk	Diacylglycerol kinase
dNTPs	Deoxyribonucleoside triphosphates
DEPC	Diethyl pyrocarbonate
DOC	Sodium deoxycholate
Erm	Erythromycin
GAPDH	Glycerolaldehyde-3-phosphate dehydrogenase
GalE	UDP-glucose 4-epimerase
GDH	Glucose-6-phosphate dehydrogenase
Gen	Gentamycin
Gki	Glucose kinase
HMM	Hidden Markov Model
HylA	Hyaluronate lyase
ICE	Integrative conjugative element
IFN-γ	Interferon γ
IL-1	Interleukin-1
i.n.	Intranasal
i.p.	Intraperitoneal
IPD	Invasive pneumococcal disease
IR	Input ratio
KEGG	Kyoto Encyclopaedia for Genes and Genomes
LD	Limit of detection
LTA	Lipoteichoic acid
LytA	N-acetylmuramoyl-L-alanine amidase
MLST	Multi-locus sequence typing

MQ	MilliQ
MSHR	Menzie’s School of Health Research
NAL	N-acetylneuraminate lyase
NanA	Neuraminidase A
NCBI	National Center for Biotechnology Information
NEB	New England Biolabs
NET	Neutrophil extracellular trap
NmlR$_{sp}$	MerR-like regulator
Nov	Novobiocin
NplT	Neopullulanase
OM	Otitis media
O/N	Overnight
OR	Ouput ratio
ORF	Open reading frame
PavA	Pneumococcal adherence and virulence factor A
PblB	Platelet-binding protein B
PBS	Phosphate buffered saline
PBP	Penicillin-binding protein
PCV7	7-valent pneumococcal conjugate vaccine
PezAT	PezA-PezT TA system
PFGE	Pulsed-field gel electrophoresis
Pht	Pneumococcal histidine triad protein
Pit	Pneumococcal iron transport
Ply	Pneumolysin
PPI-1	Pneumococcal pathogenicity island 1
PPSV23	23 valent pneumococcal polysaccharide vaccine

PsaA	Pneumococcal surface adhesion A
PspA	Pneumococcal surface protein A
PspC	Pneumococcal surface protein C
PsrP	Pneumococcal serine rich protein
PTS	Phosphotransferase system
RBS	Ribosome-binding sites
RecP	Transketolase
$\mathrm{Rel}_{\mathrm{sp}}$	RelA/SpoT homologue
rPAF	Platelet-activating factor receptor
RT	Room temperature
SB	Serum broth
SD	Standard deviation
SDg	Shine-Dalgarno
SDS	Sodium dodecyl sulphate
SEM	Standard error of the mean
SNPs	Single nucleotide polymorphisms
Spe	Spectinomycin
Spi	Signal peptidase I
SpxB	Pyruvate oxidase
ST	Sequence type
Strep	Streptomycin
StrH	β-N-acetylglucosaminidase
TA	Toxin-antitoxin
TBE	Tris borate and EDTA
TE	Tris EDTA
Tet	Tetracycline

THY	Todd-Hewitt broth supplemented with yeast extract
TLR-4	Toll-like receptor 4
TMP	Tympanic membrane perforation
TNF	Tumour necrosis factor
TSB	Tryptic soy broth
WCH	Women’s and Children’s Hospital
WHO	World Health Organisation
Xpt	Xanthine phosphoribosyltransferase
ZmpB	Zinc metalloproteinase B

Chapter 1 – Introduction

1.1 Significance of *Streptococcus pneumoniae*

Streptococcus pneumoniae (the pneumococcus) is a leading cause of pneumonia, otitis media (OM) and invasive disease (IPD) such as bacteraemia and meningitis, and is responsible for more deaths worldwide than any other single pathogen (Forrest *et al.*, 2000). Recent estimates have indicated that in the year 2000 approximately 14.5 million episodes of serious pneumococcal disease occurred in children <5 years of age, with approximately 826,000 fatalities (O'Brien *et al*., 2009). In fact, 11% of all deaths in children <5 were estimated to be due to the pneumococcus (O'Brien *et al*., 2009). Furthermore, in 2000 there were approximately 13.8 million cases of pneumococcal pneumonia, 103,000 cases of meningitis and 540,000 cases of other pneumococcal diseases, such as primary bacteraemia, in children <5 years. Case fatality rates of 5%, 59%, and 45% were recorded for pneumonia, meningitis and other severe pneumococcal diseases, such as primary bacteraemia, respectively (O'Brien *et al*., 2009). Of the total number of deaths, 61% occurred within 10 African and Asian countries and the case fatality rates were particularly high in Africa (Figure 1.1).

The ability of the pneumococcus to behave as both a primary and an opportunistic pathogen (Sjostrom *et al.*, 2006) has often led to the underestimation of the total burden of pneumococcal disease. A significant example of opportunistic pneumococcal disease occurred during the 'Spanish Flu' pandemic of 1918 – 1919, which has been attributed with approximately 50 million deaths. A recent retrospective study of both histological and bacteriological evidence found that the vast majority of deaths attributed to the influenza pandemic were actually due to secondary bacterial pneumonia, of which the pneumococcus was the predominant species (Morens *et al.*, 2008). In particular, it was shown that in the majority of cases the viral infection was

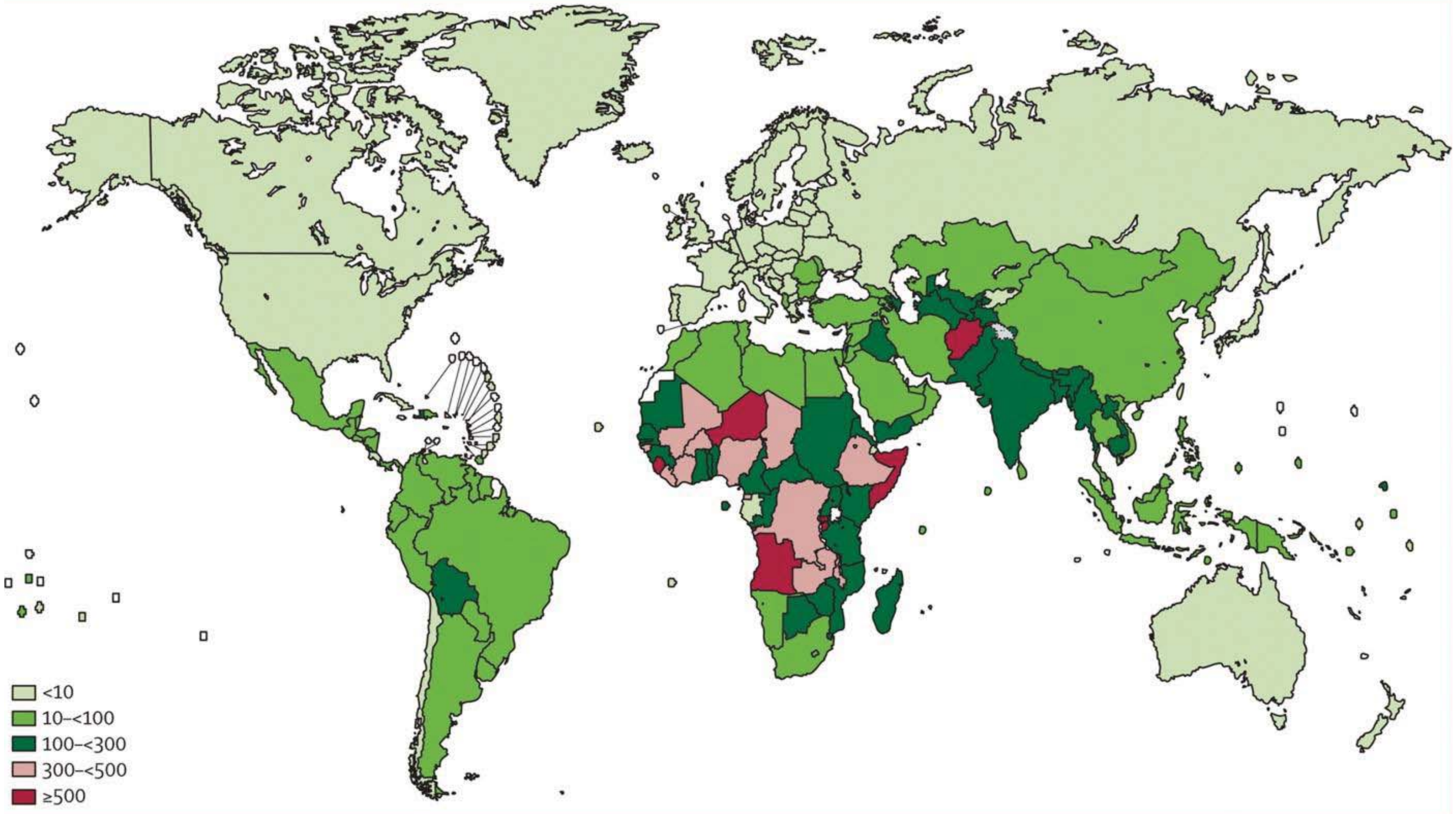


Figure 1.1 Global mortality rate of pneumococcal disease in children <5 years by country (O'Brien, *et al.*, 2009)
The number of deaths in children <5 years per 100,000 due to pneumococcal disease in each country. Deaths also associated with HIV infection are not included.

resolving at the time of death and had been replaced by extensive pulmonary inflammation characteristic of bacterial, rather than viral, infection. Such findings highlight the continued need to develop both preventative and treatment strategies against respiratory pathogens, such as the pneumococcus, especially given the current fears of a new influenza pandemic.

In addition to the mortality of IPD and pneumonia, morbidity due to OM has both a significant social and economic impact on developed and developing countries (Rodgers *et al*., 2009). OM has been reported to be the most common reason for children to go to the doctor, with an estimated economic cost of $3 billion annually in the US (Rodgers *et al*., 2009). The social impact of OM is highlighted by the dire situation that exists in children in remote indigenous communities of Australia (Section 1.1.1).

1.1.1 Pneumococcal disease in indigenous Australians

In terms of the burden of pneumococcal disease, indigenous Australians are amongst the worst affected in the world (Greenwood, 1999; Forrest *et al*., 2000). The burden of disease in remote indigenous communities is significantly greater than that of non-indigenous Australians. For example, the notification rate of IPD within the non-indigenous population in 2006 was approximately 21 cases per 100,000 children <2, which is significantly lower than the notification rate of 73 per 100,000 amongst indigenous children <2 (Roche *et al*., 2008). The same report shows that across all age groups the incidence of IPD is 4.3 times greater in the indigenous than the non-indigenous population. The incidence of serious OM, such as tympanic membrane perforation (TMP) and chronic suppurative otitis media (CSOM) is also very high in children of remote indigenous communities. The incidence of TMP in children <5 has been shown to be 24% and the incidence of CSOM between 15% and 24% (Morris *et al*., 2005; Rothstein *et al*., 2007). Given that the World Health Organisation (WHO) considers a CSOM prevalence of greater than 4% to be a “massive health problem” (Coates, 2002), it is clear that the situation is indeed serious amongst indigenous children, particularly when recent surveys have shown that the 7-valent pneumococcal conjugate vaccine (PCV7) (Section 1.4.2) is completely ineffective at offering protection against pneumococcal OM in these communities (MacKenzie *et al*., 2009).

1.2 Pneumococcal infection

S. pneumoniae is a gram positive, encapsulated, diplococcus, most commonly found colonising the nasopharynx of healthy children <5 years of age. The pneumococcus is almost exclusively a human pathogen. However, a number of animal models have been used for research on pneumococcal pathogenesis, which include mice, chinchillas, rabbits and rats (reviewed in Siber *et al*. [2008]).

1.2.1 Asymptomatic carriage and transmission

Whilst the pneumococcus is a very significant cause of human mortality and morbidity (Section 1.1), the bacterium is more likely to be found being carried asymptomatically in the nasopharynx (Gray *et al*., 1980). The carriage rate of *S. pneumoniae* is highest in children <5 years of age, and in a Finnish study has been found to start within the first year of life at a rate of approximately 13% carriage in those <6 months of age, before increasing to 43% carriage in children >19 months of age (Syrjanen *et al*., 2001). However, as reviewed by Bogaert *et al*. (2004), the carriage rate can vary considerably between regions and can be as high as 90% in some settings. The rate of carriage is influenced by environmental and socioeconomic factors, which include family size (particularly the number of older siblings), income, recent antibiotic usage, smoking (including passive smoking) and crowding (reviewed in Bogaert *et al*. [2004]). The effect of crowding is particularly clear in child care centres, where the relative risk of pneumococcal infection is 1.6 when compared to children at home (Bogaert *et al*., 2001). Similarly, a survey of pneumococcal carriage over a year in a French orphanage showed a carriage rate of 82% in children <24 months of age (Raymond *et al*., 2000). In the same study serotyping and molecular typing by pulsed-field gel electrophoresis (PFGE) showed that children were colonised by a limited number of clones, which suggested considerable horizontal transmission. Long-term carriage has also been shown to promote low-level serotype-specific immunity, which is thought to provide short-term protection from reinfection by the same serotype (Musher *et al*., 1998; Simell *et al*., 2001).

1.2.2 Pneumococcal disease

Pneumococcal disease is usually thought to be caused by a strain of the bacterium that has only recently commenced nasopharyngeal colonisation of the infected person (Gray *et al*., 1980). The mechanism by which the pneumococcus progresses from asymptomatic carriage to disease is not clear. However, whilst progression to IPD is often thought to be primarily due to host factors (Alanee *et al*., 2007; Chen *et al*., 2009), the situation is further complicated by differences in the ability of different serotypes and genotypes to cause disease (Section 1.5.2). Therefore, it is likely that factors of both the host and the infecting bacterium are important in determining whether or not disease will occur.

Those most at risk from pneumococcal disease include children <5 years, the elderly >65 years, immunocompromised individuals and people with chronic underlying illness (Butler *et al*., 2004). In addition, people with a defective innate immune system, particularly regarding complement and immunoglobulin production are especially susceptible to pneumococcal disease (Bruyn *et al*., 1992), which highlights the immune defences required for protection against pneumococcal disease. However, the association of certain serotypes, such as serotypes 1 and 7F, with disease in people usually without any underlying condition (Sjostrom *et al*., 2006), shows that it is not only the immunocompromised that suffer from pneumococcal disease. In addition, the increased susceptibility to pneumococcal disease following infection by respiratory viruses, such as influenza, has also been shown to contribute to the incidence of pneumococcal disease (reviewed in McCullers [2006]; Morens *et al*., 2008).

1.3 Molecular mechanisms of pneumococcal carriage and disease

In order for the pneumococcus to migrate and survive in a new niche and cause disease, the bacterium must navigate through and evade a number of host defences, which range from physical barriers to molecular and cellular defences. In addition, the pneumococcus must adapt to a new nutrient environment, which may differ from the nasopharynx in terms of changes in the availability of certain sugars, amino acids and metal ions. Unsurprisingly, the pneumococcus has developed a complex array of molecular mechanisms required for survival in the host. These mechanisms include an array of specific factors with roles in survival *in vivo* that are at least partially understood. However, a number of poorly understood broader regulatory networks and adaptations, which include pneumococcal phase variation and biofilm formation, appear to also have important roles in survival in the host.

1.3.1 The contribution of virulence factors to pneumococcal survival *in vivo*

The pneumococcus has been shown to encode a broad array of factors that are required for virulence in animal models of infection and this has recently been reviewed in Mitchell and Mitchell (2010). Pneumococcal virulence factors are involved in such roles as adherence to host cells and transcellular migration across epithelial and endothelial surfaces, interference with the host's immune response, transport and sequestration of nutrients and intra- and interspecies competition. Many of the key factors described below are represented in Figure 1.2.

1.3.1.1 Adherence and transcellular migration across epithelial surfaces

Pneumococcal infection commences following successful adherence by the bacterium to the nasopharyngeal surface of a new host. It has been shown with the TIGR4 strain in neonatal rats that the numbers of pneumococci that colonise the nasopharynx appear to reach a steady state that is independent of the initial challenge dose (Margolis *et al*., 2010). In resting nasopharyngeal tissue the pneumococcus binds to host cell surface carbohydrates, such as N-acetyl-glucosamine (Andersson *et al*., 1983). A number of pneumococcal surface molecules have been shown to mediate interactions with host epithelial surfaces, such as phosphorylcholine (ChoP) moieties on lipoteichoic acid (LTA) (Cundell *et al*., 1995a). In addition, non-specific physiochemical interactions between components of the bacterial cell wall and host cell surfaces have been suggested to contribute to pneumococcal adherence (Swiatlo *et al*., 2002). It has also been suggested that the substrate-binding component of the *psa* locus pneumococcal surface adhesion A (PsaA) contributes to adherence. (Berry & Paton, 1996; Romero-Steiner *et al*., 2003; McAllister *et al*., 2004; Romero-Steiner *et al*., 2006). However, whilst it is thought that the role of PsaA in adherence is indirect and is dependent on the protein's role in manganese transport (Section 1.3.1.3), it has recently been reported that PsaA directly mediates adherence by binding E-cadherin on nasopharyngeal cells (Anderton *et al*., 2007).

The activity of pneumococcal neuraminidases has been shown to assist in the establishment of colonisation through a number of processes. Neuraminidase A (NanA), cleaves the sialic acid residues of host glycoproteins, glycolipids and oligosaccharides,

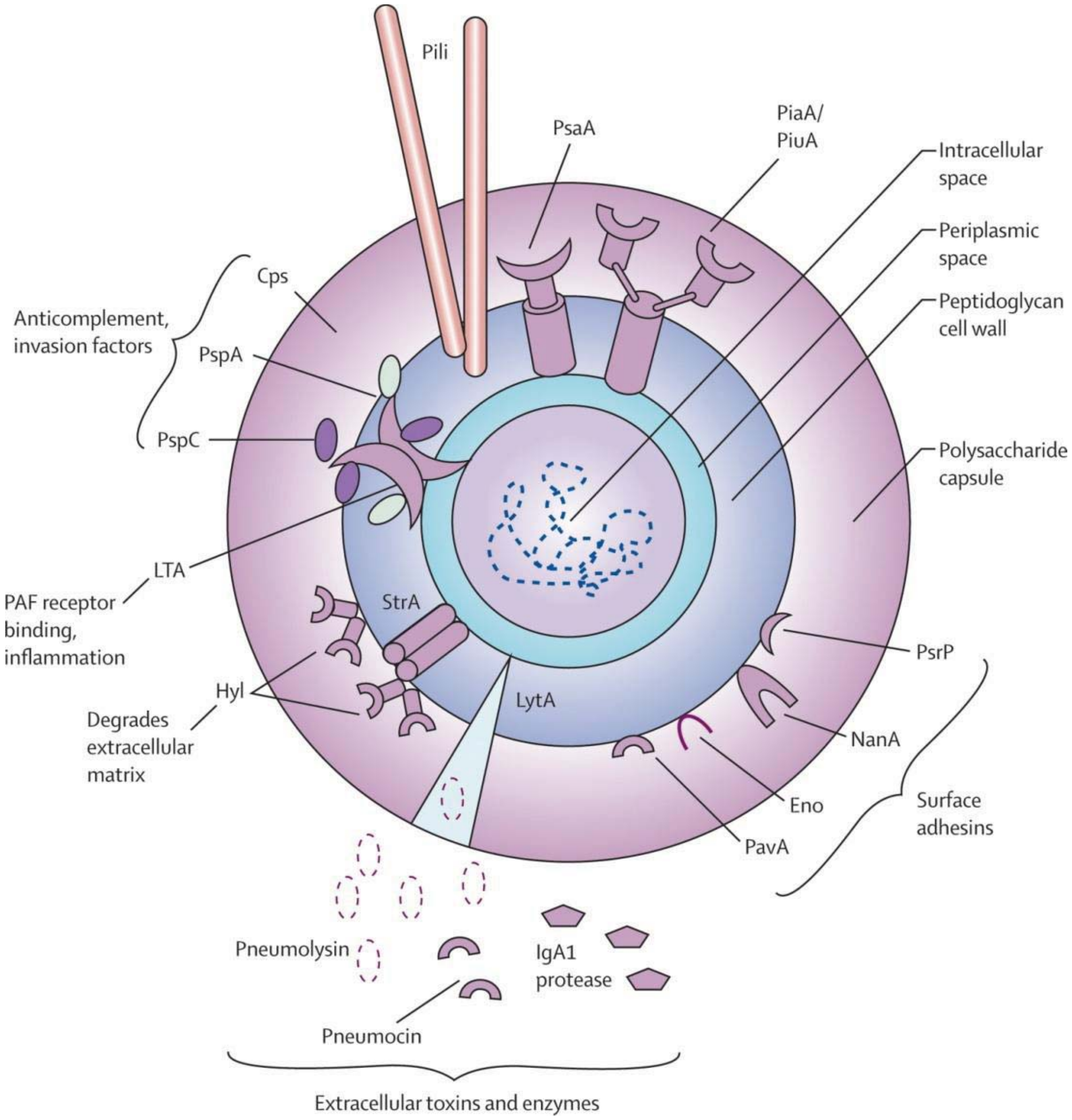


Figure 1.2 Virulence factors of *Streptococcus pneumoniae* (van der Poll & Opal 2009)
A schematic representation of key pneumococcal virulence factors, most of which are described and discussed in the Introduction. Abbreviations that are not included in the text are; Cps (polysaccharide capsule), Hyl (hylauronate lyase), StrA (sortase), Eno (enolase) and PiuA (pneumococcal iron uptake).

which is thought to increase the accessibility of N-acetyl-glucosamine receptors on the surface of host epithelial cells (Tong *et al*., 1999; Linder *et al*., 1994; Linder *et al*., 1992; Andersson *et al*., 1983). The neuraminidase activity of various viruses may also help contribute to increased adherence of the pneumococcus (McCullers *et al*., 2001). NanA-mediated desialyation of carbohydrates lining the eustachian tube has been linked to the development of OM (Linder *et al*., 1994; Linder *et al*., 1992). NanA has also been shown to promote brain endothelial invasion (Uchiyama *et al*., 2009). In addition, NanA's association with meningitis, pneumonia and OM could be facilitated by its role in biofilm formation (Parker *et al*., 2009). Interestingly, a second neuraminidase, NanB, has been shown to exhibit optimum activity at pH 4.5, compared to pH 6.5 – 7.0 for NanA, which suggests that the two enzymes exist to allow independent neuraminidase activity in different *in vivo* environments (Berry *et al*., 1996).

In addition to NanA activity, a number of other pneumococcal surface glycosidases have been identified, which modify both N-linked and O-linked glycans (Muramatsu *et al*., 2001; Umemoto *et al*., 1977; Zahner & Hakenbeck., 2000). The activities of N-glycosidases, NanA, β-galactosidases (BgaA) and β-N-acetylglucosaminidase (StrH) have recently been shown to act sequentially to deglycosylate N-linked glycans (King *et al*., 2006) and promote resistance to opsonophagocytic killing by human neutrophils (Dalia *et al*., 2010). The activities of glycosidases targeting O-linked glycans have recently been shown to increase the ability of the pneumococcus to colonise the upper respiratory tract (Marion *et al*., 2009).

Though not completely understood, an important marker of the progression from asymptomatic carriage to disease is the local production of inflammatory mediators, such as tumour necrosis factor (TNF) α and interleukin 1 (IL-1), which also exist in the presence of viral infections. These may contribute to the commencement of secondary bacterial infections (reviewed in McCullers [2006]). However, the behaviour of the pneumococcus as a primary pathogen in other cases indicates that the processes involved in both carriage and disease are complex and closely linked (Margolis & Levin, 2007; Sjostrom *et al*., 2006; Briles *et al*., 2005).

The production of inflammatory cytokines stimulates changes in the type and number of receptors on the surface of epithelial and endothelial cells and promotes an increase in the affinity of pneumococcal cell-wall ChoP for platelet-activating factor receptor (rPAF) (Cundell *et al*, 1995a). Binding to rPAF triggers transcellular migration of the pneumococcus through the epithelium and vascular endothelium, which can lead to entry into the bloodstream, and if followed by survival and replication can lead to bacteraemia. The interaction between rPAF and ChoP has also been shown to contribute to passage of the pneumococcus across the blood-brain barrier (Ring *et al*., 1998).

A number of other protein factors of the pneumococcus have also been shown to directly mediate interactions with the host. Pneumococcal surface protein C (PspC) has been shown to exhibit increased affinity for immobilised sialic acid and lacto-N-neotetraose on cytokine-activated human cells, which promotes attachment (Rosenow *et al*., 1997). In addition, PspC also binds to the polymeric immunoglobulin receptor and the secretory component of IgA, which increases migration through mucosal barriers (Zhang *et al*., 2000; Hammerschmidt *et al*., 1997). IgA1 protease, the pneumococcal pili and the pneumococcal serine rich protein (PsrP) have also been shown to contribute to adherence (Kilian *et al*., 1979; Wani *et al*., 1996; Barocchi *et al*., 2006; Bagnoli *et al*., 2008; Obert *et al*., 2006). However, the pili and PsrP are present in only a selection of strains, which indicates that they are not required by all strains and will be discussed in Section 1.5.3.2. The pneumococcal adherence and virulence factor A (PavA) has been shown to be required for full virulence in sepsis and meningitis models of infection and has been shown to bind to fibronectin and mediate attachment to endothelial cells (Holmes *et al*., 2001). In addition, the glycolytic enzymes, glycerolaldehyde-3-

phosphate dehydrogenase (GAPDH) and enolase, have been shown to mediate binding to human plasminogen, which may facilitate transmigration of pneumococci through the basement membrane (Bergmann *et al*., 2004; Bergmann *et al*., 2005).

Although not specifically required for attachment to host cells, pneumococcal hyaluronate lyase (HylA) has been shown to facilitate invasion of host tissues in other species through the enzyme's ability to degrade hyaluronan, which is a major component of the extracellular matrix (Berry *et al*., 1994; reviewed in Jedrzejas *et al*. [2004]).

1.3.1.2 Interference with the host's immune response

An important feature for any pathogen is the ability to evade and influence the host's immune system in such a way that promotes survival within the required niches of the host. As reviewed in Kadioglu & Andrew (2004), a feature of the immune response against the pneumococcus during disease is potent inflammation involving activation of complement, and phagocytosis by activated resident macrophages and recruited neutrophils. As a consequence, many virulence factors of the pneumococcus target complement and evasion of phagocytosis. The most important virulence factor for avoidance of the host's immune system is the polysaccharide capsule, which is required for both colonisation and systemic infection (Avery & Dubos, 1931; Magee *et al*., 2001), and will be specifically discussed in Section 1.3.1.5.

A number of surface proteins, such as PspC and the pneumococcal histidine triad proteins (Pht) (although the latter has been challenged [Melin *et al*., 2010]), have been shown to bind the complement regulator, factor H, which inhibits complement deposition and activation of the alternative pathway (Dave *et al*., 2001; Ogunniyi *et al*., 2009). Also targeting the alternative complement pathway is the pneumococcal surface protein A (PspA), which has been shown to block C3b deposition, thus also inhibiting complement activation (Tu *et al*., 1999; Ren *et al*., 2004a; Ren *et al*., 2004b).

Furthermore, the ability of PspA to bind human lactoferrin has been shown to reduce complement activation and inhibit the bactericidal activity of the iron-depleted form of the protein (Shaper *et al*., 2004). The importance of PspA to pneumococcal infection is highlighted by a reduction in colonisation and infection of the lungs and blood following challenge with PspA-deficient mutants (McDaniel *et al*., 1987; Ogunniyi *et al*., 2007).

An important immunomodulatory virulence factor of the pneumococcus is the pore-forming toxin, pneumolysin (Ply). Mutants lacking Ply have been shown to be attenuated in both intranasal and systemic murine models of infection (Berry *et al*., 1989a). More specifically, Ply-deficient mutants have been characterised by reduced induction of pulmonary inflammation due to delayed cell recruitment into the lungs, particularly affecting neutrophil responses and the distribution of B and T lymphocytes in and around inflamed bronchioles (Canvin *et al*., 1995; Kadioglu *et al*., 2000).

In contrast to the complement inhibitory properties of PspC and PspA, Ply has also been shown to activate the classical complement pathway in an antibody-independent fashion, leading to serum complement depletion (Paton *et al*., 1984; Alcantara *et al*., 2001). Ply has also been shown to be a potent stimulator of inflammation (Johnson *et al*., 1981; Berry *et al*., 1989a; Houldsworth *et al*., 1994; Berry *et al*., 1995; Braun *et al*., 1999; Cockeran *et al*., 2001), which has been linked to a number of properties, including the ability of Ply to bind toll-like receptor 4 (TLR-4) (Malley *et al*., 2003). In addition, Ply has been shown to induce apoptosis in murine dendritic cells (DCs) (Colino & Snapper, 2003), respiratory cells (Srivastava *et al*., 2005; Schmeck, *et al*., 2004) and neuronal cells (Bermpohl *et al*., 2005; Braun *et al*., 2002; Mitchell *et al*., 2004), due to its cytolytic properties. The cytolytic activity of Ply is characterised by the oligmerisation of approximately 50 toxin monomers, which forms a 30 nm pore in the target cell membrane (Morgan *et al*., 1995). The mutagenesis of distinct regions within Ply responsible for complement activation and cytolytic activity has shown that these activities are independent of each other and contribute separately to pathogenesis (Mitchell *et al*., 1991; Berry *et al*., 1995). Ply has also recently been shown to trigger caspase-dependent apoptosis of human DCs, which dampens the production of inflammatory cytokines, such as IL-12 and IL-1β (Littmann *et al*., 2009). Interestingly, cytolytic activity-deficient Ply strains (Kirkham *et al*., 2006; Lock *et al*., 1996) were shown to trigger a more proinflammatory cytokine response from human DCs than the full active toxin (Littmann *et al*., 2009). Ply release following autolysin-mediated autolysis (see below), has been shown to be a potent activator of production of reactive oxygen species in human neutrophils, which is released into an intracellular compartment of the targeted immune cell rather than excreted (Martner *et al*., 2008). In addition, the toxin's cytolytic activity impairs mucous-mediated clearance by the inhibition of ciliary beating in the respiratory tract, which is augmented by the activity of HylA (Steinfort *et al*., 1989; Feldman *et al*., 1990; Feldman *et al*., 2007).

The major pneumococcal autolysin (N-acetylmuramoyl-L-alanine amidase; LytA), is required for full virulence in mice, as LytA-deficient mutants are rapidly cleared from the lungs and rarely translocate into the blood (Berry & Paton, 2000; Canvin *et al*., 1995; Berry *et al*., 1989b). LytA has been shown to promote inflammation due to the release of cytoplasmic proteins, such as Ply, and by triggering the release of proinflammatory components of the bacterial cell wall during autolysis (Boulnois *et al*., 1991; Tuomanen *et al*., 1985; Chetty & Kreger, 1981). However, Ply release has more recently been shown to not be completely dependent on LytA activity in some strains (Balachandran *et al*., 2001). The remnants of pneumococci that had undergone LytA-mediated autolysis have also been shown to inhibit phagocytosis and indirectly protect unlysed pneumococci from phagocytosis (Martner *et al*., 2009). In the same study the anti-phagocytic properties of lysed bacteria included the suppression of pro-phagocytic cytokines, such as TNF, interferon γ (IFN-γ) and IL-12, which were produced in the presence of LytA-deficient mutants. The impact on cytokine production was shown to be specific as the production of IL-6, IL-8 and IL-10 was unaffected.

In addition to promoting colonisation by exposing host surface receptors, NanA-dependent desialation of pneumococcus-bound immune system components of the respiratory tract, such as lactoferrin, secretory component and IgA2, has been shown to increase *in vivo* survival (King *et al*., 2004). In addition to NanA and NanB, some strains encode a third neuraminidase, NanC, whose presence has been suggested to correlate with meningitis-causing strains rather than asymptomatic carriage (Pettigrew *et al*., 2006).

Resistance of the pneumococcus to the innate immune response has also been shown to require the regulation of specific metabolic genes, such as those acted upon by the MerR-like regulator (NmlR$_{sp}$), which is important in resistance to nitric oxide stress associated with phagocytosis and pulmonary inflammation. As such, NmlR$_{sp}$ has been shown to be required for full pneumococcal systemic virulence in a murine model of infection (Stroeher *et al*., 2007).

1.3.1.3 Transport and sequestration of nutrients in the host

For a bacterium such as the pneumococcus to survive in diverse *in vivo* niches, it is important for it to be able to adapt to changes in nutrient availability between different environments. Given that the pneumococcus lacks an electron transport chain, the bacterium is particularly reliant on carbohydrates for energy, which is highlighted by the vast array of sugar transport systems and pathways for carbohydrate metabolism encoded within the genome (Tettelin *et al*., 2001; Hoskins *et al*., 2001). The catabolite control protein A (CcpA) has been shown to regulate the hierarchical utilisation of some sugars in order to promote optimal growth in a process called carbon catabolite repression (CCR) (Iyer *et al*., 2005). CcpA has been shown to be required for wild-type equivalent systemic virulence (Giammarinaro & Paton, 2002). In addition, CcpA-deficient mutants have been shown to exhibit attenuated virulence in a pneumonia model of infection and have a reduced capacity to colonise the nasopharynx (Iyer *et al*., 2005). The ability of the pneumococcus to take advantage of complex structural host-derived sugars has been suggested by the NanA-mediated degradation of mucin (Yesilkaya *et al*., 2008). Minimal medium supplemented with mucin was shown to allow growth at a rate equivalent to that of a nutrient-rich broth, which was thought to be at least in part facilitated by NanA (Yesilkaya *et al*., 2008).

In addition, the stringent response has recently been characterised in the pneumococcus and has been shown to be regulated by a RelA/SpoT homologue (Rel$_{sp}$). Rel$_{sp}$ regulates the production of the alarmone, (p)ppGpp, which mediates a global response to nutrient limitation and other stresses (Kazmierczak *et al*., 2009). The importance of Rel$_{sp}$ to virulence was highlighted by the severe attenuation of Rel$_{sp}$-deficient mutants *in vivo*. Interestingly, the same study showed that Ply was up-regulated during the stringent response in a Rel$_{sp}$-dependent manner (Kazmierczak *et al*., 2009).

A number of transporters have also been implicated in obtaining limited nutrients including metals. For example the *pia* locus of the pneumococcal pathogenicity island 1 (PPI-1) has been shown to encode an important iron acquisition ABC transporter (Brown *et al*., 2002), and is required for full pulmonary and systemic virulence in murine models of infection (Brown *et al*., 2001). In addition, an ABC transporter required for full virulence is encoded by the *psa* locus, which includes the substrate-binding protein PsaA. PsaA has been shown to be important for manganese transport as PsaA-deficient mutants only grow in media supplemented with manganese (Dintilhac *et al*., 1997). Whilst PsaA has been shown to be important for colonisation (Section 1.3.1.1), PsaA-deficient mutants have also been shown to have massively reduced virulence, which has been suggested to be due to impaired growth in a manganese-poor environment, reduced adherence and hypersensitivity to oxidative stress (Berry & Paton, 1996; Dintilhac *et al*., 1997; Tseng *et al*., 2002; McAllister *et al*., 2004).

1.3.1.4 Intra- and interspecies competition *in vivo*

The pneumococcus encounters much competition from commensal bacteria for space and nutrients within the human nasopharynx. Interspecies competition has been found to occur between the pneumococcus and commensals, such as other α-haemolytic streptocococci, leading to a balanced coexistence (Ghaffar *et al*., 1999). The competition also extends to other species capable of causing disease such as *Haemophilus influenzae*, *Moraxella catarrhalis*, *Staphylococcus aureus* and *Neisseria meningitidis* (reviewed in Bogaert *et al*. [2004]). Pyruvate oxidase (SpxB)-mediated production of hydrogen peroxide by the pneumococcus has been shown to inhibit the growth of *H. influenzae*, *M. catarrhalis*, *S. aureus* and *N. meningitidis in vitro*, which can be reversed by the introduction of catalase into the culture medium (Pericone *et al*., 2000; McLeod & Gordon 1922). However, the extent to which interspecies competition is successful varies considerably between different strains (Margolis *et al*., 2010; Lysenko *et al*., 2005). In addition to its role in virulence described above, NanA has also been proposed to mediate interspecies competition by cleaving the terminal sialic acid residues of lipooligosaccharide of *H. influenzae* and *N. meningitidis* (Shakhnovich *et al*., 2002). However, co-colonisation of *H. influenzae* with the pneumococcus generally led to rapid clearance of the pneumococcus in a murine model of infection (Lysenko *et al*., 2005). The production of bacteriocins by the pneumococcus, such as that encoded within the *blp* locus, has also been implicated in both intra- and interspecies competition, by targeting strains that lack the specific immunity system to these bacteriocins (Dawid *et al*., 2007).

1.3.1.5 Polysaccharide capsule of the pneumococcus

The polysaccharide capsule has often been considered to be the most important virulence factor of the pneumococcus, as strains lacking the capsule are avirulent in animal models (Avery & Dubos, 1931; Watson *et al*., 1990) and exhibit a significantly reduced capacity to colonise the nasopharynx (Magee *et al*., 2001). The capsule is important for avoidance of the host's immune defences, which is achieved by a number of different mechanisms. At least 90 immunologically distinct capsular serotypes have been identified, which promote avoidance of the host's immune system by providing population-wide antigenic variability to limit immune recognition at the species level (Henrichsen, 1995). The capsule also interferes with mucous-mediated clearance by electrostatic repulsion due to the net negative charge of most serotypes (Nelson *et al*., 2007) and promotes avoidance of the host's immune system by significantly reducing opsonophagocytosis (Magee *et al*., 2001; Hardy *et al*., 2001; Brown, 1985; Winkelstein, 1984). The capsule provides an inert barrier surrounding the vulnerable cell wall, which contains numerous surface molecules, such as LTA and peptidoglycan that readily activate the alternative complement pathway (reviewed in Kenzel & Henneke [2006]; Guan & Mariuzza, 2007). Reduced trapping by neutrophil extracellular traps (NETs) adds to the repertoire of protective functions provided by the capsule (Wartha *et al*., 2007).

As will be discussed in Section 1.5, variation exists between serotypes in prevalence (Hausdorff *et al*., 2000a; Hausdorff *et al*., 2000b; Hausdorff *et al*., 2005), invasive potential (Brueggemann *et al*., 2003; Brueggemann *et al*., 2004; Hanage *et al*., 2005; Sandgren *et al*., 2004; Austrian *et al*., 1981), age distribution (Hausdorff *et al*., 2005), tendency to cause outbreaks (Hausdorff *et al*., 2005; Gleich *et al*., 2000) and degree of association with antimicrobial resistance (McCormick *et al*., 2003). Increased production of capsule at the surface of the pneumococcus *in vitro* has been shown to increase virulence *in vivo* (MacLeod *et al*., 1950). Specifically, increased encapsulation has been shown to significantly reduce opsonophagocytosis, which is a feature of opaque-phase pneumococci (Kim *et al*., 1999) (Section 1.3.2). Additional factors such as metabolic cost (number of carbons or number of high energy molecules consumed per repeat unit), degree of encapsulation (Weinberger *et al*., 2009) and deposition of complement (Hyams *et al*., 2010) have been shown to differ between serotypes following comparisons between a number of isogenic capsule-swap mutants in various *in vitro* and *in vivo* assays. Differences relating to encapsulation and metabolic cost were shown in most cases to be associated with greater persistence in nasopharyngeal colonisation in both mice and humans (Weinberger *et al*., 2009). By contrast, Hyams *et al*., 2010 showed that when the thickness of the capsule layer was kept constant, serotypes 4 and 7F were found to have been bound by less complement than serotype 6A and 23F, which was inversely correlated with neutrophil-mediated killing *in vitro* and increased virulence *in vivo*. Therefore, differences between serotypes in such characteristics as IPD-potential and carriage prevalence appear to have a capsule-dependent component. However, there is also evidence that the serotype of the capsule does not alone determine the IPD-potential or carriage prevalence of a given strain. For example, a study where the serotype 3 capsule locus was used to replace the wild-type locus of a serotype 2, 5 and 6B strain, showed vastly different effects on virulence (Kelly *et al*., 1994). The serotype 3 capsule had no effect on the serotype 2 strain’s virulence, eliminated the virulence of the serotype 5 strain and resulted in a 100-fold reduction of the LD_{50} of the serotype 6B strain (Kelly *et al*., 1994). Such a wide variety of outcomes following replacement of different wild-type capsule loci with the serotype 3 capsule locus highlights the importance of both the serotype and background genotype in IPD potential.

In summary, the pneumococcus possesses a vast array of factors that are required for virulence by promoting survival *in vivo* through a variety of different mechanisms. Whilst the above discussion of virulence factors and their roles *in vivo* is by no means exhaustive, understanding the range of mechanisms that promote IPD is important when attempting to identify new factors that predispose some strains to cause IPD, whilst others are carried asymptomatically.

1.3.2 Pneumococcal phase variation

Weiser *et al*., 1994 first identified pneumococcal phase variation by showing that under oblique, transmitted light on a transparent medium, three different opacity phenotypes of pneumococci could be observed. These included a transparent, semitransparent (an intermediate phase) and opaque phenotype. The frequency of switching between opaque and transparent phases during *in vitro* exponential growth has been shown to be highly variable between different isolates, varying from 10^{-3} to 10^{-6} per generation (Weiser *et al*., 1994). The importance of phase variation in pneumococcal pathogenesis has been highlighted by studies showing selection for the transparent phenotype during nasopharyngeal colonisation (Weiser *et al*., 1994; Weiser *et al*., 1996) and for adherence to lung epithelium following cytokine stimulation (Cundell *et al*., 1995b). By contrast, cytokine stimulation was shown to have no effect on adherence by opaque-phase pneumococci. Whilst adherence to cytokine-activated lung epithelial cells has been shown to be greatest in the transparent phase, prolonged presence of pneumococci in the lungs of infected mice has been shown to select for the opaque phase with one study showing that 90% of pneumococci recovered at 21 days post-challenge were of the opaque phase (Briles *et al*., 2003).

Molecular comparisons between opaque and transparent variants have uncovered a range of differences relating to encapsulation, cell wall composition and protein expression. Opaque-phase variants have been shown to exhibit 1.2- to 5.6-fold greater encapsulation than those of the transparent phase, which is thought to contribute to the greater systemic virulence of opaque-phase pneumococci (Kim & Weiser, 1998). This increased encapsulation has been shown to play a role in increased resistance to complement-mediated phagocytosis by reducing complement deposition (Hyams, *et al*., 2010; Kim *et al*., 1999) and is likely to also have a role in reduced mucous-mediated clearance (Nelson *et al*., 2007). Opaque-phase variants are also characterised by increased PspA expression, contributing to increased inhibition of complement (Kim & Weiser, 1998; Tu *et al*., 1999; Ren *et al*., 2004a; Ren *et al*., 2004b). By contrast, transparent-phase pneumococci are characterised by reduced encapsulation and increased expression of proteins such as PspC, SpxB, LytA and NanA (Weiser *et al*., 1996; Rosenow *et al*., 1997; Overweg *et al*., 2000; King *et al*., 2004). In particular, greater NanA-dependent desialation of bacterial clearance components of the respiratory tract (Section 1.3.1.2) has been observed in association with transparent-phase pneumococci (King *et al*., 2004). Opaque-phase variants have been shown to possess reduced membrane fluidity when compared to those of the transparent phase, most likely due to differences in the proportions of various fatty acids within the membrane (Aricha *et al*., 2004). Of particular significance is the increased presence of ChoP in the cell wall of transparent-phase variants, which has been suggested to contribute to increased adherence to the nasopharyngeal epithelium and cytokine-activated lung epithelial surfaces (Kim & Weiser, 1998; Cundell *et al*., 1995b). Transcytosis across the blood-brain barrier has also been shown to be greater in transparent-phase pneumococci through a process at least partly dependent on rPAF (Ring *et al*., 1998). A role for pneumococcal phase variation in long term colonisation has been proposed following the identification of subpopulations of transparent- and opaque-phase pneumococci at the surface and within the tissue of the nasopharynx, respectively (Briles *et al*., 2005). These observations support the idea that progression to systemic disease might be an accidental side-effect due to an overzealous bacterium or a break down in the host defences. The features of the opaque phase that promote survival in the deeper nasopharyngeal tissue might overlap with the requirements for survival in the blood.

The mechanism behind the regulation of pneumococcal phase variation has yet to be determined. However, the presence of an A-C box element in the genome has been shown to increase the rate of phase switching by 10^3 fold in some strains (Saluja & Weiser, 1995). Mathematical modelling has been used for *N. meningitidis* to propose that an increased rate of phase switching within specified ‘contingency loci’ (reviewed in Moxon *et al*. [2006]) could increase the ability of the bacterium to colonise diverse host environments and inadvertently increase the invasive potential of a given strain (Meyers *et al*., 2003). Such an idea would suggest that randomly arriving at a pattern of gene expression that promotes survival in the blood would promote invasion, which is supported by work in *H. influenzae* that suggests that blood-derived bacteria originate from a single cell invading and surviving in the blood, rather than as a cooperative process mediated by multiple cells (Moxon *et al*., 1978; Margolis & Levin, 2006). In such a model it is suggested that invasion and subsequent survival and proliferation in the blood is a result of short-sighted within-host evolution. However, the precise dynamics of invasion of the blood have yet to be established for the pneumococcus. Clearly much work remains to be done in order to fully understand phase variation in the pneumococcus.

An alternative pneumococcal phase variation phenomenon has been observed in strains of serotypes 3, 8 and 37, which has been associated with the initial attachment stage of biofilm formation (Waite *et al*., 2001; Waite *et al*., 2003). This variation occurs between encapsulated and unencapsulated variants, is apparently independent of opacity variation and appears to be restricted to serotypes where a mucoid phenotype exists

(Waite *et al*., 2001; Waite *et al*., 2003). Interestingly, the unencapsulated phenotype has been observed during the early stages of adherence and invasion of the lung epithelium (Hammerschmidt *et al*., 2005). Capsule phase variation has been shown to be caused by the random gain or loss of tandem duplications within the capsule locus (Domenech *et al*., 2009; McEllistrem *et al*., 2007; Waite *et al*., 2003).

1.3.3 Physiological states of the pneumococcus

The pneumococcus has recently been observed to exist in two physiological states, which *in vitro* exist as planktonic growth, such as in liquid culture, or a sessile state, such as on agar or in a submerged biofilm (Oggioni *et al*., 2006). These different physiological states exhibit distinct patterns of gene expression that are similar both *in vivo* and *in vitro* and are associated with different types of disease. The sessile state is usually associated with colonisation of the nasopharynx, pneumonia, OM and meningitis (Hall-Stoodley *et al*., 2006; Oggioni *et al*., 2006). In contrast, pneumococci in the blood during bacteraemia are in the planktonic state (Oggioni *et al*., 2006).

Biofilm formation occurs in a number of stages, commencing with the attachment of planktonic cells to the relevant surface, such as the nasopharynx, followed by the formation of cellular aggregates and finally the formation of a mature biofilm (Allegrucci *et al*., 2006). The extracellular matrix of a mature biofilm is thought to consist largely of polysaccharide and excreted DNA (Hall-Stoodley *et al*., 2006; Donlan *et al*., 2004). In addition, a competence-associated mechanism of programmed-cell death has been shown to be involved in the formation of the biofilm structure, by targeting cells based on their spatial location within the forming biofilm. In this system of microbial fratricide, non-competent cells undergo LytA- and CbpD-dependent autolysis, whilst competent cells remain immune (Havarstein *et al*., 2006). In addition, the switch between sessile and planktonic states, though not fully understood, has been suggested to be linked to genetic competence (Oggioni *et al*., 2006). The expression of genes required for genetic competence has been shown to be much greater in the sessile state than in the planktonic state.

The importance of biofilm formation is highlighted by increased resistance to environmental stresses, such as oxidative stress (Bortoni *et al*., 2009) and increased resistance to antibiotics (Hall-Stoodley *et al*., 2008). Furthermore, pneumococcal biofilms are more resistant to phagocytosis, which is thought to be facilitated by a greater resistance against NETs (reviewed in Urban *et al*. [2006]). However, an association between invasive potential and a tendency to undergo early biofilm formation was not identified in a selection of invasive and non-invasive serotype 6A and 6B isolates (Lizcano *et al*., 2010). Clearly a more comprehensive comparison of the kinetics of biofilm formation between different strains is required to assess the role of biofilm formation in invasive potential.

1.3.4 Genetic competence in *S. pneumoniae*

A biologically and historically important feature of the pneumococcus is its ability to undergo natural genetic transformation. Natural transformation contributes to the spread of antibiotic resistance (Section 1.4.1), the horizontal acquisition of virulence genes (Section 1.5.3.2) and vaccine escape through serotype-switching (Section 1.4.2). Natural transformation of the pneumococcus is achieved through a tightly regulated process leading to the development of a genetically competent state. Competence commences and ends abruptly with a window of approximately 30 mins during the early exponential phase of *in vitro* growth (reviewed in Johnsborg & Havarstein [2009]).

Figure 1.3 summarises what is currently understood of the regulation of competence. Initially, competence regulation occurs through the quorum sensing activities of competence stimulating peptide (CSP), which is encoded by *comC* and exported by a specific transporter encoded by *comAB* (Havarstein *et al*., 1995). At a particular culture density, the concentration of extracellular CSP reaches a threshold

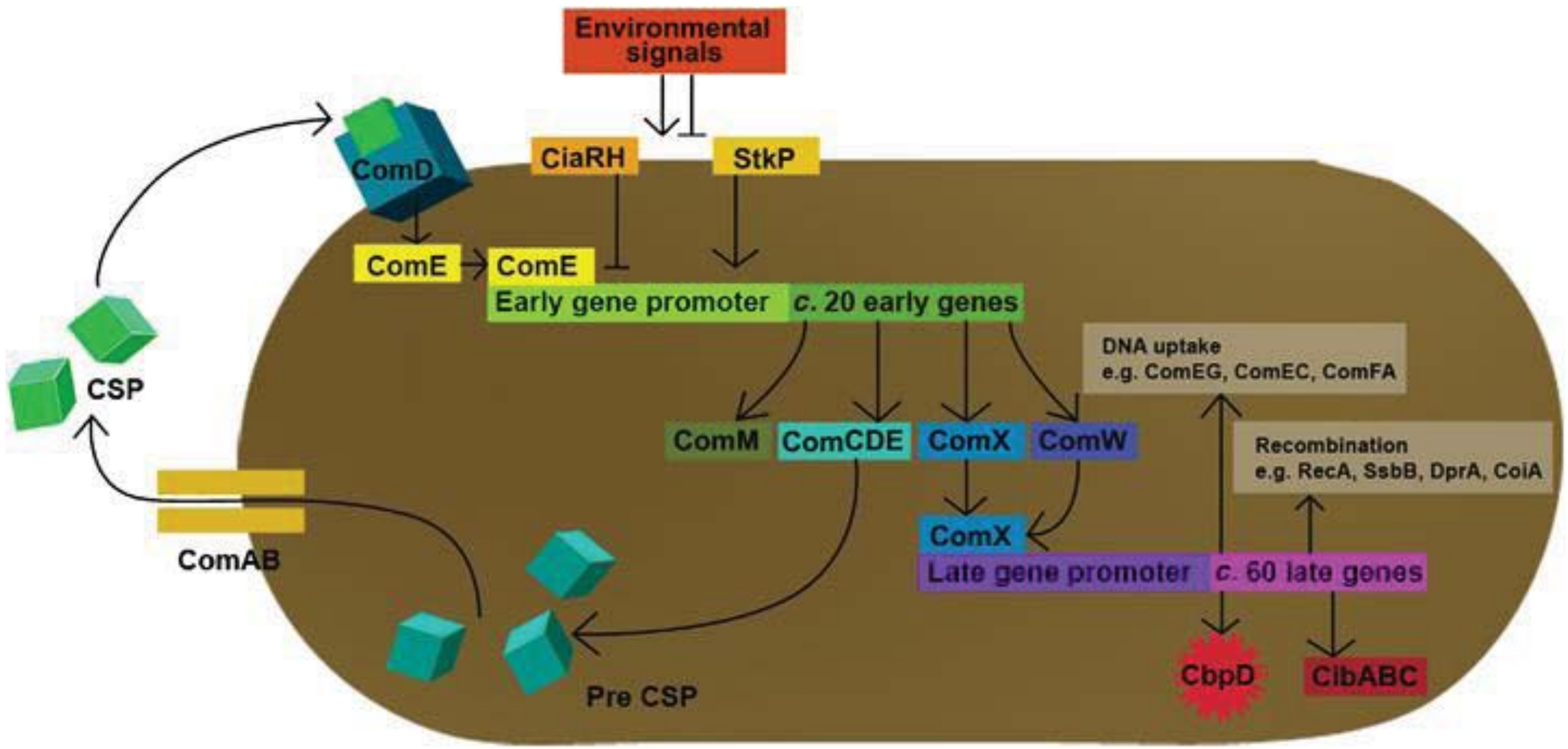


Figure 1.3 Summary of *comCDE*-dependent competence regulation (Johnsborg & Havarstein, 2009)
Schematic representation of competence regulation, as described and discussed in the Introduction and reveiwed by Johnsborg & Havarstein (2009). Unknown environmental stimuli regulate the basal expression of *comCDE*. Accumulation of CSP in the environment at a particular cell density activates the two-component system encoded by *comDE*, which triggers the expression of the early competence genes. Early competence gene expression involves autoinduction of *comCDE* expression and the regulation of late competence gene expression by ComX and ComW. Late genes, including those required for DNA uptake and recombination and those required for fratricide, are activated by ComX.

which activates a two-component system, encoded by *comD* and *comE* (Tomasz & Hotchkiss, 1964; Pestova *et al*., 1996; Havarstein *et al*., 1996). The basal expression of *comCDE* is at least partially regulated by CiaRH and StkP, which respond to unknown environmental signals and control global patterns of gene expression (Guenzi *et al*., 1994; Giammarinaro *et al*., 1999; Mascher *et al*., 2003; Halfmann *et al*., 2007). However, potential environmental stimuli might include high phosphate concentrations, bovine serum albumin, $CaCl_2$, alkaline pH and DNA-damaging agents, such as mitomycin C, which have been shown to stimulate competence induction under certain conditions (reviewed by Claverys & Harvestein [2002]; Prudhomme *et al*., 2006). Activation of the response regulator encoded by *comE* triggers the transcription of early competence genes, which include those encoding ComX and ComW, the latter of which is thought to protect ComX from proteolytic degradation (Lee & Morrison, 1999; Peterson *et al*., 2004; Luo *et al*., 2004; Sung & Morrison, 2005). ComX has been shown to activate the expression of late competence genes, which include those required for the processing and uptake of DNA from the environment and genes required for recombination of the acquired DNA into the chromosome (Peterson *et al*., 2004). In addition, genes such as *cbpD*, which are involved in the autolysis of non-competent sister cells and closely-related streptococcal species, are expressed and are thought to facilitate the acquisition of DNA from the environment (reviewed in Claverys & Harvestein [2007]). Presumably the acquisition of DNA from lysed non-competent cells is followed by selection for the optimum complement of old and newly acquired genetic material. Protection of competent cells is thought to be mediated by the expression of the early competence gene *comM*, which encodes an immunity protein (Havarstein *et al*., 2006).

However, despite what is known of competence, it is still not entirely clear how the competent state is abruptly switched 'on' or 'off' (reviewed in Johnsborg &

Harvastein [2009]). In addition, it is not clear why a large number of clinical isolates have been found to be non-transformable at least *in vitro* (Pozzi *et al*., 1996).

1.4 Treatment and prevention of pneumococcal disease and its effect on epidemiology

1.4.1 Treatment with antibiotics

Since the 1940s penicillin has been the treatment of choice for pneumococcal pneumonia. However, since the first penicillin-resistant isolate was detected in the mid 1970s, the detection of resistant isolates has rapidly increased (reviewed in Watson *et al*. [1993]). Resistance to penicillin has come about through the modification of the penicillin-binding proteins to a point where they can continue to produce peptidoglycan in the presence of the antibiotic (reviewed in van der Poll & Opal [2009]). Of the IPD isolates tested for penicillin resistance in Australia in 2006, 10.6% had reduced susceptibility to penicillin and 2.4% of isolates were completely resistant (Roche *et al*., 2008). In addition, reduced susceptibility to ceftriaxone was reported in 2.9% of IPD isolates tested in 2006 and 0.9% were completely resistant (Roche *et al*., 2008).

The prevalence of penicillin resistant pneumococci varies widely between regions. A recent study has shown that the rate of resistance was approximately 74% in South Africa, 63% in East Asia, 54% in the Middle East, 25.7% in Southern Europe, 6% in Northern Europe and 25% in Australia in isolates from pneumococcal community-acquired pneumonia (Felmingham *et al*., 2007). Resistance to antimicrobial agents, such as macrolides, fluoroquinolones, vancomycin and trimethoprim has increasingly been detected globally (reviewed in van der Poll & Opal [2009]).

The genetic transformability of the pneumococcus (Section 1.3.4) has largely been thought to mediate the spread of antibiotic resistance (Reichmann *et al*., 1997), though certain antibiotic resistance genes, such as those for chloramphenicol (cml) and tetracycline (tet), have been found associated with integrative conjugative elements (ICEs), such as Tn*916* and Tn*5252* (Croucher *et al*., 2009; Ding *et al*., 2009).

The expansion of some *S. pneumoniae* clones in terms of the global distribution and number of isolates as a proportion of the total, has been linked to the presence of antibiotic resistance, such as in the pandemic carriage Spanish^{23F} clone with the sequence type (ST) 81, which is resistant to penicillin, tet and cml, and frequently associated with fluoroquinolone resistance (reviewed in Croucher *et al*. [2009]). STs are determined following multi-locus sequence typing (MLST), which is performed to assess the genetic relatedness of the tested strains in a manner that is dependent on genetic drift over time within the genome (Section 1.5.1). Another widely distributed clone is the serotype 9V Spanish^{9V} ST156 clone, which has been associated with resistance to penicillin and trimethoprim-sulfametoxazole (Zemlickova *et al*., 2006). Interestingly, the success of the Spanish^{9V} clone has been suggested to be due to the *rlrA* locus, which encodes a pilus-like structure (Sjostrom *et al*., 2007; Barocchi *et al*., 2006). However, in this case it appears that *rlrA* has facilitated the spread of antibiotic resistance rather than selection for resistance to the antibiotic itself.

In summary, continued work is required to ensure effective antibiotics remain as treatment options for pneumococcal disease. In addition, preventative treatment strategies such as vaccination are particularly important to reduce the reliance on current antibiotics and those in development, and are also required in the parts of the world where access to effective antibiotics is limited.

1.4.2 Vaccination targeting the capsule

As discussed in Section 1.3.1.5, unencapsulated strains of pneumococci are largely avirulent. Therefore, the first generation of pneumococcal vaccines was designed to target the capsule. The first such vaccines were the 14- and then 23-valent purified polysaccharide vaccine. In particular, the 23-valent polysaccharide vaccine, Pneumovax® (PPSV23), was designed to include polysaccharide of the 23 serotypes most commonly associated with IPD (1, 2, 3, 4, 5, 6B, 7F, 8, 9N, 9V, 10A, 11A, 12F, 14, 15B, 17F, 18C, 19A, 19F, 20, 22F, 23F and 33F). However, whilst purified polysaccharide antigens can elicit some protection in adults, they are poorly immunogenic in infants, due to an inability to stimulate a T cell-dependent immune response (Mond *et al*., 1995). In addition, whilst PPSV23 affords short-term protection in the elderly against IPD, the vaccine is less effective against pneumonia, which is the primary cause of pneumococcus-associated disease in the elderly (Mangtani *et al*., 2003). In order to provide protection in infants, PCV7 was developed and has been licensed in the United States since 1999. PCV7 consists of the polysaccharide of the 7 serotypes most commonly associated with IPD (4, 6B, 9V, 14, 18C, 19F and 23F) in the United States, conjugated to a protein carrier (CRM197), which functions to promote a T cell-dependent response against the included serotypes (Ahmad & Chapni, 1999). Prior to the introduction of PCV7, the 7 serotypes included in PCV7 represented 80 – 90% of disease-causing serotypes in the United States, Canada and Australia, 70 – 75% in Europe and Africa, 65% in Latin America and 50% in Asia, which highlighted geographic variability in the distribution of serotypes (Hausdorff *et al*., 2000a). Whilst PCV7 affords good protection against IPD caused by the included serotypes in young children, protection was not much better than for PPSV23 in the elderly and so PPSV23 remains the recommended vaccine for the elderly in most countries (reviewed in Siber *et al*., 2008).

Since the introduction of PCV7 in Australia for high risk groups, such as indigenous children in 2001, and for universal vaccination in 2004, the incidence of IPD has reduced by 89.6% in children <2 years and by 82% in children 2 – 14 years, to 2006, as shown in Figure 1.4.a (Roche *et al*., 2008). Similarly in the United States the

NOTE:
This figure is included on page 29
of the print copy of the thesis held in
the University of Adelaide Library.

Figure 1.4 Notification rates of IPD caused by total, PCV7 and non-PCV7 serotypes in indigenous and non-indigenous children in Australia (Roche *et al.*, 2006)
(a) The rate of total IPD in Australia in the indicated years per 100,000 people in each age group. Universal PCV7 vaccination of children <5 years commenced in 2004.
(b) The rate of IPD in each indicated year per 100,000 in indigenous and non-indigenous children <2 years due to either PCV7 included serotypes (7vPCV) or PCV7 non-included serotypes (Non-7vPCV).

incidence of IPD decreased by 76% in children <5 between the introduction of the vaccine in 1999 and 2007 (Pilishvili *et al*., 2010). However, the incidence of IPD due to non-PCV7 serotypes has increased by 45% in the United States, with the non-PCV7 serotype, serotype 19A, currently responsible for 42% of all cases, which is a 324% increase compared to pre-vaccination (Pilishvili *et al*., 2010). An increase in IPD caused by non-PCV7 serotypes has been observed in a number of studies (Steenhoff *et al*., 2006; Hsu *et al*., 2010; Giele *et al*., 2009), which also found a significant increase in the contribution of serotype 19A as a proportion of all IPD.

In contrast, such an increase in the prevalence of non-PCV7 serotypes was not yet apparent in the most recent Australian national survey (2006), as the rate of IPD due to non-PCV7 remained stable in both indigenous and non-indigenous children (Figure 1.4.b; Roche *et al*., 2008). However, a more recent report from Western Australia detected an increase in the contribution of serotype 19A to IPD, particularly in non-indigenous children (Giele *et al*., 2009). Of concern for the long-term effectiveness of PCV7 in Australia was the significant contribution of non-PCV7 serotypes to IPD in indigenous communities (Figure 1.4.b), which was approximately 69.8 cases per 100,000 compared to 13.5 per 100,000 in non-indigenous children.

In the Finnish trials of PCV7, cases of acute OM due to PCV7-included serotypes were reduced by 51% in children that had received 4 doses of PCV7 compared to non-vaccinated children. However, the incidence of acute OM due to non-PCV7 serotypes increased by 33% in children that were vaccinated compared to those who were not vaccinated (Eskola *et al*., 2001). Following introduction of the vaccine in the United States an increase was also observed in the incidence of acute OM due to non-PCV7 serotypes, which increased by 20 % between 1999 and 2002 (McEllistrem *et al*., 2005).

In response to concerns over the increased rate of IPD and OM due to non-vaccine serotypes, 10- and 13-valent conjugate vaccines have been recently licensed in Europe and the US, which include Synflorix® (PCV10) and Prevnar13™ (PCV13). The serotypes included in PCV10 in addition to those of PCV7 are 1, 5 and 7F and the serotypes included in PCV13, in addition to those of PCV7, are 1, 3, 5, 6A, 7F and 19A. Whilst the increased valency of these vaccines could afford protection against emerging dominant serotypes, the problem of serotype specificity will probably remain a long-term concern, requiring repeated reformulation.

In addition to the problem of serotype-specificity, conjugate vaccines are expensive to manufacture (Butler *et al*., 2004). Currently the three required doses of PCV7 cost almost $300 ($A) per child, which is well above what is feasible for many developing countries, where the burden of pneumococcal disease is greatest (Section 1.1).

Given the limitations and cost of conjugate vaccines, alternatives such as protein-based vaccines are currently nearing clinical trials. Formulations include virulence factors such as PdB (Ply toxoid), PspC and PspA (Ogunniyi *et al*., 2007) and the more recently identified surface proteins PscB and StkP (Giefing *et al*., 2008). Protein-based vaccines should theoretically provide species-wide protection due to the serotype-independent nature of the selected surface proteins.

1.5 The contribution of serotype and genomic diversity to invasive disease potential

1.5.1 Multi-locus sequence typing for the genotyping of *S. pneumoniae*

MLST was developed as a serotype-independent method for comparing the genetic relatedness of different strains of *S. pneumoniae* by sequencing defined regions within seven housekeeping genes, as described in Section 2.10 (Enright & Spratt, 1998). By comparing the sequence of these genes with a central database (http://www.mlst.net), individual genes are given an allele number that is used to generate an allelic profile and determine the ST. MLST measures the amount of genetic drift in the sequenced genes in order to predict the relative time that has passed since the STs had separated from a single clone.

1.5.2 Association between serotype, genotype and invasive disease

As discussed in Section 1.3.1.5, the different structures of the polysaccharide capsule of different serotypes have been shown to at least partly contribute to variation in resistance to complement deposition and opsonophagocytic killing by neutrophils (Hyams *et al*., 2010; Weinberger *et al*., 2009). Serotypes 1, 7F 14, 18C, 9V and 4 have generally been found to be associated with high or at least medium invasive potential and are not commonly recovered from healthy carriers (Brueggemann *et al*., 2003; Hanage *et al*., 2005; Sandgren *et al*., 2004). In particular, serotypes 1 and 7F have been suggested to behave as primary pathogens, as they are readily isolated from people not suffering from an underlying illness (Sjostrom *et al*., 2006). In contrast, serotypes 3, 6A, 6B, 8, 19F and 23F are thought of as having a lower invasive potential, as they are more commonly recovered from healthy carriers (Brueggemann *et al*., 2003; Hanage *et al*., 2005; Sandgren *et al*., 2004). In addition, serotypes of low invasive potential are most commonly found to cause disease in people with underlying illness and so behave as opportunistic pathogens (Sjostrom *et al*., 2006). Interestingly, it has been suggested that the disease caused by serotypes that behave as primary pathogens tends to be less severe in both mice and humans, as defined by case fatality rate and acute physiology and chronic health evaluation II scores, compared with that of serotypes that behave as opportunitistic pathogens (Sjostrom *et al*., 2006). This may be due to the lack of underlying illness contributing to disease severity.

In addition to the role of serotype, the role of genotype in invasive potential has also been shown to be significant, and in fact the considerable overlap between genotype and serotype might actually be responsible for much of the association between serotype and invasive potential (reviewed in Henriques-Normark *et al*. [2008]). The role of genotype is highlighted by differences in invasive potential between strains of the same serotype. For example, in one study, serotype 14 ST307 was recovered from only healthy carriers, whereas serotype 14 ST230, was recovered from people suffering from disease (Sandgren *et al*., 2004). In addition, closely-related clones of serotype 6B, STs 188 and 176, have been found to be associated with IPD in Sweden, the United Kingdom and Denmark, whereas less closely-related clones of the same serotype were found to be predominantly carried (Sandgren *et al*., 2004). Similarly, strong associations with IPD have been identified for STs 482, 191, 124, 138 and 156, with ST156 associated with both serotypes 9V and 14. In contrast, STs 485 and 62 were found to be associated mainly with healthy carriers (Hanage *et al*., 2005).

The role of bacterial factors as predictors of IPD potential and associated morbidity and mortality is controversial. Alanee *et al*. (2007) identified the serotype of isolates from 796 cases of IPD in 10 countries and found no correlation between serotype and IPD, or between serotype and disease severity allowing for underlying illness. Whilst there is some merit to the author’s counterclaims of associations between underlying illness, disease severity and serotype that were claimed by Sjostrom *et al*. (2006), the lack of data collected for asymptomatic carriage by Alanee *et al*. (2007), makes the prediction of invasive potential impossible in this study. However, it is also worth noting that in the study by Sandgren *et al*. (2004), carriage isolates were taken from children at day-care centres, whereas the majority of disease isolates (257 of 273) were from hospitalised adults. Therefore, it is not possible to differentiate between the impact of serotype in predicting invasive potential compared to the impact of age on the prevalence of a particular serotype in disease. Nevertheless, the findings of Sandgren *et al*. (2004) generally correlated with that found by Brueggemann *et al*. (2003) and Hanage *et al*. (2005), where both carriage and invasive isolates were only collected from children <5 years and <2 years respectively.

Whilst there is some controversy regarding the role of bacterial factors in predicting the invasive potential of pneumococcal isolates, there is compelling evidence that differences do exist between both serotypes and genotypes in their capacity to cause disease.

1.5.2.1 Serotype one

An important serotype of the pneumococcus is serotype 1, due to its high association with IPD, whilst rarely being isolated from healthy carriers (Brueggemann *et al*., 2003; Sandgren *et al*., 2004).

Serotype 1 strains have recently been found in a number of studies to be amongst the leading causes of pneumococcal pulmonary empyema, peritonitis and severe meningitis (Hernandez-Bou *et al*., 2009; Goldbart *et al*., 2009; Lagos *et al*., 2008; Yaro *et al*., 2006) and have a higher ratio of hospitalisation versus ambulatory care when compared to other serotypes (Alpern *et al*., 2001).

Serotype 1 is ranked amongst the most commonly isolated pneumococcal serotypes in a number of continents including Europe, Asia, South America, Africa and in indigenous communities of Australia (Hausdorff *et al*., 2000a; Hausdorff *et al*., 2000b; Hanna *et al*., 2008). Despite the importance of serotype 1 disease in many parts of the world, PCV7 does not include serotype 1 (Section 1.4.2). However, serotype 1 is now included in the recently licensed PCV10 and PCV13 (CDC, 2010). The lack of inclusion of serotype 1 in PCV7 has seen an increase in serotype 1-associated disease in some jurisdictions (Byington *et al*., 2010; Sa-Leao *et al*., 2009; Kirkham *et al*., 2006).

In addition, serotype 1 strains are often associated with outbreaks of IPD (Hausdorf *et al*., 2000; Dagan *et al*., 2000; Yaro *et al*., 2006; Leimkugel *et al*., 2005; Gratten *et al*., 1993). A particularly severe outbreak of meningitis occurred between 2000 and 2003 in the Kassena-Nankana District of northern Ghana, which featured a case-fatality rate of approximately 44.4%. This was found to have been caused predominantly by serotype 1 ST217 and the single-locus variants ST303 and ST612 (Leimkugel *et al*., 2005). Clonal expansion of the contribution of ST306 to serotype 1-associated disease in Scotland has been proposed to be associated with the presence of a non-haemolytic allele of Ply (Kirkham *et al*., 2006; Jefferies *et al*., 2007).

Given the association of serotype 1 with outbreaks of IPD and very low carriage, it is not surprising that the geographic distribution of serotype 1 clones has been shown to be quite defined (Brueggemann & Spratt, 2003). Three lineages of serotype 1 clones were established based on STs sharing at least four of seven alleles (Figure 1.5). Interestingly, the lineages reflected geographic isolation, where lineage A STs were predominantly from developed regions, such as the US and Western Europe, lineage B STs were from Africa and the small number of lineage C STs were of South American origin. As discussed above, Sandgren *et al*. (2005) suggested that whilst serotype 1 strains were often associated with disease in patients without underlying illness, the disease was not severe, which was corroborated in mice. However, both Sandgren *et al*. (2005) and Sjostrom *et al*. (2006) only considered STs of lineage A, such as ST306 and

NOTE:
This figure is included on page 36
of the print copy of the thesis held in
the University of Adelaide Library.

Figure 1.5 Serotype 1 lineages as determined by Brueggemann & Spratt, 2003
Dendogram modified from Brueggemann & Spratt, 2003. A linkage distance of 0.14 or 0.29 indicates identical alleles at 6 of 7 or 5 of 7 loci, respectively. Lineage A STs were of North American and Western European origin, lineage B STs were of African origin and lineage C STs were mostly of South American origin (Brueggemann & Spratt, 2003).

ST228, whereas STs of lineage B, such as ST217, ST303 and ST618 have been associated with severe cases of IPD and have been described as hypervirulent, with case-fatality rates of almost 50% (Yaro *et al*., 2006; Leimkugel *et al*., 2005; Antonio *et al*., 2008). Therefore, whilst host-associated differences between Europe and Africa cannot be ruled out, it might be possible that isolates of lineage B are inherently more invasive than those of lineage A.

An interesting recent finding has been the increasing detection of serotype 1 carriage (Smith-Vaughan *et al*., 2009; Nunes *et al*., 2008). Nunes *et al*. (2008) detected an increase in carriage of serotype 1 from 0% in 2001 and 2002 to 0.4% in 2003 and 3.1% in 2006, which paralleled the introduction and increase in the use of PCV7 in Portugal. The carried clones were primarily ST306 and the double-locus variant ST228. In the Northern Territory of Australia, intermittent carriage of ST304 of serotype 1 was detected between 1992 and 2005, which was followed by clusters of serotype 1 carriage of ST227 in 2002 and 2003 (Smith-Vaughan *et al*., 2009). In the latter study, the detection of serotype 1 carriage was not correlated with an increase in serotype 1-associated disease, indicating that the increased carriage did not lead to increased disease. The cluster of serotype 1 carriage was observed following the introduction of PCV7 in these communities. The relationship between the introduction of PCV7 and the detection of serotype 1 carriage in both studies, suggests that serotype replacement has occurred in these regions, as discussed in Section 1.4.2.

As was discussed above, serotype 1 strains have been suggested to behave as primary pathogens due to the ability to cause disease in people without underlying illness (Sjostrom *et al*., 2006), which suggests that many serotype 1 isolates possess an inherent capacity for causing IPD. Hence, the study of serotype 1 isolates provides a good model for understanding the progression from colonisation of the nasopharynx to IPD, without the added complexity that comes with opportunistic and secondary infections. Therefore, having both non-invasive and invasive serotype 1 isolates would provide a clear basis for the identification of serotype-independent properties that might contribute to differences in invasive potential.

1.5.3 Contribution of genomic diversity to invasive potential

Genomic diversity can occur via three different ways; gene gain, gene loss and gene change (reviewed in Pallen & Wren [2007]). Such changes can range from large scale acquisition or loss of genomic islands, to allelic differences between genes and single nucleotide polymorphisms (SNPs) within both coding and non-coding sequence. Such diversity has been reported between different pneumococcal isolates (Tettelin *et al*., 2001; Bruckner *et al*., 2004; Silva *et al*., 2006; Obert *et al*., 2006; Iannelli *et al*., 2002). The ubiquitous presence of transposases, site-specific recombinases/integrases and insertion sequences, highlights the importance of horizontal genetic transfer between strains of the pneumococcus and other closely related species (Tettelin *et al*., 2001; Hoskins *et al*., 2001). An important driver of genetic diversity in the pneumococcus is its capacity to undergo natural transformation (Section 1.3.4). Furthermore, ICEs and bacteriophages have also been found to be important drivers of horizontal genetic exchange (Croucher *et al*., 2009; Romero *et al*., 2009a; Romero *et al*, 2009b). Rapid emergence of uniformity has been suggested to occur amongst isolates associated with higher invasive potential (Pallen & Wren, 2007), which is supported by evidence in the pneumococcus that highlights greater clonal relatedness amongst serotypes associated with a high invasive potential, such as serotypes 1 and 7F (Dagerhamn *et al*., 2008; Sandgren *et al*., 2005; Sandgren *et al*., 2004). This phenomenon is particularly pronounced during outbreaks where large groups of people can be infected by single or closely-related clones, such as STs 217, 303 and 618 in meningitis outbreaks in Ghana (Leimkugel *et al*., 2005).

Genome projects undertaken to date have shown that the pneumococcal genome can be divided into core- and non-core (accessory) regions (Tettelin *et al*., 2001; Bruckner, *et al*., 2004; Silva *et al*., 2006; Obert *et al*., 2006). The core genome contains regions that are presumably required for the viability of an individual cell and are present in all isolates of *S. pneumoniae*. By contrast, the accessory regions (ARs) appear to form a large hypothetical pool of genetic potential, which can provide genetic variability and adaptability. It has been predicted that approximately 50% of the pneumococcal supragenome represents the core genome, which suggests that the equivalent of half a genome worth of AR sequence exists, and is similar to other naturally transformable species such as *H. influenzae* (Hiller *et al*., 2007). To date 41 ARs have been identified, using the TIGR4, R6 and G54 genomes as a reference (Blomberg *et al*., 2009). The putative association of some of these regions with virulence will be discussed in Section 1.5.3.2.

1.5.3.1 Small-scale genetic variability in the pneumococcus

Whilst the focus of the work of this thesis is on the genetic variability of relatively large regions of the genome, allelic variation of individual virulence factors and SNPs play a role in the genetic variability of the pneumococcus (reviewed in Mitchell & Mitchell [2010]).

Allelic variation has been observed in a number of virulence factors, which includes PspC, NanA, PspA and Ply (Iannelli *et al*., 2002; King *et al*., 2005; Hollingshead *et al*., 2000; Jefferies *et al*., 2007; Kirkham *et al*., 2006; Lock *et al*., 1996). For surface proteins, such as PspC, NanA and PspA, variation has been proposed to be driven by the need to avoid immunological detection (Iannelli *et al*., 2002; King *et al*., 2005; Hollingshead *et al*., 2000). The purpose for allelic variation in Ply is not clear. However, differences in the inflammatory response to non-haemolytic Ply (Littmann *et* *al*., 2009), may indicate a potential Ply-mediated mechanism for differential modulation of the immune system.

1.5.3.2 Accessory regions of the pneumococcal genome associated with invasive potential

As mentioned above, the most recent report has identified 41 ARs, using the genomes of TIGR4, R6 and G54 as references (Blomberg *et al*., 2009). Of the 41 ARs identified, 24 contained genes previously identified by signature-tagged mutagenesis as being associated with virulence (Hava *et al*., 2002; Lau *et al*., 2001; Polissi *et al*., 1998). A number of ARs have also been identified with links to increased invasive potential (Silva *et al*., 2006, Obert *et al*., 2006; Blomberg *et al*., 2009). ARs associated with virulence have been found to encode genes of a variety of putative functions. A particularly important AR encodes the capsule locus, the contents of which vary between serotypes and has been studied extensively (Coffey *et al*., 1998; Nesin *et al*., 1998). Horizontal genetic transfer between the capsule loci of different serotypes has been shown to result in serotype-switching, which can promote immune evasion, as described in Sections 1.3.1.5 and 1.4.2. In addition, an AR encoding the PPI-1 has been found to be associated with the *pia* locus (Section 1.3.1.3), and the PezAT chromosomal toxin-antitoxin (TA) system (encoded by *pezAT*), which is required for full systemic virulence in TIGR4 (Brown *et al*., 2001; Brown *et al*., 2004).

A number of ARs have been suggested to be important for adherence to human epithelial cells and for full virulence in mice, such as the pilus encoded by the *rlrA* islet (Barocchi *et al*., 2006) and PsrP (Obert *et al*., 2006). The pilus-like structure encoded by the *rlrA* islet has been shown to be involved in adherence to the lung epithelium, nasopharyngeal colonisation and is a stimulator of inflammation *in vivo* (Barocchi *et al*., 2006). However, the *rlrA* islet appears to be a clonal property and has not been shown to be preferentially associated with invasive isolates (Basset *et al*., 2007). Interestingly, *rlrA* is highly represented amongst some PCV7-included serotypes (4, 6B, 9V 14) and is also strongly associated with pneumococci non-susceptible to penicillin (Sjostrom *et al*., 2007; Basset *et al*., 2007; Moschioni *et al*., 2008; Aguiar *et al*., 2008). PsrP has been shown to contribute to adherence to lung epithelium by binding to keratin on lung cells (Shivshankar *et al*., 2009), and has been claimed to be present in a large proportion of cases of IPD caused by some serotypes, including 89% of IPD due to serotype 1, 88% caused by serotype 15B/C, 85% caused by serotype 14, 79% caused by 10A and 72% caused by 19A in the United States (Orihuela, 2009). However, such claims were dependent on *psrP* being present in all isolates of a particular cluster of clones and there have been conflicting reports of the importance of the gene to virulence (Obert *et al*., 2006; Blomberg *et al*., 2009).

Four large zinc metalloproteinases have been described for the pneumococcus of which *zmpC* and *zmpD* are found to vary in their presence between strains (Chiavolini *et al*., 2003). ZmpC has been shown to cleave human matrix metalloproteinase 9, which might contribute to the development of invasive disease (Oggioni *et al*., 2003; Camilli *et al*., 2006). However, the function of ZmpD is currently unknown. In addition, ARs encoding genes, such as bacteriocins with roles in intra- and interspecies competition have also been identified (Section 1.3.1.4), as have ARs associated with putative functions in the transport and degradation of sugars, such as NanC and various PTS systems (Pettigrew *et al*., 2006; Obert *et al*., 2006). Furthermore, a significant number of ARs encode proteins of unknown function.

As discussed above, ARs have generally been reported using the genomes of TIGR4, G54 and R6 as references in comparative genomic hybridisation (CGH). However, given that Hiller *et al*. (2007) predicted that the sequence of approximately 33 genomes would be required to reflect the entire pneumococcal supragenome, many more ARs are likely to await discovery.

In addition, by concentrating on a small group of relatively closely-related strains of the same serotype with well-characterised virulence profiles in both humans and mice, true genetic predictors might be identified that could be extrapolated to include proteins of similar function across different isolates of the pneumococcus.

1.6 Preliminary studies of invasive and non-invasive serotype 1 pneumococci

As discussed in Section 1.5.2.1, asymptomatic carriage of serotype 1 strains of STs 304 and 227 was detected in a number of indigenous communities of the Northern Territory of Australia (Smith-Vaughan *et al*., 2009). In previous work (Harvey, 2006), four such non-invasive isolates were obtained from the Menzie's School of Health Research (MSHR) in Darwin, comprising nasopharyngeal carriage isolates (strains 1 & 4) and acute OM isolates (strains 2 & 3) that were not associated with IPD.

In addition, serotype 1 blood-derived isolates of indigenous origin were obtained from the MSHR, including strains 1662, 2977, 3415 and 5482. Similarly blood-derived serotype 1 isolates of non-indigenous origin were obtained from the Women's and Children's Hospital (WCH) in Adelaide, including strains 4496 and 5173.

1.6.1 Characterisation of serotype 1 isolates in murine models of infection

Initially the virulence of the clinical isolates was compared in mice, using an intraperitoneal (i.p.) model of infection, which found that one strain of indigenous origin (1861) and one strain of non-indigenous origin (4496) were significantly more virulent than the next most virulent strain, strain 3415. This indicated that strains 1861 and 4496 were highly virulent and as such strain 3415 was designated intermediately virulent (Figure 1.6). Strain 5482 was also chosen as an intermediately virulent strain

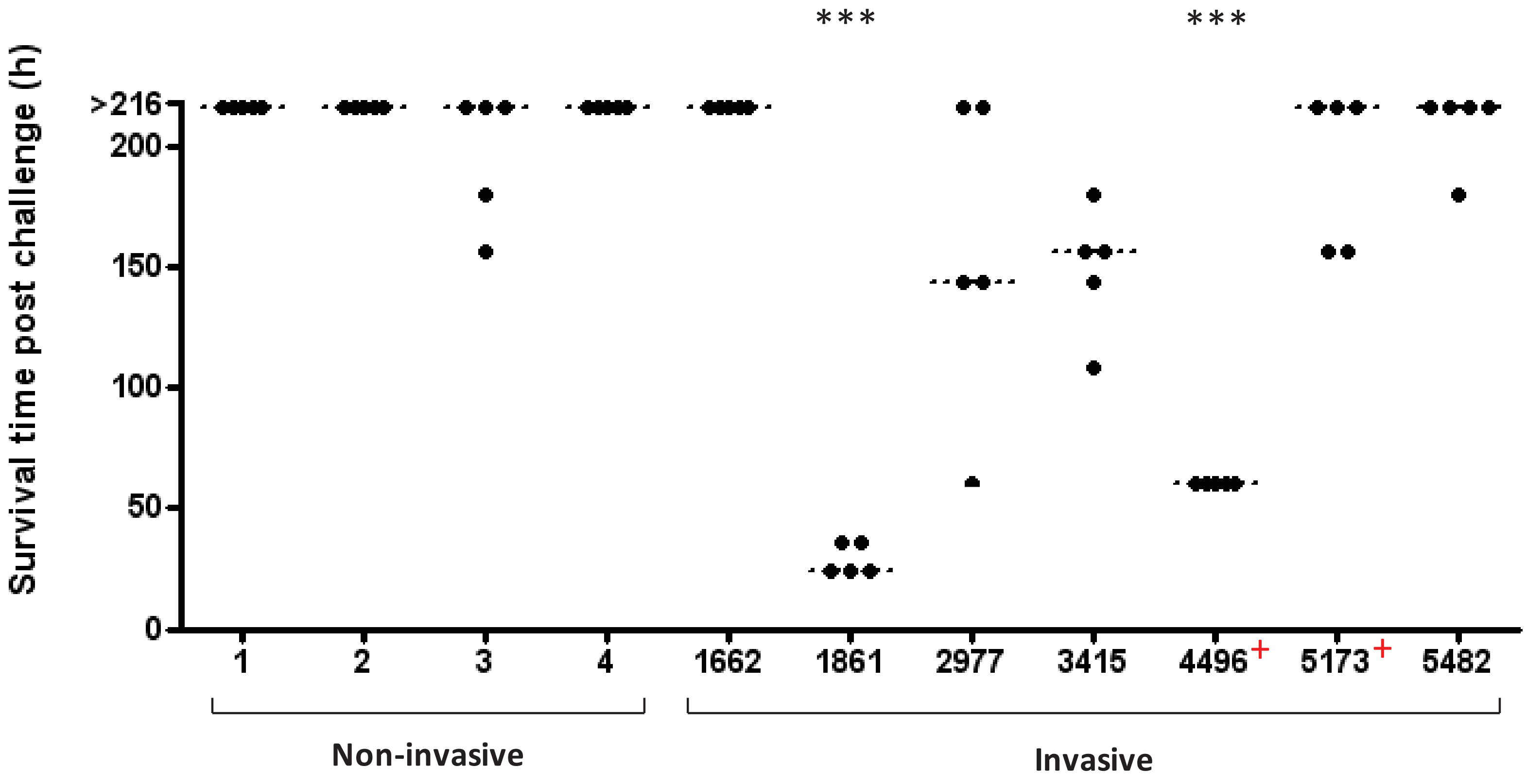


Figure 1.6 Intraperitoneal challenge with serotype 1 non-invasive and invasive isolates of indigenous and non-indigenous origin
In previous work groups of 5 Balb/c mice were challenged with ~10^4 CFU of the relevant opaque-phase isolate via the i.p. route. Mice were monitored regularly for 216 h for signs of illness and their survival times were recorded. Median survival times for each isolate are indicated with a broken line and differences between median survival times between strain 3415 and strains 1861 and 4496 was determined using a two-tailed *t*-test (*, *P*<0.05; **, *P*<0.01; ***, *P*<0.001). Non-indigenous isolates are indicated by +.

due to its status as an invasive isolate in humans. In addition, strains 1 and 4 were selected to represent the non-invasive isolates as they represented both ST304 and ST227 and were of nasopharyngeal origin. However, due to the significantly retarded growth of strain 4, strain 2 was included in some of this work of this study. An alternative intranasal (i.n.) virulence model (Section 2.13.2) also confirmed that strains 1861 and 4496 were significantly more virulent than strains 1 and 4. In addition, regular assessment of bacteraemia detected pneumococci only in the blood of mice challenged with strains 1861 or 4496, which indicated that strains 1 and 4 were unable to invade the blood in this model of infection (data not shown). Therefore, it was clear that, as in the human situation, strains 1 and 4 were unable to enter the blood, which was in contrast to the high level of bacteraemia achieved by both strains 1861 and 4496.

1.6.2 Examination of known virulence factors and relationship to invasive potential

As discussed in Section 1.3.1.5, the amount of capsule has been shown to influence resistance to opsonphagocytosis. Therefore, the amount of capsule produced was compared using a uronic acid assay between the different serotype 1 isolates (1, 4, 3415, 5482, 1861 and 4496). However, no significant differences were detected and as a result, the capsule was considered to be unlikely to be responsible for the differences in invasive potential observed between these serotype 1 isolates (Harvey, 2006).

Subsequently, Western blot analysis was used to assess the relative expression and apparent molecular weight of the virulence factors NanA, CbpA, LytA, PsaA, PspA, PiaA and Ply (Section 1.3.1). However, whilst some differences were observed, there were no patterns that were consistent with virulence profile. Interestingly, Ply of strain 4496 was found to exhibit retarded electrophoretic mobility. Subsequently, specific haemolytic activity was found to be significantly reduced and sequencing revealed nucleotide changes in *ply* of 4496 that were similar to those found previously

(Lock *et al*., 1996; Kirkham *et al*., 2006). It was interesting that despite the lack of significant haemolytic activity, strain 4496 was highly virulent in mice.

1.6.3 Chromosomal toxin-antitoxin system PezAT was found only in strains 1861 and 4496

Following the analysis described in Section 1.6.2, preliminary CGH was carried out between the highly virulent strains (1861 & 4496) and the non-invasive strains (1 & 4) in order to identify a genetic basis for the differences observed in invasive potential in both humans and mice.

Of interest though, was a small locus of genes encoding *pezAT*, which was identified in only the highly virulent strains, had previously been associated with virulence and is located within a variable region of PPI-1 (Brown *et al*., 2004). Closer analysis of the PPI-1 variable region and surrounding sequence from the CGH data suggested that both the highly virulent and non-invasive strains possessed ORFs homologous to TIGR4 *SP_1041 – SP_1046* at the 5' end and *SP_1067* at the 3' end of PPI-1. However, only strains 1861 and 4496 possessed ORFs *SP_1050 – SP_1053*, which encodes *pezAT*. In addition, Southern hybridisation analysis revealed that the intermediately virulent strains lacked *pezAT*. Subsequent PCR analysis of the entire PPI-1 variable region in the tested strains suggested that in addition to *pezAT* the overall organisation of the region encompassing PPI-1 was different between the *pezAT*-positive and *pezAT*-negative strains.

Given the association of *pezAT* and the organisation of the PPI-1 variable region with heightened virulence in these serotype 1 strains, it was considered important to characterise the sequence of the region in both groups of strains in order to determine its contribution to virulence. This will form a large part of the present study as detailed in Section 1.7.

1.7 Aims

As discussed in Section 1.5.2, it is not clear how differences in the genomic content of pneumococcal strains influence invasive potential in a serotype-independent manner. In particular serotype 1 isolates have historically been found to have a high invasive potential with only brief carriage preceding disease (Section 1.5.2.1). However, since the introduction of PCV7, an increase in the detection of serotype 1 isolates has occurred worldwide and in particular, asymptomatic carriage has been detected in some regions (Section 1.5.2.1). Therefore, preliminary genomic comparisons between non-invasive and invasive serotype 1 isolates were performed in order to identify a genetic basis for differences in invasive potential. As a result, an association between the organisation of the PPI-1 variable region, and virulence was identified (Section 1.6.3). Therefore, the aims of this project were to characterise the role of the PPI-1 variable region in virulence and identify additional regions of variation between the genomes of serotype 1 isolates of differing invasive potential. The specific objectives of this study are as follows;

1. Sequence and annotate the PPI-1 variable region in non-invasive, intermediately virulent and highly virulent serotype 1 clinical isolates.
2. Compare the PPI-1 variable region of the serotype 1 isolates of this study with isolates of other serotypes from a collection of laboratory strains and publicly available genomes.
3. Determine the transcriptomic structure of the PPI-1 variable region, by identifying co-transcripted genes.
4. Compare the *in vivo* expression patterns of PPI-1 variable region genes of strains 1861 and 4496 between niches in a mouse model.

5. Assess the effect of mutagenesis of the PPI-1 variable region on virulence in a mouse model of infection.
6. Identify additional genomic differences between non-invasive, intermediately virulent and highly virulent serotype 1 clinical isolates by CGH.
7. Identify genomic differences between the highly virulent and non-invasive strains by sequencing the genomes of strain 1 and strain 1861, followed by verification by PCR in other non-invasive, intermediately virulent and highly virulent strains.
8. Assess the niche-sensitive expression of key newly-identified genes from the genomic comparisons of aim 7.

Chapter 2 - Materials and Methods

2.1 Bacterial strains and plasmids

The bacterial strains and plasmids used in this study are listed in Table 2.1.

Table 2.1 Strains and plasmids used in the work of this thesis

Strain	Description	Reference/Source
D39	Serotype 2	Avery *et al*., (1944)
TIGR4	Serotype 4	Tim Mitchell, University of Glasgow, Scotland (Aaberge *et al*., 1981)
1	Serotype 1 (ST304), nasopharyngeal isolate, avirulent in both i.p. and i.n. mouse models	Amanda Leach, MSHR
2	Serotype 1 (ST304), otitis media, avirulent in both i.p. and i.n. mouse models	MSHR
4	Serotype 1 (ST227), nasopharyngeal isolate, avirulent in both i.p. and i.n. mouse models	MSHR
3415	Serotype 1, blood isolate, virulent in i.p. mouse model	MSHR
5482	Serotype 1, blood isolate, virulent in i.p. mouse model	MSHR
1861	Serotype 1, blood isolate, highly virulent in both i.p. and i.n. mouse models	MSHR
4496	Serotype 1, blood isolate, highly virulent in both i.p. and i.n. mouse models	Andrew Lawrence, WCH
63	Serotype 18	Laboratory collection[b]
94	Serotype 18C	Laboratory collection[b]
160	Serotype 23F, ST81 Spanish^{23F} pandemic carriage clone, resistant to Penicillin, cml & tet	Tracey Coffey, University of Sussex, UK
WCH211	Serotype 11 (ST3020)[^], sinusitis isolate	WCH
3773	Serotype 15B (ST199)[a], trachea isolate	Michael Watson, Path West, Western Australia
WU2	Serotype 3 (ST378)[^]	David Briles, University of Alabama, USA (Briles *et al*., 1981)
4104	Serotype 19A (ST199)	Path West
G54	Serotype 19F isolate	Laboratory collection[b]
3518	Serotype 11A (ST62)[a]	Path West
2663	Serotype 11A (ST 3019)[a]	Path West
MSHR5	11 (ST62)[^] nasopharyngeal isolate	MSHR
WCH43	Serotype 4 (ST205)[a] blood isolate	WCH
WCH16	Serotype 6A	WCH

WCH206	Serotype 3 (ST180)[^] otitis media isolate	WCH
73	Serotype 5	Laboratory collection[*b*]
49	Serotype 5	Laboratory collection[*b*]
171	Serotype 19A	Laboratory collection[*b*]
71	Serotype 5	Laboratory collection[*b*]
141	Serotype 16	Laboratory collection[*b*]
MSHR17	Serotype 3 (ST458)[^] otitis media isolate	MSHR
MSHR1	11A (ST3021)[^] nasopharyngeal isolate	MSHR
EF3030	Serotype 19F	University of Alabama
164	Serotype 7C	Laboratory collection[*b*]
140	Serotype 16	Laboratory collection[*b*]
153	Serotype 9V	Laboratory collection[*b*]
163	Serotype 35F	Laboratory collection[*b*]
67	Serotype 23	Laboratory collection[*b*]
D39:BS:cΔt	Tet^R, Spe^R D39 mutant with *cpsC* replaced with *tetM* (tet^R) and *aad9* (spe^R) with *cps* promoter inserted into *bgaA*	Byrne, J. (unpublished)
Rx1 –ply promoter	Rx1 Pneumolysin promoter sequence replaced with cml^R	Berry, A.M., (unpublished)
D39ΔPezT	Erm^R partial PPI-1 variable region deletion mutant (nt 942,891 - 962,708 replaced with erm^R)[*]	This study
D39^{1861}	Spe^R D39 PPI-1 variable region replacement mutant with PPI-1 variable region of 1861 (replaced D39 nt 945,320 - 962,708 [#])[*]	This study
D39ΔPPI-1	Cml^R D39 PPI-1 variable region deletion mutant (nt 939,867 - 962,708 replaced with cml^R)[*]	This study
D39^1	Spe^R D39 PPI-1 variable region replacement mutant with PPI-1 variable region of strain 1 (replaced D39 nt 939,867 - 962,708)[*]	This study
DP1617	Erm^R, Nov^R, Strep^R derivative of Rx1	Shoemaker *et al*., 1979
pVA891	Erm^R, Cml^R, *S. pneumoniae* and *E. coli* OriR	Macrina *et al*., 1983

* D39 genome sequence, Genbank Accession number CP1000410
# Indicates where coordinates are approximate
^ MLST on these strains was carried out by L.J. McAllister
a. MLST on these strains was carried out by R.M. Harvey in work not included in this thesis
b. Laboratory collection of J.C. Paton, University of Adelaide, Adelaide, Australia
MSHR – Menzie's School of Health Research, Darwin, Australia
WCH – Women's and Children's Hospital, North Adelaide, Australia
Abbreviations: Chloramphenicol (cml), Tetracycline (tet), Spectinomycin (spe), Erythromycin (erm), Novobiocin (nov), Streptomycin (strep), Origin of replication (OriR)

2.2 Growth Media

S. pneumoniae strains were grown in THY (Todd-Hewitt broth [Oxoid, Hampshire, England] with 1% Bacto yeast extract), SB (Serum broth [10% (v/v] donor horse serum in nutrient broth [10 g/l peptone (Oxoid), 10 g/l Lab Lemco powder

(Oxoid) and 5 g/l NaCl]), BHI (Brain Heart Infusion broth [Oxoid], TSB (Tryptic soy broth [Bacto]) or on blood agar (BA) plates (39 g/l Columbia base agar [Oxoid], 5% [v/v] defibrinated horse blood).

Overnight (O/N) cultures of *S. pneumoniae* were grown in either THY or BHI + 0.5% choline chloride (w/v) at 37°C for >16 h. Plates were also grown + 5% CO_2 for >16 h, but in 95% air/5% CO_2.

To identify colonies of opaque or transparent opacity phenotypes (Section 1.3.2), *S. pneumoniae* were grown on THY plates (THY broth with 1.5% Bacto agar) supplemented with 6300 U of catalase (Roche Diagnostics, Mannheim, Germany) per plate. After incubation at 37°C in 95% air/5% CO_2 for 24 h, the colony phenotype was determined using oblique transmitted light, as described by Weiser *et al*. (1994).

For storage, *S. pneumoniae* from an appropriate O/N BA plate were grown in SB for approximately 1-3 h or were resuspended in THY + 15% (v/v) glycerol and stored at -80°C.

2.3 Chemicals, reagents and enzymes

Most chemicals used were AnalaR grade and were purchased from Ajax Chemicals (NSW, Australia). Tris was purchased from Progen Industries (QLD, Australia). Sodium dodecyl sulphate (SDS) was purchased from Sigma Chemical Company (MD, USA). Deoxyribonucleoside triphosphates (dNTPs) were purchased from Roche Diagnostics Australia (NSW, Australia). Sodium deoxycholate (DOC) was purchased from BDH Biochemicals (Poole, England). Enzymes were purchased from Roche or New England Biolabs (NEB) (distributed in Australia by Genesearch, QLD, Australia), where indicated.

2.3.1 Antibiotics

Erm and cml were purchased from Roche and gentamycin (gen) and spe were purchased from Sigma.

2.3.2 Oligodeoxynucleotides

The oligodeoxynucleotides (primers) used in this study were purchased from Sigma Genosys (NSW, Australia). Primers used for real-time RT-PCR were designed using the online primer designing software, OligoPerfect^{TM} (Invitrogen, http:///www.invitrogen.com). The desired primer parameters used when designing primers for real-time RT-PCR are shown in Table 2.2.

Table 2.2 Parameters for design of real-time RT-PCR primers using OligoPerfect^{TM}

Primer Sizes (bases)			Product size (bases)	
Minimum	**Optimum**	**Maximum**	**Minimum**	**Maximum**
18	20	27	140	160
Primer melt temperature (°C)			**Experimental conditions**	
Minimum	**Optimum**	**Maximum**	**[Salt]**	**[Primer]**
30	50	80	50 mM	200 nM
Primer % GC				
Minimum	**Optimum**	**Maximum**		
30	50	80		

Primers used in this study are listed in Table 2.3.

Table 2.3 Primers used in this study

Code$^{\#}$	Primer name	Primer sequence	Primer coordinates
a	RH1045F	5’- GTC CTA GCT CAG CGA GTA GAA G -3	939,290 – 939,311^{a} 3,771 – 3,792$^{b\ \&\ c}$
aa	RHx5*	5’- CCG TTG CCT TTG CCC ACC TG -3’	8,728 – 8,747^{c}
ab	RHy4	5’- CCG TTA CCA TTA ATA GCA TTC TGG -3’	7,131 – 7,154^{c}
ac	RH1051R*	5’- CTA CCT GAC TCC ACT CTC C -3’	943,595 – 943,580^{a} 8,056 – 8,074^{b}
ad	RH1052F	5’- GTT TTC TTA TGT TAG CAG AGG C -3’	943,956 – 943,977^{a} 8,435 – 8,456^{b}
af	RH1042F	5’- GGA AAG AGA TTT CAA ATT TAT CCC -3’	935,778 – 935,801^{a} 257 – 280$^{b\ \&\ c}$
ag	RH1045eryR*$^{\#}$	5’- ttg ttc atg taa tca ctC CTT CTC GCC ACC AGA AGT AGG -3’	938,501 – 938,480^{a} 2,961 – 2,983^{b} 2,960 – 2,982^{c}

aj	RHy5	5’- CTA ACC AAG TTG TTG GGG GAG -3’	7,708 – 7,728[c]
ak	RHx6*	5’- GAA ATG CGA CAA CCT CAC CC -3’	8,132 – 8,151[c]
al	RH1050R*	5’- CAT CAT CAA TGC TAA TTC CCT G -3’	942,732 – 942,711[a] 7,190 – 7,210[b]
am	RH1053for	5’- CTA AGC CTG AAT ATC GTG ATG TAG -3’	944,955 – 944,978[a] 9,435 – 9,458[b]
an	RH1042F$_{(2)}$	5’- GAG GTG ATG AAA ATG GCA ACC -3’	936,410 – 936,430 890 – 910[b] 889 – 909[c]
ao	RH1041R*	5’- GGG ATA AAT TTC AAA TCT CTT TCC -3’	935,801 – 935,778[a] 257 – 280[b & c]
ap	RH1048R*	5’- CCG CTC TGT AGC CAT TAA TC -3’	941,868 – 941,849[a] 6,328 – 6,347[b]
aq	RH1053R*	5’- GTC ACG ATT CTG TTT GTA GAA CC -3’	990,643 – 990,665[a] 9,488 – 9,510[b]
ar	RH1044R*	5’- CAG CAT TCG GAT ACA TTC TGC C -3’	945,030 – 945,008[a] 2,157 – 2,177[b] 2,156 – 2,176[c]
as	RH1045F$_{(2)}$	5’- CCT ACT TCT GGT GGC GAG AAG GC -3’	938,480 – 938,502[a] 2,961 – 2,983[b] 2,960 – 2,982[c]
at	RH1053F$_{(2)}$	5’- GTC AGA TGG AGT TAA CGG ATG G -3’	944691 – 944712[a] 9,171 – 9,192[b]
au	RH1053R$_{(2)}$*	5’- CTT GAA GGA ATC TTT TCC TCC CTC -3’	944,796 – 944,773[a] 9,253 – 9,276[b]
av	RHz1*	5’- GGC GCT CTA AAT TCT CTA TAT CAG -3’	21,279 – 21,302[b]
aw	RHz2*	5’- GAT TCC TCT TAT CGT CAT ATT CTC -3’	20,638 – 20,661[b]
ax	RHw1	5’- GGA CGG TAC ATG CAG TGG TG -3’	9,815 – 9,834[b]
ay	RH1045gap	5’- GCG TAT TAA TCT CCA GTA TGT C -3’	2,814 – 2,835[b] 2,813 – 2,834[c]
b	RH1066R*	5’- CCA TGA GGA GAT CGT CTA GGC TT -3’	1,000,963 – 1,000,985[a] 12,769 – 12,791[c]
ba	RHw2	5’- CTC AGT CTA CGC CAG AAA AG -3’	10,603 – 10,622[b]
bb	RHz3*	5’- CAA CCA CTG ATA GAG ATG TTG G -3’	20,164 – 20,185[b]
bc	RHw3	5’- GTT GTA CGA GTC CAG AAA TGA C -3’	11,045 – 11,066[b]
bd	RHz4*	5’- GTC CTT ACA AGA TAA TAT CCA CTA G -3’	19,456 – 19,480[b]
be	RHz5	5’- GAT AAT CAA AAG ATA AGT TTG GGT C -3’	11,775 – 11,799[b]
bf	RHw4*	5’- GCA TGT TAA TCC CAA TCC TTT C -3’	18,786 – 18,807[b]
bg	RHz6	5’- GAG CTT GTA TTA GAA ATT GCA GAG -3’	12,384 – 12,407[b]
bh	RHw5*	5’- CCT CTT AGG CAC CAA CTT GG -3’	18,019 – 18,038[b]
bi	RHz7	5’- CCC CAT TTT TCA ACT TTA AAG AG -3’	13,177 – 13,199[b]
bj	RHw6*	5’- GAT TTG AAA CAT ATG ATA CAA CCG G -3’	17,337 – 17,361[b]
bk	RHz8*	5’- CAC TGT GTG CCC TTT ATA ATT TCC -3’	16,696 – 16,719[b]
bl	RHw7	5’- GGG TAT GGT ATA ATG ACT CGG G -3’	13,900 – 13,921[b]
bm	RHz9*	5’- GAA CCA ATC ACC TCT CAA AAA ACC -3’	16,033 – 16,056[b]
bn	RHw8	5’- CGT ATG TAA TGA TGG TTG GAT GAG -3’	14,499 – 14,522[b]
bo	RHw9	5’- GTA ATC CAA GAA TCA AGG ATG GG -3’	14,844 – 14,866[b]
bp	RHz10*	5’- GAG ATG AGG GAT TTA GTA TCT TTG -3’	15,374 – 15,397[b]
bq	RHz11	5’- CCT GTA GCT TTC GAG AGT AG -3’	15,598 – 15,617[b]
br	1067F$_{(2)}$	5’- CCA TTC TGG CGC TCG TAT -3’	964,511 – 964,528[a] 23,672 – 23,689[b] 13,686 – 13,704[c]
c	RH1067R*	5’- CGA ATA GAC CAC AAT CAA GCC C -3’	963,996 – 963,975[a] 23,136 – 23,157[b] 13,151 – 13,172[c]
cb	RH1066gapF	5’- CTG GAC ATC ATT AAT CCC TTC C -3’	21,688 – 21,709[b]
ce	RH1067gapF	5’- CGA AGG ATT GTA AGG TAA GAG -3’	963,786 – 963,806[a] 22,947 – 22,967[b]

			12,961 – 12,981[c]
cd	RH1(5)R*	5’- CCC CTG TAA ATC ACC ACC AAA G -3’	4,519 – 4,498[c]
ce	RH1(8)F	5’- CTA AGT GTG ATG CCA AGT TTT G -3’	13,588 – 13,609[b] 6,105 – 6,126[c]
cf	RH1(10)F	5’- GTA ATG AAA GGT GGT AAG ATT GTT G -3’	8,925 – 8,949[c]
cg	RH1(11)F	5’- GTC GAA ATC TGG AAG AAG CCT AC -3’	9,910 – 9,932[c]
ch	RH1(12)F	5’- GCT TTT AGA GTG ACT GAA AGT C -3’	10,576 – 10,597[c]
ci	RH7F	5’- GGA ATG CTA AGC CGG ATA CAC -3’	6,186 – 6,206[b]
cj	RHrt18F	5’- GAT AAT ACA CGA CAA TTG GAA GAA C -3’	7,924 – 7,948[b]
ck	RH11F	5’- CTA ACC GCA GAT GAG CTA AAA C -3’	10,173 – 10,194[b]
cl	RH11R*	5’- GCA CAT CTC GAC TGA TCT GTT C -3’	10,347 – 10,326[b]
cm	RHrt12R*	5’- CGA CCT CCT GTC AAT TCC ATC -3’	11,197– 11,177[b]
cn	RHrt13R*	5’- CTC TTG AAG ACT TTT TAA AGT TGG -3’	12,314 – 12,291[b]
co	RHrt5F	5’- CCG GTT GCA TAG AGA AAA ACT G -3’	12,903 – 12,924[b]
cp	RHrt6R*	5’- GCG GAT TGT CAC TCA GCT AGC -3’	13,055 – 13,035[b]
cq	RHrt17R*	5’- GCC CTA TTA ATT TTA TAA TTA CAA TTC -3’	15,746 – 15,720[b]
cr	RHrt15F	5’- GGA AAT ACC TCT AAT TTC AAT ATA G -3’	16,570 – 15,594[b]
cs	RHrt16F	5’- CAA TCA GTA GAA TCA TTC TAT CTT AC -3’	17,208 – 17,233[b]
ct	RHrt14F	5’- CAG ATA TCC AAG AAG AGT AGA AAT C -3’	17,890 – 17,914[b]
cu	RHrt1R*	5’- CAC CAC CAA AGA AAT CAT CGC -3’	4,507 – 4,487[b]
cv	RH1048F	5’- GAT TAA TGG CTA CAG AGC GG -3’	6,328 – 6,347[b]
cw	RHrt3R*	5’- GGA TTG AGC ATT TTA TCT TCT CC -3’	7,098 – 7,076[b]
cx	RHrt20R*	5’- GTG TGG ATG CTG AGA ACG AAA AC -3’	7,590 – 7,568[b]
cy	RHrt21R*	5’- GTT TTA GCT CAT CTG CGG TTA G -3’	10,194 – 10,173[b]
cz	RHrt22F	5’- GCT CAG CCA CTC AGT CGT G -3’	10,449 – 10,467[b]
d	RH0972R*	5’- GGT CAA ATC GTG ATC TTC CCT C -3’	962,828 – 962,807[a] 21,964 – 21,985[b] 11,984 – 12,005[c]
da	RHrt4R*	5’- CAA TCT CTA CTT TCG GAT TCC TC -3’	13,331 – 13,309[b]
db	RHrt8R*	5’- GAT TAT AAA ATC AAC CTG ATT TAT AGC -3’	14,061 – 14,035[b]
dc	RHrt23R*	5’- GGA TTA CCT CTT GCC AAG GAA C -3’	14,850 – 14,829[b]
dd	RHrt24F	5’- GAG ATG AGG GAT TTA GTA TCT TTG -3’	15,374 – 15,398[b]
de	RHrt15F	5’- GGA AAT ACC TCT AAT TTC AAT ATA G -3’	16,570 – 16,594[b]
df	RHrt16F	5’- CAA TCA GTA GAA TCA TTC TAT CTT AC -3’	17,208 – 17,233[b]
dg	RHrt25F	5’- GTG CCA AGA TAC GTG GAA ATG G -3’	18,726 – 18,747[b]
dh	RHrt10F	5’- GAG GAA TAC ATT GAT ATC GAT AAC -3’	19,341 – 19,364[b]
di	RHrt11F	5’- GTA GGA TGA TAT CTC GAA TCG G -3’	20,093 – 20,114[b]
dj	RH1(9)R*	5’- GTA GTA TGG ACA AGC AGC AAC -3’	6,952 – 6,973[b]
dk	RH6F^	5’- ATA CAG AGC GAA TCC TGT GG -3’	5,273 – 5,292 [b]
dl	RH6R*^	5’- GAC GAC AAT CTG GAT CTG GT -3’	5,419 – 5,400 [b]
dm	RH8F^	5’- TCA GGG AAT TAG CAT TGA TGA -3’	7,189 – 7,209 [b]
dn	RH8R*^	5’- CAT ACG TTC AAT TCC ATC CA -3’	7,336 – 7,317 [b]
do	RH10F^	5’- AAA GTT GTA GCC ATG CTT GA -3’	8,947 – 8,966 [b]
dp	RH10R*^	5’- TGA CGT TGA GCT TTC ACA TC -3’	9,101 – 9,028 [b]
dq	RH15F^	5’- TTA AGT TGG CAA ATG ACT CAG -3’	13,379 – 13,399 [b]
dr	RH15R*^	5’- CAA ATA TCC CGA TTT GAA CG -3’	13,523 – 13,504[b]
ds	RH16F^	5’- CGG AAC CAA CAA AAT TCA AG -3’	14,375 – 14,394 [b]
du	RH16R*^	5’- CAT CCA ACC ATC ATT ACA TAC G -3’	14,520 – 14,499 [b]
dv	RH19F^	5’- TTG TTA TTA CTT CAA ACG AAC ACC -3’	16,295 – 16,318 [b]
dw	RH19R*^	5’- AAC TGT CCT CCA TTT TTC ACA -3’	16,436 – 16,416 [b]
dx	RH22F^	5’- AAA GCA TTG GAA GAA TGT CG -3’	19,524 – 19,543 [b]
dy	RH22R*^	5’- AAT ATC TCC CCC TCC AAA TC -3’	19,670 – 19,651 [b]
dz	RH23F^	5’- CAG ATA GTT TCA AGC GGA AAA -3’	20,249 – 20,269 [b]
ea	RH23R*	5’- GCC CTG AAA AAG CAC TCA TA -3’	20,404 – 20,385 [b]
eb	RH5eryR*[a]	5’- ttg ttc atg taa tca ctc ctt cCA TTA AGC TAC CAT CCG CTT TTC -3’	942,892 – 942,870[a] 7,372 – 7,349[b]
ec	RH7eryF[b]	5’- cgg gag gaa ata att cta tga gGT TCA AAA TGT CCA GTT	21,865 – 21,888[b]

		TGG ACA -3’	962,708 – 962,730[a]
ed	RH3cmlREagI*[c]	5’- gag atc cgg ccg ATG AAG ATT TTC TAG AGA ATT TTC -3’	939,832 – 939,809[a] 4,313 – 4,290[b]
ee	RH7cml$_{(a)}$xhoI[d]	5’- ata tat ctc gag GTT CAA AAT GTC CAG TTT GGA CA -3’	962,730 – 962,708[a] 21,865 – 21,887[b]
ef	RH4	5’- GGG ATT ATC CCA GGC GGT AC -3’	938,711 – 938,730[a] 3,192 – 3,211[b]
eg	RHspec1861aR*[c]	5’- tac agt cgg ccg CCA TCA CTA GCA TCA AAA TTG GG -3’	13,743 – 13,721[b]
eh	RHspec1861aF[d]	5’- tgc ata ctc gag GTT GGT GGT TAA ATC ACT TAG GTG -3’	13,758 – 13,781[b]
ei	RHspec1R$_{(2)}$*[c]	5’- gta tat cgg ccg CGT GTG CTA AAC CTG TTA CTG -3’	6,338 – 6,318[c]
ej	RHspec1F[d]	5’- gac tat ctc gag GGT GTT ATG GAG CTC CGG GC -3’	6,510 – 6,529[c]
ek	RH9F(1)	5’- GGG AAA AGT ATA TGG TTT TGT TG -3’	9,320 – 9,342[c]
el	RH8F(1)	5’- CAG GTG GGC AAA GGC AAC GG -3’	8,720 – 8,739[c]
ef	RH8R(1)*	5’- CAA CAA TCT TAC CAC CTT TCA TT -3’	8,949 – 8,927[c]
eg	RH6R(1)*	5’- CGT TTT GAA CGT GGG GAG TC -3’	6,031 – 6,012[c]
eh	RH11F(1)	5’- GAT ATT TCG GCA CCA TAC TCA G -3’	10,816 – 10,837[c]
ei	RH1051F	CGAACAGTTGATGTTCCAAAGA	7,723 – 7,744[b]
ej	RH24F	5’- CAA GTG GGC TAT ATG GAA CC -3’	20,807 – 20,827[b]
ek	RH24R*	5’- CCA AAT ACT CTG CCA AGA AAC TC -3’	21,098 – 21,076[b]
g	RH1068R$_{(2)}$*	5’- CCT GAT AAT CTT CCT GCG TTG -3’	965,317 – 965,301[a] 24,462 – 24,468[b] 14,476 – 14,482[c]
J	RH1052R*	5’- CAA TAT GCA ATT TAT CTG CTG CAC -3’	944150 – 944127[a] 8,606 – 8,629[b]
k	RH1046F(2)	5’- GAT GCC TGC AAG GTT CCT GAC TG -3’	939,875 – 939,897[a] 4,354 – 4,376[b]
n	RHx1*	5’- CAG TTC CCA TAC CAC CTA ATT G -3’	11,077 – 11,098[c]
o	RHy1	5’- CGA TGC TTG GCG TTT GGA TGT G -3’	4,968 – 4,989[b] 4,969 – 4,990[c]
p	RHx2*	5’- GCC TTC CCA ATA ACG TAT AAG C -3’	10,328 – 10,349[c]
q	RH1067eryF[#]	5’- cgg gag gaa ata att cta TGA GCG CCT TGC AGT TGG TTC -3’	964,028 – 964,048[a] 23,190 – 23,209[b] 13,204 – 13,223[c]
s	RH1046F$_{(4)}$	5’- GGC CTG AGT CGG AGT ATG CC -3’	939,525 – 939,544[a] 4,006 – 4,025[b] 4,005 – 4,024[c]
t	RH1041F	5’- GAG AGA TTA TCG ATG GTT ATA TTG G -3’	935,522 – 935,546[a] 1 – 25[b & c]
u	RH1046R*	5’- GGC ATA CTC CGA CTC AGG CC -3’	4,006 – 4,025[b] 4,005 – 4,024[c]
v	RHx3	5’- GTC TAC ATT CGA GAA ACG TCT T -3’	12,162 – 12,183[b] 5,835 – 5,856[c]
w	RHy2	5’- CAT GAA AGT CAT GCT GGA TGC GG -3’	4,695 – 4,714[b] 4,696 – 4,715[c]
y	RHy3*	5’- GAA GGT CCC AAA ACT AAT CCT ATC -3’	9,467 – 9,490[c]
z	RHx4	5’- GGT GTT ATG GAG CTC CGG GC -3’	6,510 – 6,529[c]
-	RH16SF$_{(3)}$^	5’- CAT GCA AGT AGA ACG CTG AA -3’	-
-	RH16SR$_{(3)}$^	5’- TGT CAT GCA ACA TCC ACT CT -3’	-
-	J293a[e]	5’- gat cat cgg ccg GGT TCG CGG GAA GTC TAC TAA G -3’	-
-	J254a*[f]	5’- tgc ata ctc gag TTA TAC CTT CTT CAA TCT GTT ATT TAA ATA GTT TAT AGT TA -3’	-
-	J214	5’- GAA GGA GTG ATT ACA TGA ACA A -3’	5,103 – 5,125[d]
-	J215*	5’- CTC ATA GAA TTA TTT CCT CCC G -3’	4,384 – 4,363[d]
-	RHcatF[e^]	5’- tat aat cgg ccg CAA CAG CGT GAC CGA AAA TTG -3’	-
-	RHcatR*[f^]	5’- tat aat ctc gag GGG TTC CGA GGC TCA ACG TC -3’	-
-	RHrtcomDF^	5’- GGT TCG TAT CAT GAG CGT TT -3’	2,041,784 – 2,041,803[a]

-	RHrtcomDR*^	5’- CCT GAA GGA GTC ATC GTC AT -3’	2,041,654 – 2,041,673[a]
-	RHrtcomXF^	5’- AAG GCA TGC TCT GCT TAC AT -3’	14,351 – 14,370[a] & 1,803,041 – 1,803,060[a]
-	RHrtcomXR*^	5’- TCT ACG CTT CTG ACT TTC CTG -3’	14,473 – 14,493[a] & 1,802,918 – 1,802,938[a]
-	RHrtcomWF^	5’- GGT GAT AAT TTT GAC TGG GAA C -3’	22,240 – 22,261[a]
-	RHrtcomWR*^	5’- AAC CCC GAT TCA TTA CCA T -3’	22,373 – 22,388[a]
-	RHrtcglAF^	5’- TCA GTT GCA GTT GAA CGA AG -3’	1,842,485 – 1,842,504[a]
-	RHrtcglAR*^	5’- CTG TCG CAC CTG TCA AAC TA -3’	1,842,353 – 1,842,372[a]
-	RHrtcoiAF^	5’- AGT CCC TTG CCT CAG AAA GT -3’	879,521 – 879,540[a]
-	RHrtcoiAR*^	5’- GCC CAT GTT TTG ACT GAA AT -3’	879,654 – 879,673[a]
-	3*	5’- TGA CAG TTG AGA GAA TCT TT -3’	Whatmore *et al.*, (1999)
-	4	5’- CTT TTC TAT TTA TTT GAC CT -3’	Whatmore *et al.*, (1999)
-	5	5’- CTT TTC TAT TTA TTT GAC CT -3’	Whatmore *et al.*, (1999)
MLST	aroE-up	5’- GCC TTT GAG GCG ACA GC -3’	Enright & Spratt, 1998
MLST	aroE-dn	5’- TGC AGT TCA (G/A)AA ACA T(A/T)T TCT AA -3	Enright & Spratt, 1998
MLST	gdh-up	5’- ATG GAC AAA CCA GC(G/A/T/C) AG(C/T) TT -3	Enright & Spratt, 1998
MLST	gdh-dn	5’- GCT TGA GGT CCC AT(G/A) CT(G/A/T/C) CC -3	Enright & Spratt, 1998
MLST	gki-up	5’- GGC ATT GGA ATG GGA TCA CC -3	Enright & Spratt, 1998
MLST	gki-dn	5’- TCT CCC GCA GCT GAC AC -3	Enright & Spratt, 1998
MLST	recP-up	5’- GCC AAC TCA GGT CAT CCA GG -3	Enright & Spratt, 1998
MLST	recP-dn	5’- TGC AAC CGT AGC ATT GTA AC -3	Enright & Spratt, 1998
MLST	spi-up	5’- TTA TTC CTC CTG ATT CTG TC -3	Enright & Spratt, 1998
MLST	spi-dn	5’- GTG ATT GGC CAG AAG CGG AA -3	Enright & Spratt, 1998
MLST	xpt-up	5’- TTA TTA GAA GAG CGC ATC CT -3	Enright & Spratt, 1998
MLST	xpt-dn	5’- AGA TCT GCC TCC TTA AAT AC -3	Enright & Spratt, 1998
MLST	ddl-up	5’- TGC (C/T)CA AGT TCC TTA TGT GG -3	Enright & Spratt, 1998
MLST	ddl-dn	5’- CAC TGG GT(G/A) AAA CC(A/T) GGC AT -3	Enright & Spratt, 1998
-	RHrtcomDF^	5’- GGT TCG TAT CAT GAG CGT TT -3’	2,041,784 – 2,041,803[i]
-	RHrtcomDR^	5’- CCT GAA GGA GTC ATC GTC AT -3’	2,041,654 – 2,041,673[i]
-	RHrtcomXF^	5’- AAG GCA TGC TCT GCT TAC AT -3’	14,351 – 14,370[i] & 1,803,041 – 1,803,060[i]
-	RHrtcomXR^	5’- TCT ACG CTT CTG ACT TTC CTG -3’	14,473 – 14,493[i] & 1,802,918 – 1,802,938[i]
-	RHrtcomWF^	5’- GGT GAT AAT TTT GAC TGG GAA C -3’	22,240 – 22,261[i]
-	RHrtcomWR^	5’- AAC CCC GAT TCA TTA CCA T -3’	22,373 – 22,388[i]
-	RHrtcglAF^	5’- TCA GTT GCA GTT GAA CGA AG -3’	1,842,485 – 1,842,504[i]
-	RHrtcglAR^	5’- CTG TCG CAC CTG TCA AAC TA -3’	1,842,353 – 1,842,372[i]
-	RHrtcoiAF^	5’- AGT CCC TTG CCT CAG AAA GT -3’	879,521 – 879,540[i]
-	RHrtcoiAR^	5’- GCC CAT GTT TTG ACT GAA AT -3’	879,654 – 879,673[i]
-	RH06Fa	5’- GGT CCT GGT CGT GAA CAA AC -3’	24,086 – 24,105[j] 24,093 – 24,112[k]
-	RH06R*	5’- GCT AAA CTC CCT GTA TCA AGC G -3’	58,233 – 58,212[j]
-	RH06R$_{(3)}$*	5’- GCC TCT CTC AAA GCC TCC CTC -3’	59,883 – 59,863[k]
-	RHrtPblBF^	5’- GAA CGT TCA CAA CAG GAG GT -3’	51,879 – 51,898[j]
-	RHrtPblBR*^	5’- TCA TTT ACT AGG GCG ACA GG -3’	52,020 – 52,001[k]
-	RH152F	5’- CAT GTA CAC CTG ATA TGA TGC G -3’	528,307 – 528,328[j] 528,333 – 528,354[k]
-	RH152Ra*	5’- GGT TCC TGC TCA CTT GTA TTA G -3’	531,406 – 531,385[j] 531,434 – 531,413[k]

-	RH208/09F	5’- GAA GAT AGG GGA GCG GAA GAG -3’	625,104 – 625,124[j] 625,146 – 625,166[k]
-	RH208/09R*	5’- GGC AAA GGT ATA TCT CTC CCC -3’	639,242 – 639,622[j] 629,967 – 629,947[k]
-	RH224/5F	5’- GAG GAA GAG GAC ATA GAA AAT GG -3’	628,775 - 628,797[j] 682,818 – 682,840[k]
-	RH224/5R*	5’- GCC CTC TGT ACC GGC TGG G -3’	685,213 – 685,195[j] 685,255 – 685,237[k]
-	RH344F	5’- GTT GAA CCC GCA ATT CAG CCT G -3’	1,053,577 – 1,053,598[j] 1,053,650 – 1,053,671[k]
-	RH344R*	5’- GAC ATG TGA CCA GGA AAC CAT TG -3’	1,117,411 – 1,117,389[j] 1,117,485 – 1,117,463[k]
-	RHrtzmpDF$^\wedge$	5’- GAA CCG GAA GTT CAT GAG AA -3’	1,059,407 – 1,059,426[j]
-	RHrtzmpDR*$^\wedge$	5’- CAC ATA CGG AAA GCA TTG TG -3’	1,063,677 – 1,063,658[k]
-	RHumuCF	5’- CAT CGA TTG CGG ACT GGA GCC T -3’	Henderson-Begg *et al.*, 2009
-	RHumuCR*	5’- CGC TTT GTG TTA TGA GTC GTG C -3’	Henderson-Begg *et al.*, 2009
-	RHint5252F	5’- TTA AGC CAA TGC AGA AAT GC -3’	Henderson-Begg *et al.*, 2009
-	RHint5252R*	5’- CCA TGC AAT GGA AAA ACC TC -3’	Henderson-Begg *et al.*, 2009
-	RHint916F	5’- GCG TGA TTG TAT CTC ACT -3’	Henderson-Begg *et al.*, 2009
-	RHint916R*	5’- GAC GCT CCT GTT GCT TCT -3’	Henderson-Begg *et al.*, 2009
-	RH383F	5’- CAA CAC GTC CAT AAT GAG CTG TAC -3’	1,252,783 – 1,252,806[j] 1,252,850 – 1,252,873[k]
-	RH383R*	5’- CAG AGC GTT ATT ATA AGC TCC CC -3’	1,278,553 – 1,278,531[j] 1,278,617 – 1,278,595[k]
-	RH386Fa	5’- GGG ACT GTA TAT GAA CGA ACG C -3’	1,260,270 – 1,260,291[j] 1,260,337 – 1,260,358[k]
-	RH386F$_{(2)}$	5’- YGA ATC CTG TGG AAC TAC TCC -3’	1,257,753 – 1,257,773[j] 1,257,820 – 1,257,840[k]
-	RH386R$_{(2)}$*	5’- GCC ATA TAG CAT TGT CCA TAA CG -3’	1,272,092 – 1,272,070[j] 1,272,137 – 1,272,159[k]
-	RH386F$_{(3)}$	5’- GAA GGA GRT GAT AAA GTC CAT C -3’	1,272,256 – 1,272,277[j] 1,272,323 – 1,272,344[k]
-	RH386R$_{(3)}$*	5’- GGG GCA GTA TGG GAC TAC ATT TGG -3’ 5’- GGG GAA GTA TGG GTC TAC ATT TGG -3’	1,068,657 – 1,068,680[k]
-	RH430F	5’- CTC TAT TAC CTC TTA TTA TAC CAC -3’	1,506,959 – 1,506,982[j] 1,507,033 – 1,507,056[k]
-	RH430R*	5’- CAA TAC AGA ACT AAA TCC TTC GG -3’	1,508,581 – 1,508,560[j] 1,507,656 – 1,508,634[k]
-	RH434Fa	5’- GTT AAG AAG TAT CGT ACG ACT TGG -3’	1,509,535 – 1,509,558[j] 1,509,610 – 1,509,632[k]
-	RH434Ra*	5’- GGT GTC ATG ATG TTG GTC TTT AC -3’	1,517,257 – 1,517,235[j] 1,517,332 – 1,517,310[k]
-	RH467F	5’- GAT ACA TGG TCA ATA CCT CTA C -3’	1,653,619 – 1,653,640[j] 1,653,702 – 1,653,723[k]
-	RH467R*	5’- CCA CTC CGT GAT ACC ATC C -3’	1,656,637 – 1,656,619[j] 1,656,718 – 1,656,700[k]
-	RH469F	5’- GCG CAT AAC CTG AAA CAT ATG C -3’	1,659,021 – 1,659,042[j] 1,659,102 – 1,659,123[k]
-	RH469R*	5’- CTT GTT AGC AGA GCT GGA AAA G -3’	1,660,672 – 1,660,651[j] 1,660,751 – 1,660,730[k]
-	RH513F	5’- GAA GAA TTA CCT CAT CCA ACT TG -3’	1,957,613 – 1,957,635[j] 1,957,707 – 1,957,729[k]

-	RH513R*	5’- GTC TTG TTC GTA GGA CTC GCC -3’	1,959,167 – 1,959,147^{j} 1,959,260 – 1,959,240^{k}
-	RHPspAF	5’- CCC GTA AGT CAG TCT AAA G -3’	153,339 – 153,356^{j} 153,356 - 153,370^{k}
-	RHPspAR*	5’- CCA TGA ATC GTT GTA TTG GAG CC -3’	154,559 – 154,537^{j} 154,581 – 154,559^{k}
-	RHPrtAF	5’- CTT CAA AGG ATA CTG AAG AAA AG -3’	593,541 – 593,563^{j} 593,578 – 593,600^{k}
-	RHPrtAR*	5’- CAG AAT AGA TTT TAC CAG ATT CC -3’	598,121 – 598,099^{j} 598,159 – 598,137^{k}
-	RH316F	5’- GTG TTG ACC TTA ACG TAG GTT AAG -3’	931,379 – 931,102^{j} 931,441 – 931,464^{k}
-	RH316R*	5’- GGG CTC CAA AAT CAA GCA GAT AC -3’	931,925 – 931,903^{j} 931,970 – 931,949^{k}
-	RHseqhylAF$_{(2)}$	5’- GAG CAA GTG TTC AAG TTG GT -3’	305,949 – 305,965^{j} 305,975 – 305,994^{k}
-	RHhylAF	5’- GTC GGA TGG AAG ATA TTG TTG AAG -3’	304,294 – 304,317^{j} 304,320 – 304,343^{k}
-	RHhylAR*	5’- CCG AAG ATA TTT AGC CTT CTT CG -3’	307,754 – 307,731^{j} 307,747 – 307,769^{k}
-	RHPhtD(a)F	5’- CCT TAT GAT GCC ATC ATC AGT G -3’	922,342 – 922,363^{j} 922,403 – 922,424^{k}
-	RHPhtD(a)R*	5’- GGT AAT GGT CAT AAT GAG GTA TG -3’	923,986 – 923,964^{j} 924,047 – 924,025^{k}
-	RHPhtD(b)F	5’- CCG TCT TCT GTT GTG TAC TTG C -3’	1,137,802 – 1,137,827^{j} 1,137,874 – 1,137,895^{k}
-	RHPhtD(b)R*	5’- CTC ATG CGG ATA ATG TCC GTA C -3’	1,139,038 – 1,139,017^{j} 1,139,110 – 1,139,089^{k}
-	RHrt0083F^	5’- TGG CTC AGG CTA TAT GCT TT -3’	58,480 – 58,499^{j}
-	RHrt0083R*^	5’- TCA TGG CAC CTT CTA CAT CA -3’	58,622 – 58,603^{j}
-	RHrt0747F*^	5’- TCC AGT CAG GAA TCT CCA TC -3’	682,959 – 682,940^{j}
-	RHrt0747R^	5’- AAA GAG TTG AGG CAA GAC GA -3’	682,817 – 682,836^{j}
-	RHrt1340F*^	5’- GTT GAC CTA GAT GCG GAA AA -3’	1,252,969 – 1,252,950^{j}
-	RHrt1340R^	5’- CAC GAA CCA CAC TTT CAT TG -3’	1,252,813 – 1,252,832^{j}
-	RHrt1353F*^	5’- CGT CCA GAA TAT CCA GGT GT -3’	1,263,338 – 1,263,319^{j}
-	RHrt1353R^	5’- TTC CTT CGA TTG CGT AGA TT -3’	1,263,077 – 1,263,096^{j}

* Reverse complementary to target

# Only PPI-1 primers were designated a code

^ Primers designed with parameters in Table 2.2

Lowercase type indicates non-homologous sequence to target

Underlined type indicates *e*. *Eag*I site; *f. Xho*I site; *g*. anneals to J214; *h.* anneals to J215

Coordinates derived from *a.* TIGR4 Genbank accession number AAGY00000000; *b.* strain 1861 (PCR *t* – *g*); *c.* strain 1 (PCR *t* – *g*); *d.* pVA891 (Macrina *et al.*, 1983); *i.* D39 Genbank accession number CP1000410; *j.* strain 1861 consensus sequence (Section 5.3); *k.* strain 1 consensus sequence (Section 5.3)

2.4 Serotyping of pneumococcal strains

Serotyping of clinical isolates was performed by the MSHR, Darwin, Australia and the WCH, North Adelaide, Australia.

2.5 Optochin sensitivity

In order to distinguish between *S. pneumoniae* and other bacterial species with similar colony morphology, colonies were tested for optochin sensitivity (an intrinsic characteristic of *S. pneumoniae*) by plating on BA in the presence of a 5 μg optochin disc (Oxoid) and incubating at 37°C in 95% air/5% CO_2 for >16 h. A zone of growth inhibition surrounding the optochin disc confirmed *S. pneumoniae*.

2.6 Bioinformatic software

2.6.1 Sequence analysis

Genomic sequence reads were analysed using SeqMan Pro from Lasergene® version 8 (DNASTAR Inc.). Prediction of open reading frames (ORF) and searches for direct and inverted repeats were performed using DNAMAN version 4.15 (Lynnon BioSoft, Quebec, Canada). Homology profiles of large regions of sequence were undertaken using the Artemis Comparison Tool (ACT) version 8 (Genome Research Limited, http://www.sanger.ac.uk/Software/ACT/v8). Comparison files for ACT were generated using Double ACT version 2 (Health Protection Agency, http://www.hpa-bioinfotools.org.uk/pise/double_act.html). ClustalW alignments were performed using the alignment tool at http://align.genome.jp/. Promoter prediction searches were performed using the tool at http://www.fruitfly.org/seq_tools/promoter.html. Rho-independent transcription terminators were predicted using the tool at http://www.softberry.com.

2.6.2 Search engines

The sequence search engines of the online databases at the National Center for Biotechnology Information (NCBI) (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and the

Kyoto Encyclopaedia for Genes and Genomes (KEGG) (http://www.genome.jp/kegg) were used to identify homologous nucleotide or amino acid sequences of predicted ORFs and motifs. In addition, the HHpred search engine (http://toolkit.tuebingen.mpg.de/hhpred) was used to confirm the results of sequence searches performed on the NCBI and KEGG databases and for proteins for which a known ortholog was not present in either the KEGG or NCBI databases. The HHpred search engine performs Hidden Markov Model (HMM) comparisons between the query sequence and databases of protein structures. HMMs contain information on the conservation of each residue position, which increases the sensitivity of the searches when compared to sequence-only search engines.

2.7 Preparation of frozen stock cultures for *in vitro* growth measurements

Frozen stock cultures were made by growing the relevant strain in 10 ml THY to A_{600} 0.5 – 0.6. The precise culture density to which strains were grown was kept consistent within experiments. Cultures at the desired density were placed on ice and then concentrated at 3,200 × *g* for 20 mins at 4°C and resuspended in 1/10th of the starting volume with THY + 15% glycerol, which were kept at -80°C in 100 μl aliquots until required. A_{600} readings were taken using 1 ml aliquots of the relevant culture in 1.5 ml disposable cuvettes (PLASTIBRAND®) and the OD600 DiluPhotometer™ (Implen, Munich, Germany).

2.8 Transformation of *S. pneumoniae*

2.8.1 Preparation of competent cells

Pneumococci were grown in complete-CAT (cCAT) medium (10 g/l Bacto Casamino acids [Difco], 5 g/l Bacto Tryptone, 5 g/l NaCl, 10 g/l Bacto yeast extract, 4% [v/v] 0.4 M K_2HPO_4, 0.002% [w/v] glucose, 150 mg/l glutamine) to an A_{600} of 0.5 – 0.6, then diluted to an A_{600} of 0.02 in CTM medium (cCAT supplemented with 0.2% [w/v] bovine serum albumin [BSA], 1% [v/v] 0.01 M $CaCl_2$) and grown to an A_{600} of 0.2. Cells were subsequently pelleted at 3,200 × g for 20 min, resuspended in 1/10th volume CTM pH 7.8 + 15% [v/v] glycerol, and stored at -80°C in 50 μl aliquots.

2.8.2 Transformation of *S. pneumoniae*

500 μl of CTM-pH 7.8 and CSP-1 to a final concentration of 100ng/ml (Havarstein *et al*., 1995) (obtained from Chirontech, Victoria, Australia) were added to a 50 μl aliquot of competent pneumococci. The cells were subsequently incubated for 5 – 20 mins (optimised for each batch of competent cells) at 37°C, before addition of approximately 1 ng donor DNA. Subsequently the transformation mix was incubated for 30 min at 32°C before incubating further at 37°C for 2 – 4 h. After incubation, cells were plated onto BA supplemented with the appropriate antibiotic, and incubated for >16 h at 37°C in 95% air/5% CO_2.

2.8.3 Electrotransformation of *S. pneumoniae*

The method for the electrotransformation of *S. pneumoniae* was based on the protocol described by Lefrancois and Sicard (1997). Strains were grown in cCAT to OD_{600} 0.6 – 0.8, and then were washed and resuspended in 1/10th starting volume of electroporation medium (0.5 M Sucrose; 7 mM potassium phosphate, pH 7.5; 1 mM magnesium chloride), whilst on ice. Electroporation was carried out at 1.25 kV and 25 μF, following the addition of the relevant DNA. Immediately following electroporation, pre-warmed SB was added to the mixture of cells, which were then incubated for 2 h at 37°C. Subsequently, cells were spun down and resuspended in 100 μl SB before spreading onto BA supplemented with the relevant antibiotic.

2.9 DNA isolation and manipulation

2.9.1 Agarose gel electrophoresis

DNA was electrophoresed through horizontal agarose gels (0.6-2.5% [w/v] agarose dissolved in Tris Borate and EDTA (TBE) buffer [44.5 mM Tris, 44.5 mM boric acid, 1.25 mM EDTA, pH 8.4]) immersed in TBE buffer at 180 V. Prior to loading DNA samples, a 1/10th volume of loading buffer was added (15% [w/v] Ficoll, 0.1% [w/v] bromophenol blue, 100 ng/ml Rnase A). Gel staining was undertaken with 3 × GelRed™ (Biotium) + 0.1 M NaCl in MilliQ (MQ) H_2O for ~ 30 min at room temperature (RT), as described in the manufacturer's instructions. DNA bands were visualised by transillumination with short wavelength UV using the Gel Doc XR system (Bio-Rad, NSW, Australia). Approximate sizes of visualised fragments were calculated by comparison of their mobility with that of DNA size markers. The markers used in this study included *Eco*R1-Digested *Bacillus subtilis* bacteriophage SPP-1 DNA (Geneworks, SA, Australia) with fragment sizes of 8.56, 7.43, 6.11, 4.90, 3.64, 2.80, 1.95, 1.88, 1.52, 1.41, 1.16, 0.99, 0.71, 0.49, 0.36 and 0.08 kb and 1 kb-plus ladder (Invitrogen, Vic, Australia) with fragment sizes of 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1.65, 1, 0.85, 0.65, 0.5, 0.4, 0.3, 0.2 and 0.1 kb. Also used was *Hin*dIII restricted bacteriophage lambda DNA (Geneworks) with fragment sizes 23.1, 9.41, 6.55, 4.31, 2.32, 2.02, 0.56 and 0.12 kb.

2.9.2 *S. pneumoniae* chromosomal DNA isolation for applications other than CGH and next generation sequencing

S. pneumoniae chromosomal DNA was isolated using the Wizard Genomic DNA purification kit (Promega, WI, USA), according to the manufacturer's instructions with the following modifications; cells were lysed by re-suspension in 200 μl 50 mM EDTA, 0.1% (v/v) DOC. Following the removal of supernatant after protein precipitation, DNA was precipitated using 2.5 × supernatant volume of 100% ice-cold ethanol. DNA was spooled using a Pasteur pipette, washed in 70% (w/v) ethanol and resuspended in 1 × Tris EDTA (TE).

2.9.3 *S. pneumoniae* chromosomal DNA isolation for CGH and next generation sequencing

20 ml O/N cultures of *S. pneumoniae* were split into two, centrifuged at 3,200 × *g* for 20 min at RT and pellets were resuspended in 4 ml buffer (0.1 M sodium citrate, 0.14 M sodium chloride, pH 7.8). DNA-containing buffer was extracted with Tris-HCl equilibrated Phenol (pH 7.5), until any white interface material between the aqueous and phenol phases had been completely removed. Each extraction included adding an equal volume of phenol to buffer, mixing gently for 1-2 min and separating the aqueous and phenol phases by centrifugation at 1,620 × *g* for 5 min at RT. Following the phenol extractions, contaminating phenol was removed from the aqueous phase by extracting twice with an equal volume of chloroform. DNA was precipitated in 2 × volume 100% ice cold ethanol and 1/10th volume 3 M sodium acetate. Precipitated DNA was spooled using a Pasteur pipette, washed in 70% (w/v) ethanol and resuspended in ~500 μl 1 × TE per initial starting 20 ml culture.

2.9.4 Restriction endonuclease digestion of DNA

Restriction endonuclease digestion of DNA was carried out at a DNA concentration determined by its subsequent use and with 1-2 U of the relevant restriction enzyme, in the appropriate buffer according to the manufacturer's instructions (NEB). Reactions took place at the manufacturer's suggested temperature for 1-16 h. Restriction digests were analysed by agarose gel electrophoresis, as described in Section 2.9.1.

2.9.5 DNA Ligation

DNA ligation reactions were performed using T4 DNA ligase (Roche Diagnostics). Each reaction contained approximately 3 U of T4 DNA ligase, 10 × T4 DNA ligase buffer and 500 ng of each ligation substrate and was incubated at RT O/N to complete the reaction.

2.9.6 Polymerase chain reaction

PCR reactions were performed using either G-STORM GS482 (Gene Technologies, UK), PCRSprint (Hybaid) or Mastercycler (Eppendorf, NSW, Australia) thermal cyclers, in a final volume of either 25μl or 50μl. Standard reactions were carried out using *Taq* DNA polymerase (Roche diagnostics), as described by the manufacturer's instructions, with the exception that the buffer was used at 1.6 × concentration. The Expand™ Long Template PCR System and Expand™ High Fidelity PCR System (Roche Diagnostics) were used where either amplification of long templates or high-fidelity amplification was required. Long-range PCRs included 3% (v/v) DMSO. Typical reaction conditions comprised 25 cycles of denaturation at 94°C (92°C for long templates >10 kb) for 30 s, annealing at 55°C for 30 s and extension at 68°C for various lengths of time (approximately 1 min for each kb of expected PCR product).

2.9.7 PCR product purification

PCR products were purified using the MinElute® PCR purification kit as described in the manufacturer’s instructions (Qiagen, Vic, Australia). When DNA was to be used for transformation or sequencing reactions, it was eluted in MQ H_2O.

2.9.8 DNA sequencing

DNA sequencing was performed using the thermal cyclers described in Section 2.9.6. Reactions were performed in 20 μl volumes, consisting of ~25 ng template DNA, 10 μM oligonucleotide primer, 3 μ1 BIG DYE terminator buffer and 1 μl BIG DYE terminator mix (version 4). The sequencing reaction consisted of 25 cycles at 96°C for 30s, 55°C for 15s and 60°C for 4 min, followed by holding at 4°C. The reaction product was purified by adding 1 μl glycogen, 20 μl MQ H_2O, and 60 μl 100% isopropanol. DNA was precipitated at RT for 15 – 180 min, and then centrifuged at 16,000 × *g* for 30 min at 4°C. Pelleted DNA was washed with 250 μl of 75% isopropanol and centrifuged for 20 min as before. The supernatant was removed and the pellet dried at 65°C for 10 min. Sequencing reactions were run by SA Pathology, Adelaide, using an Applied Biosystems 3730 model automated sequencer.

2.9.9 Next generation sequencing

The genomes of *S. pneumoniae* strains were sequenced by Geneworks, Adelaide, using the Illumina® Genome Sequence Analyzer *II* as 35-bp reads from chromosomal DNA, which was prepared as described in Section 2.9.3. Generated sequence was assembled by Geneworks using SeqMan NGen™ version 1.2 (DNASTAR Inc.). Sequence data obtained from Geneworks was analysed using the software described in Section 2.6.1 and by methods which are described in more detail in Chapter 5.

2.10 Multi-locus sequencing typing of *S. pneumoniae* strains

STs for *S. pneumoniae* strains were determined as per the instructions at http:///www.mlst.net by amplifying and sequencing defined regions of seven housekeeping genes; shikimate dehydrogenase (*aorE*), glucose-6-phosphate dehydrogenase (*gdh*), glucose kinase (*gki*), transketolase (*recP*), signal peptidase I (*spi*), xanthine phosphoribosyltransferase (*xpt*) and D-alanine-D-alanine dehydrogenase (*ddl*) (Enright & Spratt, 1998). Primers used for amplification and sequencing of the seven target genes are shown in Table 2.3. The nucleotide sequence derived from each gene was subsequently compared to the *S. pneumoniae* database, where each gene was designated a number that corresponds to a previously identified identical sequence. The allele numbers for all seven genes together form the allelic profile, which was compared to the database for determination of the ST. Where either a unique allele was sequenced or a new allelic profile was identified, the relevant information was submitted to the database curator for designation with a unique ST number.

2.11 Comparative genomic hybridisation

2.11.1 Generation of microarray probes

Approximately 20 μg of genomic DNA was isolated (Section 2.9.3), digested with *Sau*3a (Section 2.9.4) and purified (Section 2.9.7). Labelling of genomic DNA was performed using the Genisphere Array 900 DNA™ DNA labelling kit (Genisphere, P.A., USA), as described in the manufacturer's instructions.

Extension products were purified as described in Section 2.9.7 and eluted in 20 μl MQ H_2O. Eluted samples were dried and resuspended in 4.5 μl MQ H_2O containing either suspended Alexa Fluor® 555 Reactive Dye (Invitrogen) or Alexa Fluor® 647 Reactive Dye (Invitrogen). The labelling reaction was allowed to proceed for at least 1 h at RT and then quenched with 4.5 µl 4 M hydroxylamine for 15 min. Both samples were then mixed, purified (Section 2.9.7) and eluted with 20 µl of MQ H_2O.

2.11.2 Hybridisation to microarray slides

Microarray slides were obtained from the Bacterial Microarray Group at St. George's Hospital, University of London, and consisted of PCR products for each of the identified ORFs of the *S. pneumoniae* TIGR4 genome, supplemented with additional PCR products of ORFs identified in the sequence of *S. pneumoniae* R6.

The microarray slide was blocked with blocking solution (1 % [w/v] BSA, 0.1 % [v/v] SDS and 3.5 × SSC [20 × SSC contains 0.15 M NaCl, 0.15 M sodium citrate]), at 65°C for 30 min. Following blocking, the slide was washed twice with MQ H_2O and twice with isopropanol. The slide was dried using nitrogen gas.

The hybridisation mix (20 µl probe, 3.4 µl 20 × SSC and 0.6 µl 10% [v/v] SDS) was heated at 95°C for 5 min, dispensed on the slide and covered with a cover slip. The slide was incubated at 65°C O/N in a humidified chamber. The following day the slides were washed for 15 min in 2 × SSC + 0.03% (v/v) SDS at 65°C, then for 15 min in 1 × SSC at RT and finally for 15 min in 0.2 × SSC at RT. The slides were dried using nitrogen gas and scanned using a GenePix 4000B scanner (Molecular Devices CA, USA). Images were acquired using Gene Pix Pro 6.0 (Axon).

2.12 RNA Isolation, Manipulation and Analysis

2.12.1 Acid-phenol RNA extractions

RNA was isolated from bacterial cell pellets with acid-phenol:chloroform:isoamyl alcohol (25:24:1; pH 4.5, Ambion, Austin, TX., USA) as described previously (Ogunniyi *et al*., 2002). Cells were pelleted by centrifugation at 15,500 × *g* for 20 min at 4°C and after removing the supernatant, the pellet was resuspended in 400 μl pre-warmed acid-phenol and incubated at 65°C for 5 min before adding 400 μl pre-warmed NAES (50 mM sodium acetate, 10mM EDTA, 1% [w/v] SDS, pH 5.1, treated with diethyl pyrocarbonate [DEPC]), mixing and incubating at 65°C for a further 5 min. Subsequently, a 1 min incubation on ice was carried out, followed by centrifugation at 15,500 × *g* for 2 min at 4°C. The supernatant was then removed and washed with 400 μl of pre-warmed acid phenol to commence the second round of extraction. 2-3 rounds of extraction were performed until any white interface material between the aqueous and organic phases had been completely removed. Following the extractions, RNA was precipitated by adding 2 volumes of 100% ethanol and 1/10th volume 0.05% (w/v) DEPC-treated 3 M sodium acetate to the remaining aqueous phase, in the presence of 40 ng/μl glycogen (Roche) (*in vivo* samples only) and incubated at -80°C for >2 h. Following precipitation, the RNA was pelleted by centrifugation at 15,500 × *g* for 30 min at 4°C, the supernatant removed, the pellet washed in 70% (w/v) ethanol by pipetting up and down and then repelleted by centrifugation at 15,500 × *g* for 20 min at 4°C. Subsequently, the supernatant was removed and residual ethanol in the pellet was removed by drying in a SpeedVac for 5 min. The dried pellet was then resuspended in 18 μl nuclease-free water (Roche). Prior to DNase treatment, resuspended RNA was heated at 95°C for 3 min to denature any RNA:DNA complexes that could otherwise escape DNase treatment. Contaminating

DNA was degraded by treatment with 10 U RNase-free DNase (Roche) at 37°C for 1 h in the presence of 1 U/µl RNasin ribonuclease inhibitor (Promega Life Sciences, WI., USA). Successful removal of DNA contamination was assessed by one-step RT-PCR using the Access RT-PCR system kit (Promega), which was carried out as described in Section 2.12.5, using primers RH16sF$_{(3)}$ and RH16sR$_{(3)}$ (Table 2.3). Successful treatment was determined by the absence of amplification in reactions lacking the reverse transcriptase enzyme.

2.12.2 Enrichment for prokaryotic RNA

Extracted *in vivo* RNA samples (Section 2.12.1) were initially purified by using the RNeasy^{TM} minikit (Qiagen). Prokaryotic RNA derived from the lungs of infected mice was enriched using the MICROB*Enrich*^{TM} kit (Ambion), as described in the manufacturer's instructions.

2.12.3 RNA amplification

Following RNA extraction (blood- or nasal wash-derived RNA) or prokaryotic RNA enrichment (lung-derived RNA), samples were purified using the RNeasy^{TM} mini kit (Qiagen). RNA was then amplified in a linear fashion using the SenseAmp^{TM} RNA amplification kit (Genisphere, Hatfield, PA., USA), according to the manufacturer's instructions. In most cases a second round of amplification was required and this was undertaken in accordance with the manufacturer's instructions. Amplified RNA was stored at -80°C until required.

2.12.4 Generation of cDNA

The generation of cDNA from RNA was performed by adding equal weight of random primers (Invitrogen) to RNA, followed by incubation at 65°C for 5 min and incubation on ice for 1 min. The mix was made up to a total volume of 14 µl with MQ

H_2O before adding 4 μl first strand buffer, 1 μl 0.1 M DTT and 1 μl Superscript® III (200 U) and incubating at RT for 5 mins. Reverse transcription was undertaken by incubating the mix at 50°C for 1 h. Reverse transcription was performed in a thermal cycler to reduce condensation on the tube lids. Following reverse transcription, cDNA was purified using the MinElute® PCR purification kit as described in Section 2.9.7, with an additional wash with 80 % ethanol (w/v) following the PE buffer wash.

2.12.5 Reverse transcription polymerase chain reaction (RT-PCR)

RT-PCR for the purpose of detecting DNA contamination of samples was carried out using the Access RT-PCR System (Promega) according to the manufacturer's instructions. A standard 12 μl RT-PCR reaction contained DEPC-treated water, 0.2 mM of each dNTP, 1 μM of each primer, 1mM $MgSO_4$, AMV/*Tfl* reaction buffer, 0.1 U/μl AMV Reverse Transcriptase, 0.1 U/μl *Tfl* DNA polymerase, 0.25 μl RQ1 DNase stop solution and 0.25 μl of the relevant template. A control reaction was run in tandem, lacking reverse transcriptase and thus amplification would only be possible if DNA contamination was present.

RT-PCR was carried out initially with a reverse transcription cycle at 48°C for 45 min, followed by 25 cycles of amplification comprising denaturation at 95°C for 30s, annealing at 55°C for 30 s and extension at 68°C for 20s.

2.12.6 Real-time RT-PCR

Real-time RT-PCR was performed using the Superscript™ III Plantinum® SYBR® Green One-Step qRT-PCRkit (Invitrogen), with specifically designed primers (Section 2.3.2).

Real-time RT-PCR was carried out using a LightCycler® 480 II (Roche). The cycling program used for real-time RT-PCR was 50°C for 15 min, 95 °C for 5 min followed by 40 cycles of 95°C for 15 s, 55°C for 30 s and 72°C for 15 s. The acquisition of fluorescence was undertaken at the 72°C step of each cycle. The cycling phase was followed by 40°C for 1 min. Melt curves were generated by 95°C for 5 s, 65°C for 1 s, which was increased at a rate of 0.11 °C/s to 97°C. The acquisition of fluorescence was carried out continuously during the gradual temperature increase phase.

2.12.7 Real-time RT-PCR analysis

The relative amount of target mRNA present in different RNA samples was calculated using the comparative cycle threshold ($2^{\Delta\Delta Ct}$) method (Livak & Schmittgen, 2001). The amount of target mRNA in one sample was compared with the amount of the same target mRNA in another sample, relative to the internal control, 16S rRNA. Alternatively, the amount of target mRNA relative to the amount of 16S rRNA was also calculated in some instances.

The standard deviations (SD) were determined as $\sqrt{}((\text{SDsample})^2 + (\text{SD16S})^2)$, and this was applied to the formulas: SD+ $=2^{\Delta\Delta Ct\text{-}SD} - 2^{\Delta\Delta Ct}$ and SD– = $2^{\Delta\Delta Ct+SD}$.

2.13 Challenge of mice

2.13.1 Growth of challenge strain

Strains of *S. pneumoniae* to be used for challenge studies were inoculated from an O/N BA plate into sterile SB and grown to A_{600} 0.16. At this density, the concentration of bacteria is approximately 10^8 CFU per ml. Bacteria were then diluted to the appropriate concentration with sterile SB such that 50 μl aliquots used for i.n. administration contained the required challenge dose. The actual dose delivered was determined retrospectively by plating serial dilutions of the challenge inocula following mouse challenge.

2.13.2 Intranasal challenge

Mice were anaesthetised by i.p. injection of Nembutal (pentobarbitone sodium; Rhône Mérieux, QLD, Australia), at a dose of 66 μg per 1 g of body-weight. Following anaesthesia, the appropriate dose (~10^7 CFU) of *S. pneumoniae* was administered in a 50 μl volume by pipetting onto the nares and allowing the animal to inhale the culture deep into the lungs. After challenge, mice were monitored regularly for signs of disease and survival time was recorded.

2.13.3 Quantitation of *S. pneumoniae* in mouse tissues

Following i.n. challenge, mice were euthanased by CO_2 asphyxiation at the relevant time points post-challenge. After exposure of the trachea, the nasopharynx was washed with 1 ml 0.5% (w/v) trypsin, 0.02% (v/v) EDTA, in sterile phosphate buffered saline (PBS), by insertion of a 26-gauge needle sheathed in tubing into the tracheal end of the upper respiratory tract. The solution was allowed to drip in to the nasopharynx slowly and collected from the nose. 40 μl of the collected samples was taken for the enumeration of pneumococci. Additionally, the upper palate and nasopharynx were excised and placed into 2 ml sterile PBS to ensure all pneumococci from the nasopharynx were recovered. 1 ml blood was taken from the aorta and dispensed into a heparinised tube. 40μl blood was immediately added to 160 μl sterile PBS for quantitation of bacteraemia. Both lungs were excised following extensive perfusion with sterile PBS + 0.02% EDTA and placed into 2 ml sterile PBS + 0.02% EDTA. The nasopharyngeal tissue and lungs were homogenised using ceramic beads in the Precellys®24 tissue homogeniser (Bertin Technologies, France), before 40 μl was taken for enumeration of pneumococci in each niche. 40 μl of each sample was diluted in 160 μl sterile PBS, and eight serial five-fold dilutions were performed. 25 μl of each dilution was spotted, in duplicate, onto BA supplemented with and without gen (RNA analysis only). Plates were incubated at 37°C in 95% air/5% CO_2 for >16 h and colonies were counted and recorded. Plates lacking gen were used to assess contamination of the collected samples, which would determine whether the relevant sample was suitable for later RNA analysis.

2.13.4 Preprartion of infected mouse tissues for extraction of prokaryotic RNA

Following the quantitation of *S. pneumoniae* CFU in Section 2.13.3, samples were prepared for the extraction of prokaryotic RNA. The nasal wash samples underwent centrifugation at 15,500 × *g* for 5 min at 4°C, and all but the last 40 μl of supernatant was removed before storing at -80°C. Subsequently, red blood cells, and lung tissue was separated from the remainder of their respective samples by differential centrifugation at 850 × *g*, for 6 min at 4°C. Subsequently, pneumococci were separated from the remaining blood and lung-derived samples and stored as pellets at -80°C following centrifugation at 15,500 × *g* for 5 min at 4°C. RNA extraction was performed on these samples as described in Section 2.12.1.

2.13.5 Calculation of competitive index *in vivo*

The competitive index (CI) is a measure of the competitive fitness of one strain versus another and is expressed as ratio (CI = 1 indicates a ratio of 1:1). The CI is calculated as follows; CI = output ratio (OR) / input ratio (IR). The IR is determined by calculating the ratio of the two challenge strains in the incoculum using antibiotic selection that is specific for one of the two strains. The OR is calculated in the same manner as for the IR, but from pneumococci recovered from the infected mouse tissues, as described in Section 2.13.3.

Chapter 3 – Characterisation of Pneumococcal Pathogenicity Island-1 Sequence in a Selection of Serotype 1 Clinical Isolates

3.1 Introduction

Serotype 1 isolates have been historically associated with a high invasive potential as they are rarely isolated from the nasopharynx of healthy carriers (Section 1.5.2.1). However, the recent detection of clusters of serotype 1 carriage in the Northern Territory of Australia (Smith-Vaughan *et al*., 2009) provided a rare opportunity to perform genomic comparisons between strains of the same serotype with differing invasive potential. In a previous study that performed such genomic comparisons, homologs of the TIGR4 ORFs *SP_1050 – SP_1053*, which includes the PezAT chromosomal TA system (*SP_1050* and *SP_1051*), were identified by CGH in highly virulent serotype 1 strains (1861 and 4496), but were absent from intermediately virulent and non-invasive serotype 1 strains (Harvey, 2006). *PezAT* is located within the PPI-1, downstream of the pneumococcal iron transport (*pit*/*pia*) locus (Brown *et al*., 2001). PPI-1 was defined in TIGR4 by a sudden drop in G+C content between sequence flanking the island and sequence immediately within the island, dropping from 40.1% to 31.5% at the 5' end and increasing from 28.8% to 39.4% at the 3' end, suggesting that this region was acquired through horizontal genetic transfer (Brown *et al*., 2001). The sequence between the PPI-1 ORFs homologous to *SP_1045* and *SP_1067* (*ftsW*) of TIGR4, has been shown to exhibit considerable variation between different strains (Brown *et al*., 2004; Croucher *et al*., 2009). When the variability of the 3' half of PPI-1 was first recognised, four differing genetic organisations were observed in four different serotypes; serotype 4 (TIGR4, ST205), serotype 17 (strain 142), serotype 2 (R6 – rough derivative of D39) and serotype 19F (G54) (Figure 3.1). Strain 142 had the shortest version of the region, virtually lacking any sequence between the homologs of *SP_1046* (*nplT*) and *SP_1064*. A 7-bp direct repeat was identified in TIGR4 flanking the 11.3-kb sequence and was present only in this strain, suggesting that the region may have been inserted into its current location by homologous recombination. One copy of the 7-bp sequence was found in strain 142 at the equivalent insertion site of the TIGR4-specific sequence. Further comparisons with R6 and G54 revealed that between the homologs of *SP_1054* and *SP_1067*, R6 was shown to contain 18 ORFs within 20.5 kb of sequence. Between the same TIGR4 homologs in G54, 15 ORFs were present, with eight of these absent in the equivalent region of R6. Of the four versions of the region shown in Figure 3.1, G54 and strain 142 lacked *pezAT*. Brown *et al*. (2004) also screened a number of strains representing a variety of serotypes for *pezAT* by PCR and Southern hybridisation analysis, finding that 30% of strains screened appeared to lack the system. Of particular note was the serotype 1 isolate, strain 40, included in the screen, which lacked *pezAT* in contrast to the highly virulent strains 1861 and 4496 that were *pezAT* positive. PCR data for strain 40 also showed that the homologs of TIGR4 ORFS *SP_1052*, *SP_1056*, *SP_1061*, *SP_1063* and *SP_1064* were absent. Figure 3.2 shows a sequence alignment from Croucher *et al*. (2009), of PPI-1 in strains D39 (serotype 2), TIGR4, 3-BS71 (serotype 3, ST180), G54 and ATCC 700669 (serotype 23F). In agreement with Brown *et al*. (2004), the alignment showed that the 5' half of PPI-1 was much more conserved between strains than the 3' half, which exhibited extensive sequence variability. This variability appeared to be particularly prominent between the cluster of Tn*5252*-related ORFs (*SP_1054* to *SP_1056*) and *ftsW*. The region of variability contained an alternative set of genes in each of the six aligned strains. The region containing *pezAT* appeared to be an additional pocket of variability, given its absence from 18-BS74 and G54. The region of Tn*5252*-related ORFs represented a third region of variability

NOTE:
This figure is included on page 75 of the print copy of the thesis held in the University of Adelaide Library.

Figure 3.1 Variations of the PPI-1 variable region reported by Brown *et al*. (2004)
(A) Comparison between the PPI-1 variable region organisation of TIGR4 (ii) and strain 142 (iii). The nucleotide sequence at the junction of the insertion of the PPI-1 variable region in TIGR4 (i) and strain 142 (ii). Boldface highlights the direct repeat sequences flanking the 11.3 kb TIGR4-specific sequence. Complementary nucleotide sequences flanking the insertion point in both strains are underlined. The organisation of the PPI-1 variable region in R6 (B) and G54 (C) are also shown. Pale grey ORFs are present in TIGR4, dark grey in G54 and R6 but not TIGR4, white ORFs are G54-specific, spotted ORFs are R6-specific. *FtsW* is indicated in black.

NOTE:
This figure is included on page 76 of the print copy of the thesis held in the University of Adelaide Library.

Figure 3.2 Alignment of the PPI-1 region from a selection of sequenced genomes, as reported by Croucher *et al.* (2009)
Homology between consecutive strains is indicated in red. Tn*5252*-associated sequence homology is indicated in pink. The *pit* locus is indicated with white boxes, ORFs associated with mobile genetic elements such as *pezAT* and Tn*5252*-associated ORFs such as a mobA domain protein (G54) are represented with pink or brown boxes. *FtsW* is represented by a green box.

within the 3' half of PPI-1 and displayed varying degrees of degeneration between strains due to recombination. Brown *et al*. (2004) and Croucher *et al*. (2009) also identified Tn*5252*-related sequence immediately upstream of *ftsW* in all strains with the exception of G54 (Figure 3.2). As Tn*5252*-related sequence was found to flank the primary region of hypervariable sequence, the authors hypothesised that this flanking sequence could have contributed to the sort of recombination required for the variability of this region Croucher *et al*. (2009). To date, PPI-1 comparisons have yet to identify differences between strains of the same serotype. Brown *et al*. (2004) included a selection of serotype 3 isolates representing all major STs within their screen for the presence of *pezAT*. However, all clones were found to possess *pezAT*, leading the authors to conclude that acquiring *pezAT* is a relatively stable event. Given that previous genetic comparisons between strains of this study have showed differences in the presence of *pezAT* and in the remaining content of the PPI-1 variable region, it was important to determine the ST of the tested strains in order to identify whether a correlation between ST and the organisation of PPI-1 existed.

3.2 Genetic relatedness of serotype one strains 1, 2, 3415, 5482, 1861 & 4496

3.2.1 ST of strains 1861 and 4496

As discussed in Section 1.5.2.1, Brueggemann *et al*. (2003) showed that serotype 1 strains could be divided into three lineages, where all strains within each lineage shared at least four alleles of seven MLST genes. Interestingly, the content of PPI-1 in the serotype 1 strains of this study could be divided into two groups by CGH and Southern blotting (Section 1.6.3). Therefore, it was important to assess the extent to which PPI-1 sequence correlated with genetic relatedness by determining the ST of each strain. Of the seven serotype 1 strains used in this study, the MSHR determined the ST of strains 1, 2 and 4. Strains 1 and 2 were ST304 and strain 4 was ST227 (Section 1.6.1). For strains 3415 and 5482, the MSHR sequenced six of seven MLST genes and assumed that they were ST227. For the purposes of this study, we confirmed that strains 3415 and 5482, along with strains 1, 2 and 4 were of lineage A (Figure 3.3). Since MLST had not previously been carried out on strains 1861 and 4496, STs were determined for these strains, as described in Section 2.10. Table 3.1 shows the results of the MLST. The novel ST, ST3079 was assigned to strain 1861 as it possessed the new *ddl* allele, 276. Strain 4496 was also assigned a new ST, ST3018, as its allelic profile had not previously been identified.

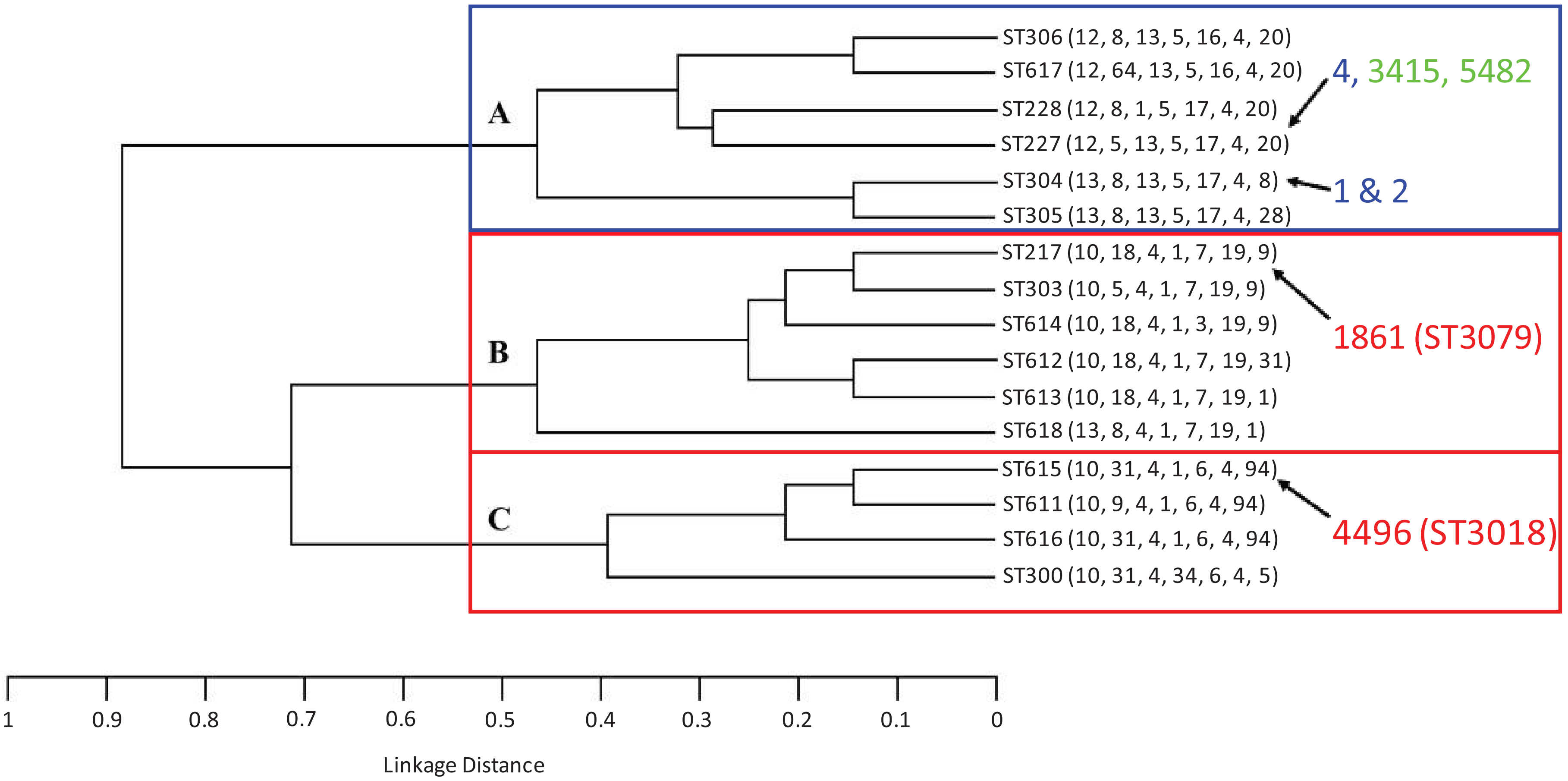
Table 3.1 Allelic profiles of strains 1861 and 4496

	aroE	**gdh**	**gki**	**recP**	**spi**	**xpt**	**ddl**	**ST**
1861	10	18	4	1	7	19	276*	3079#
4496	10	31	4	16	6	4	94	3018#

Alleles for each gene of each strain were determined as described in Section 2.10
* Novel allele assigned number after submission as described in Section 2.10
# Novel ST assigned after submission as described in Section 2.10

3.2.2 Relatedness of strains 1861 and 4496 to other serotype 1 strains

By comparing the allelic profiles of strains 1861 and 4496 with that of the STs in Brueggemann *et al*. (2003), the relatedness of each strain was determined (Figure 3.3). ST3079 was found to be most closely related to ST217 and ST613 of lineage B (Section 1.5.2.1), as their allelic profiles differed by only 1 of 7 alleles (*ddl*). ClustalW alignment (Section 2.6.1) of *ddl* alleles 1 (ST613), 9 (ST217) and 276 (ST3079) revealed ST3079 to be most closely related to ST217, as the *ddl* sequence of allele 276 differed by only 1 bp from allele 9 compared to 4 bp from allele 1. By contrast, ST3018 was most closely related to ST615 of lineage C, as their allelic profiles differed by only 1 of 7 alleles. Therefore, strain 1861 was found to be of lineage B, whereas strain 4496



A
B
C
ST306 (12, 8, 13, 5, 16, 4, 20)
ST617 (12, 64, 13, 5, 16, 4, 20)
ST228 (12, 8, 1, 5, 17, 4, 20)
ST227 (12, 5, 13, 5, 17, 4, 20)
ST304 (13, 8, 13, 5, 17, 4, 8)
ST305 (13, 8, 13, 5, 17, 4, 28)
ST217 (10, 18, 4, 1, 7, 19, 9)
ST303 (10, 5, 4, 1, 7, 19, 9)
ST614 (10, 18, 4, 1, 3, 19, 9)
ST612 (10, 18, 4, 1, 7, 19, 31)
ST613 (10, 18, 4, 1, 7, 19, 1)
ST618 (13, 8, 4, 1, 7, 19, 1)
ST615 (10, 31, 4, 1, 6, 4, 94)
ST611 (10, 9, 4, 1, 6, 4, 94)
ST616 (10, 31, 4, 1, 6, 4, 94)
ST300 (10, 31, 4, 34, 6, 4, 5)
4, 3415, 5482
1 & 2
1861 (ST3079)
4496 (ST3018)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Linkage Distance

Figure 3.3 Placement of serotype 1 isolates into lineages determined by Brueggemann *et al.*, 2003
Dendogram modified from Brueggemann *et al.*, 2003. A linkage distance of 0.14 or 0.29 indicates identical alleles at 6 of 7 or 5 of 7 loci, respectively. The STs or closely related STs of the non-invasive strains (blue), intermediately virulent strains (green) and highly virulent strains (red) are indicated above.

was of lineage C, indicating that the two highly virulent strains were of distinct genetic backgrounds as they shared only two of seven alleles, despite their similar virulence profiles (Figure 3.3). MLST data showed that the two alternative versions of PPI-1 content, predicted previously (Section 1.6.3), could be grouped into the lineage A strains and the lineage B and C strains. Therefore, divergence of the content of PPI-1 possibly occurred between the division of the lineage A strains from those of lineages B and C, and the division of the lineages B and C from each other.

3.3 Sequence and annotation of PPI-1 in strains 1, 2, 4, 3415, 5482, 1861 and 4496

3.3.1 Step-wise sequencing of the PPI-1 variable region

The PPI-1 variable region was sequenced in the serotype 1 strains 1, 2, 4, 3415, 5482, 1861 and 4496 using a selection of primers in Table 2.3. Primers *ao*, *t*, *af*, *an*, *ar*, *ag*, *ay*, *as*, *u*, *a*, *s*, *w*, *p*, *v*, *z*, *ab*, *aj*, *ak*, *aa*, *y*, *o*, *n*, *d*, *b*, *c*, *cb*, *q*, *g* and *br* were used to sequence the PPI-1 variable region in the lineage A strains. Primers *ao*, *t*, *af*, *an*, *ar*, *ag*, *ay*, *as*, *u*, *a*, *k*, *w*, *ap*, *al*, *ac*, *ad*, *j*, *at*, *au*, *aq*, *am*, *ax*, *ba*, *bc*, *be*, *bg*, *bi*, *bl*, *bn*, *bo*, *bq*, *bm*, *bk*, *bj*, *bh*, *bf*, *bd*, *bb*, *aw*, *av*, *cb*, *d*, *c*, *cb*, *q*, *g* and *br* were used to sequence the PPI-1 variable region in strains 1861 and 4496. Sequencing was carried out from primer *t* to *g*, which included sequences flanking the PPI-1 variable region. When sequencing the region in the lineage A strains, primers were designed in ORFs homologous *SP_1041 – SP_1046* and ORFs *SP_1067 – SP_1068*, which were known to be shared with TIGR4 (Section 1.6.3). The same primers were used to commence sequencing in strains 1861 and 4496, with additional primers designed in ORFs *SP_1047 – SP_1053*. PCR products were generated across regions of unknown sequence, using primers *w* and *d* in the first group of strains and primers *am* and *d* in the second group of strains. Step-wise sequencing was undertaken initially using primers within regions of known sequence to progressively enable new primers to be designed within previously unknown sequence until complete coverage of the region was achieved. Individual sequence files were edited for sequencing errors using the chromatograms and ClustalW alignments between strains of similar sequence (Section 2.6.1). Sequence files were assembled using the sequence assembly tool in DNAMAN (Section 2.6.1).

Following completion of the PPI-1 sequence between primers *t* and *g*, in all seven strains, the lineage A strains were found to share 99.94% sequence identity across 14.5-kb of sequence, while strains 1861 and 4496 shared 99.73% sequence identity over 24.5-kb of sequence. These alignments showed that a high degree of sequence identity exists between strains possessing the same organisation of PPI-1.

3.3.2 Annotation of the PPI-1 variable region in the sequenced strains

The translation overview feature of DNAMAN was used to identify ORFs in all six reading frames from the sequence generated in Section 3.3.1. ORFs encoding proteins of at least 80 amino acids from an ‘ATG’ start codon, were submitted to the BLASTp search engines of the NCBI and KEGG databases to identify previously annotated genes of similar sequence. ORFs that did not receive high-scoring hits or received hits to hypothetical proteins, were subsequently submitted to the HHpred search engine, which enables searches with greater sensitivity, as described in Section 2.6.2. The conserved domain (CD) feature of the NCBI database was also used to identify putative functional domains within amino acid sequences. Transcriptional and functional analysis of the PPI-1 variable region including the transcriptional structure and analysis of *in vivo* expression profiles will be addressed in Chapter 4.

3.3.2.1 Strains 1, 2, 4, 3415 and 5482

As expected from the high level of sequence conservation observed in the previous section, the translation overviews of each of the lineage A strains highlighted identical patterns of ORFs to that in Figure 3.4. Therefore, for the purposes of predicting and analysing proteins, the translation overview of strain 1 was assumed to be representative of the whole group. The translation overview identified 15 ORFs greater than 80 amino acids in length. 12 of these ORFs were located on the forward strand and 3 on the reverse strand. Following preliminary BLASTp searches, the 3 ORFs on the reverse strand were removed from the list of predicted coding regions due to a lack of homology to other annotated genes and their relatively short length.

Table 3.2 shows the final list of 12 predicted genes in the PPI-1 variable region, including putative functions predicted from a combination of BLASTp and HHpred analysis. ORF 1 could not be assigned a putative function as BLASTp and HHpred searches returned only hits with hypothetical proteins at a similar genomic position in other *S. pneumoniae* strains such as *SP_1041* of TIGR4. ORF 2 returned hits with only hypothetical proteins following BLASTp analysis, and did not return any high-scoring hits following HHpred analysis. ORF 3 returned high-scoring hits to ADP-ribosylglycohydrolase by both BLASTp and HHpred. The ADP-ribosylglycohydrolase of strains 3-BS71 (serotype 3, ST180) and CDC1873-00 (serotype 6A, ST376) possessed the greatest sequence identity to ORF 3. ADP-ribosylglycohydrolase is involved in post translational modification of proteins by ADP-ribosylation, which can be used to regulate the functions of target proteins (reviewed in Ludden [1994]). The most significant hit received from BLASTp searches for ORF 4 was the diacylglycerol kinase catalytic domain protein, also encoded by *SPCG_1025* of CGSP14 (serotype 14, ST15). HHpred searches also returned a high-scoring hit for a diacylglycerol kinase (Dgk) of *S. aureus*. Dgk is important in lipid-metabolism, because it phosphorylates

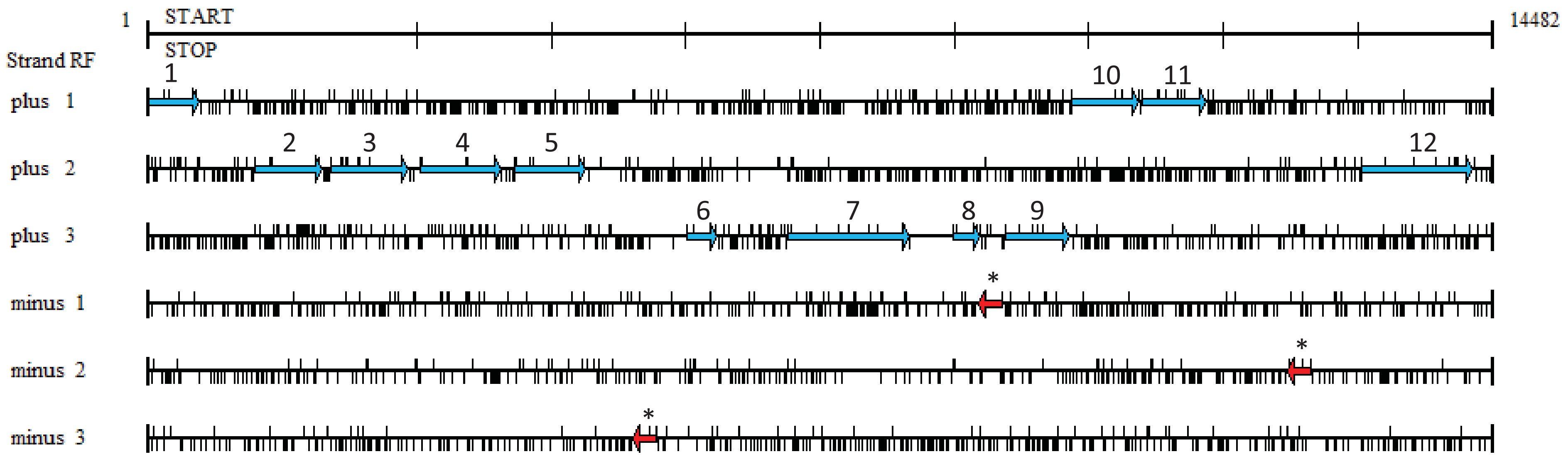


Figure 3.4 Translation overview of the PPI-1 variable and flanking regions in the lineage A strains
Translation overview was generated using DNAMAN (Section 2.6.1), from strain 1 *t* – *g* sequence generated in Section 3.3.1. ORFs containing the start codon 'ATG' and at least 80 amino acids in length are highlighted in blue (forward strand) and red (reverse strand). Vertical black bars indicate start codons (above) and stop codons (below). *Denotes ORFs >80 amino acids in size with no significant homology following BLASTp searches against the NCBI database. ORFs predicted to be true coding regions are numbered 1 – 12.

Table 3.2 Predicted ORFs from the PPI-1 variable region of the lineage A strains

ORF	Location (nt)	Annotation of homologous genes	Highest ranked hit(s) (BLASTp)#	Blastp Alignment*	Putative function or metabolic pathway
1	13 - 576	Hypothetical protein	TIGR4 (SP_1041), CGSP14 (SPCG_1021)	99%, Q1-188/188, S16-203/203	Unknown
2	1,151 - 1,907	Hypothetical protein	23-BS72	100%, Q1-252/252, S1-252/252	Unknown
3	1,967 - 2,819	ADP-ribosylglycohydrolase	3-BS71, CDC1873-00	99%, Q1-284/284, S1-284/284	ADP-ribosylation - post-translational modifications
4	2,933 - 3,815	Diacylglycerol kinase catalytic domain protein	CGSP14 (SPCG_1025)	100%, Q1-294/294, S18-311/311	Regulator of diacylglycerol metabolism
5	3,956 - 4,751	Neopullulanase	D39 (SPD_0927)	98%, Q1-265/265, S1-265/587	Oligosaccharide hydrolysis (truncated)
6	5,805 - 6,159	Chromosome segragation ATPase	D39 (SPD_0938)	98%, Q1-52/118, S4-55/198	Remnant of Tn*5252* conjugative transposon
7	6,894 - 8,208	Lantibiotic biosynthesis and transport system	SP23-BS72, CDC0288-04, CDC3059-06	99%, Q2-438/438, S1-437/707	Mersacidin modification and export (fragmented)
8	8,673 - 8,988	Lantibiotic biosynthesis and transport system	SP23-BS72, CDC0288-04, CDC3059-07	100%, Q1-105/105, S593-697/707	Mersacidin modification and export (fragmented)
9	9,237 - 9,963	Lantibiotic biosynthesis and transport system	CDC0288-04, CDC3059-06, ATCC 700669, JJA	99%, Q1-242/242, S1-242/242	Mersacidin immunity ABC transporter
10	9,958 - 10,693	Lantibiotic biosynthesis and transport system	23-BS72	100%, Q1-245/245, S1-245/245	Mersacidin immunity ABC transporter
11	10,708 - 11,416	Putative membrane protein	23-BS72	99%, Q1-236/236, S1-236/236	Mersacidin immunity ABC transporter
12	13,082 - 14,309	Cell division protein, FtsW	D39 (SPD_0952), CDC1873-00	99%, Q1-409/409, S1-409/409	Cell Division

*Indicates the amino acid sequence identity and alignment length of the total ORF legnth between the query sequence (Q [strain 1]) and the subject sequence# (S) diacylglycerol formed from the turnover of membrane phospholipids (reviewed in Sakane *et al*. [2007]). ORF 4 is homologous to *SP_1045* in TIGR4, which flanks the 5' end of the PPI-1 variable region (Section 3.1). ORF 5 received hits for a number of neopullulanases (EC 3.2.1.135), which are enzymes of the α-amylase family (Henrissat, 1991). The highest-scoring hit for ORF 5 was for neopullulanase (*nplT*), *SPD_0927* of D39 (serotype 2, ST545). Such enzymes have been shown to exhibit a unique substrate specificity for α-1,4 and α-1,6-glucosidic linkages (Takata *et al*., 1992). BLASTp searches showed that ORF 5 was unusual compared to genes of most similar sequence, due to the presence of a premature stop codon, leading to a truncated protein. Comparisons with neopullualnase of *Bacillus stearothermophilus*, which was the highest-ranked HHpred hit, showed that only the N-domain, residues 1 – 123, was complete in ORF 5, suggesting that the truncation had probably led to the loss of function of this gene in strain 1 (Hondoh *et al*., 2003). The highest-scoring hit returned for ORF 6 was for the putative chromosome segregation ATPase, *SPD_0938* of D39. Multiple alternative hits of significant scores suggested that this chromosome segregation ATPase probably originated from the conjugative transposon, Tn*5252*. Given that ORF 6 appeared to be a small fragment of a Tn*5252*-related gene and that no other complete Tn*5252* genes were identified in the near vicinity, it seemed unlikely that a functional conjugative transposon exists in the PPI-1 variable region of strain 1. ORFs 7 and 8 returned high-scoring hits that suggested the two ORFs were fragments of a single lantibiotic mersacidin transport system gene, *SPJ_0998* of strain JJA (serotype 14, ST66) and *SPN23F_09810* of ATCC 700669 (serotype 23F, ST81). Upon analysis of ORFs 7 and 8 (Figure 3.4), it was observed that the gap between the two ORFs was devoid of stop codons other than the ORF 7 stop codon, which seemed unusual given the density of stop codons present throughout other non-coding sequences of the PPI-1 variable region. Therefore, it was possible that either ORFs 7 and 8 are interrupted due to a sequencing error, or that the interruption was real due to a natural point mutation. To test for a sequencing error, the original sequencing chromatograms obtained from primer *bg* in all five lineage A strains were analysed to check for an editing error. However, the chromatograms were unambiguous for the substitution of 'C' for 'T', at position 8,208 in the *t* – *g* sequence, leading to a premature stop codon. Therefore, either the point mutation was real, or an error had been introduced by the high-fidelity polymerase during the original PCR amplification. Although it was unlikely that the polymerase had made an identical error in all five strains, the region was re-amplified from strain 1 genomic DNA (Section 2.9.6), and re-sequenced using primers *bg* and *v*. This confirmed that the point mutation was real, and that the fragmentation was real. *SPN23F_09810* and *SPJ_0998* possess two CDs, the first was of the C39-like protease superfamily (5 – 130 a.a.) and the second of the P-loop NTPase superfamily characteristic of the ATP-binding component of ABC transporters (CD database, NCBI). Such conserved domains have previously been associated with a common bacteriocin-processing endopeptidase, and function by cleaving the double-glycine leader peptide from the relevant bacteriocin precursor molecule, before exporting it from the cell. Interestingly, the peptidase domain was completely intact within ORF 7 and that if ORF 8 uses the alternative start codon 'TTG' (position 8,217) instead of the later 'ATG' codon, the P-loop NTPase superfamily CD was complete within ORF 8. Whilst it seemed unlikely that ORFs 7 and 8 could produce a functional protein, it seemed fortuitous that the premature stop codon of ORF 7 would neatly divide the original gene between the two functional domains. Therefore, it might be possible that the two separate ORFs could still allow the translation of a functional bacteriocin-processing endopetidase as two separate peptides. High-scoring hits for ORF 9 included *SPN23F_09830*, which is a component of the lantibiotic mersacidin biosynthesis and transport system. ORF 9, like ORF8, was predicted to contain the P-loop NTPase superfamily CD, which suggested that it was a component of another transporter. Highest-scoring hits returned for ORF 10 were for the hypothetical proteins *SPN23F_09840* and *SPJ_1001* and those returned for ORF11 were *SPN23F_09850* and *SPJ_1002*. Other high-scoring hits suggested that these ORFs encode membrane-spanning components of an ABC transporter. Therefore, ORFs 9 – 11, encode a complete ABC transporter, probably involved in immunity against the mersacidin lantibiotic (Draper *et al*., 2008). Finally, all significant hits received for ORF 12 were for the cell division protein, FtsW, homologous to *SP_1067* in TIGR4, which flanks the 3' end of the variable region of PPI-1 (Section 3.1).

3.3.2.2 1861 & 4496

The translation overview in Figure 3.5 showed that the PPI-1 variable region in strains 1861 and 4496 was predicted to contain 27 ORFs encoding proteins greater than 80 amino acids in length. Since both strains 1861 and 4496 possessed identical patterns of ORFs across all six reading frames, only strain 1861 was used for detailed analysis of the ORFs in this region. However, of the 27 ORFs in Figure 3.5, two were predicted to not encode a polypeptide as BLASTp searches failed to identify homology to a previously annotated gene. In addition, the sequence of these ORFs overlapped with longer ORFs that were predicted to encode previously identified products. Of the remaining 25 ORFs, listed in Table 3.3, ORFs 1 to 4, located upstream of the PPI-1 variable region, were identical to that of the lineage A strains and were discussed in Section 3.3.2.1. ORFs 5 and 6 returned high-scoring hits for *nplT*, *SPD_0927*, in D39. ORF 5 aligned with the first 132 residues of *SPD_0927* and ORF 6 aligned with residues 122 – 579 of the 579 amino acid *SPD_0927* (Table 3.3). However, BLASTn searches returned high-scoring hits for the pseudogenes *SP_1046* and *SPP_1049* of TIGR4 and P1031 (serotype 1, ST303), respectively, suggesting that during the annotation of the TIGR4 and P1031 genomes, *SP_1046* and *SPP_1049* were predicted

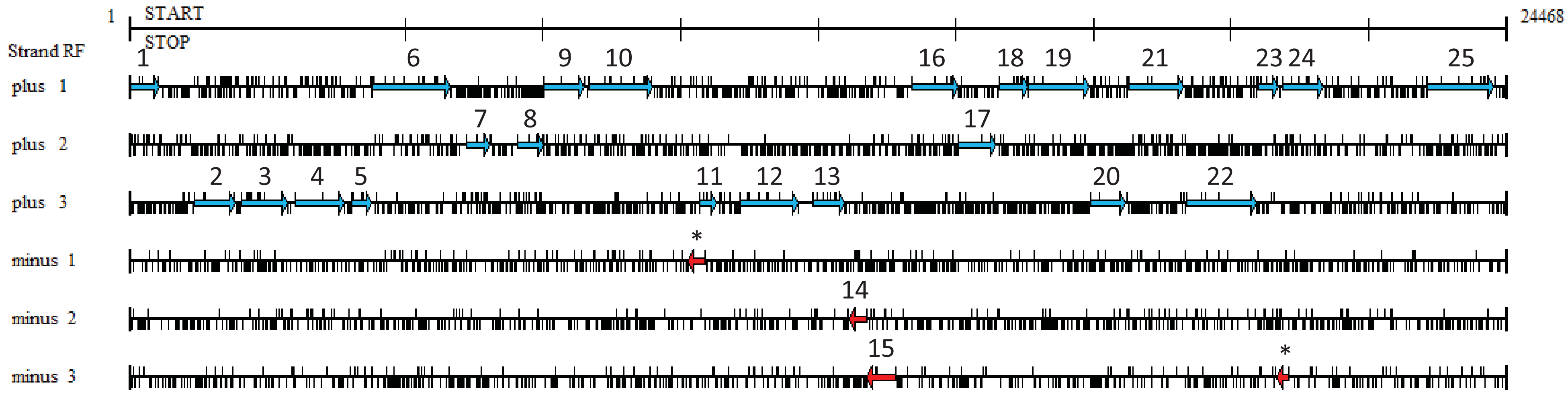


Figure 3.5 Translation overview of the PPI-1 variable and flanking regions in strains 1861 & 4496
Translation overview was generated using DNAMAN (Section 2.6.1), from strain 1861 *t* – *g* sequence generated in Section 3.3.1. ORFs containing the start codon ‘ATG’ and at least 80 amino acids in length are highlighted in blue (forward) and red (reverse). Vertical black bars indicate start codons (above) and stop codons (below). *Denotes ORFs >80 amino acids in size with no significant homology following BLASTp searches against the NCBI database. ORFs predicted to be true coding regions are numbered 1 – 25.

Table 3.3 Predicted ORFs from the PPI-1 variable region of strains 1861 & 4496

ORF	Location (nt)	Annotation of homologous genes	Highest ranked hit(s) (BLASTp)$^{\#}$	BLASTp alignment*	Putative function or metabolic pathway
1	13 - 577	Hypothetical protein	TIGR4 (SP_1041), CGSP14 (SPCG_1021)	100%, Q1-188/188, S16-203/203	Unknown
2	1,152 - 1,907	Hypothetical protein	TIGR4 (SP_1043), P1031 (SP_1046)	100%, Q1-252/252, S1-252/252	Unknown
3	1,968 - 2,819	ADP-ribosylglycohydrolase	D39 (SPD_0925), SP14-BS69	100%, Q1-284/284, S1-284/284	ADP-ribosylation - post-translational modifications
4	2,934 - 3,815	Diacylglycerol kinase catalytic domain protein	P1031 (SPP_1048)	100%, Q1-294/294, SQ1-294/294	Regulator of diacylglycerol metabolism
5	3,957 - 4,352	Neopullulanase	D39 (SPD_0927)	97%, Q1-132/132, S1-132/587	Oligosaccharide hydrolysis (fragmented)
6	4,312 - 5,715	Neopullulanase	Hungary 19A (SPH_1147)	99%, Q11-468/468, S122-579/579	Oligosaccharide hydrolysis (fragmented)
7	5,993 - 6,445	Hypothetical protein	P1031 (SPP_1051)	100%, Q1-151/151, S1-151/151	Unknown
8	6,893 - 7,367	Addiction system antitoxin, *pezA*	D39 (SPD_0930)	100%, Q1-158/158, S1-158/158	Addiction system
9	7369 - 8,128	Addiction system toxin, *pezT*	P1031 (SPP_1054)	98%, Q1-253/253, S1-253/253	Addiction system
10	8,158 - 9,312	Putative phosphoesterase	P1031 (SPP_1055)	100%, Q1-385/385, S1-385/385	Unknown
11	10,134 - 10,490	Tn*5252* , ORF 10	P1031 (SPP_1056)	100%, Q1-119/119, S1-119/119	Unknown
12	10,854 - 11,897	Tn*5252* , relaxase	3-BS71	96%, Q1-348/348, S1-348/379	Remnant of Tn*5252* conjugative transposon
13	12,132 - 12,716	Tn*5252* , relaxase	23-BS72	97%, 1-195/195, S427-621/621	Remnant of Tn*5252* conjugative transposon
14	12,764 - 13,097	Transcriptional regulator Rgg/GadR/MutR family	G54 (SPG_0976)	96%, Q1-111/111, S173-283/283	Non-functional due to fragmentation
15	13,104 - 13,626	Transcriptional regulator Rgg/GadR/MutR family	11-BS70, MLV-016	100%, Q5-173/173, S1-169/283	Non-functional due to fragmentation
16	13,912 - 14,739	3-hydroxyisobutyrate dehydrogenase	11-BS70, MLV-016	99%, Q1-276/276, S14-289/289	Valine, Leucine and Isoleucine degradation
17	14,729 - 15,448	Hypothetical protein	P1031 (SPP_1063), G54 (SPG_0978)	100%, Q1-240/240, S1-240/240	Unknown
18	15,448 - 15,976	Prephenate dehydratase	P1031 (SPP_1064)	100%, Q1-176/176, S1-176/176	L-phenylalanine biosynthesis
19	15,982 - 17,073	Hypothetical protein	P1031 (SPP_1065), MLV-016, 11-BS70	100%, Q1-364/364, S1-364/364	Unknown
20	17,079 - 17,747	Hypothetical protein	P1031 (SPP_1066)	100%, Q1-223/223, S1-223/223	Unknown
21	17,752 - 18,777	UDP-glucose, 4-epimerase	P1031 (SPP_1067)	100%, Q1-342/342, S1-342/342	Interconversion of UDP-glucose and UDP-galactose
22	18,804 - 20,045	Biotin carboxylase	P1031 (SPP_1068)	99%, Q1-414/414, S1-414/414	First committed step of fatty acid biosynthesis
23	20,050 - 20,455	Transporter, major facilitator superfamily	G54 (SPG_0985)	100%, Q1-135/135, S1-135/135	Putative importer (Lactose or glycerol -3-phosphate)
24	20,509 - 21,259	Transporter, major facilitator superfamily	G54 (SPG_0984), 11-BS70	100%, Q1-250/250, S15-264/264	Putative importer (Lactose or glycerol -3-phosphate)
25	23,068 - 24,295	Cell division protein, *ftsW*	18-BS74, MLV-016, CDC1087-00	100%, Q1-409/409, S1-409/409	Cell division

*Indicates the amino acid sequence identity and alignment length of the total ORF legnth between the query sequence (Q [strain 1]) and the subject sequence$^{\#}$ (S) to be non-functional. Pair-wise alignment of ORFs 5 and 6 with *SPD_0927* identified a 2-bp frame-shift deletion between positions 4,348 and 4,349 in strain 1861, which was also found to be present in P1031. As for ORFs 7 and 8 in Section 3.3.2.1, the 2-bp deletion responsible for the fragmentation of *nplT* was checked by re-sequencing and was found to be real. CD searches of ORFs 5 and 6 showed that the two functional CDs, the pullulan-degrading enzymes N-terminal domain and α-amylase superfamily catalytic domain of *SPD_0927*, were each complete in ORFs 5 and 6 respectively of 1861. When the structure of the *Bacillus stearothermophilus* neopullulanase was determined (Hondoh *et al*., 2003), the N-domain (Section 3.3.2.1) was found to be responsible for the specificity of neopullulanases, separating them from other members of the α-amylase superfamily. Interestingly, Hondoh *et al*. (2003) showed that the N-domain extended from residues 1 – 123, which closely matches the length of ORF 5 (a.a. 1 – 132) in strain 1861, indicating that the entire N-domain is probably intact within ORF 5. On the other hand, ORF 6 appeared to possess the three remaining neopullulanase domains. Whilst initially it seemed that fragmentation into ORFs 5 and 6 would lead to a non-functional protein, it was intriguing that the site of fragmentation is approximately between two domains. Given that theoretically all domains of neopullulanase remained uninterrupted, it is possible that the two final proteins encoded by ORFs 5 and 6 could together maintain similar function to that of *nplT* in D39. A putative function could not be determined for ORF 7 using both BLASTp and HHpred searches, as neither search engine returned hits to proteins of known function. However, high-scoring hits for ORFs 8 and 9 indicated that they encoded the *pezAT* TA system. It seemed probable that the *pezAT* system was functional in strain 1861 as alignments confirmed that the full-length system was present and that the systems' promoter was intact, as described in Khoo *et al*. (2007). Homologs of ORF 10 where found to exist as two fragments in strains TIGR4 (*SP_1052* and *SP_1053*), D39 (*SPD_0932* and

SPD_0933) and JJA (*SPJ_0989* and *SPJ_0990*), but as a single ORF in strains P1031 (*SPP_1055*) and 70585 (Serotype 5, ST289 [*SP70585_1130*]). Hits returned for ORF 10 from *S. pneumoniae* genomes were annotated as hypothetical proteins, with the exception of CGSP14 (*SPCG_1031*), which was annotated as a putative phosphoesterase. The highest-scoring hit received for ORF 10 was *SPP_1055* from P1031, with which it shared 100% sequence identity across the full-length ORF. CD search results showed the presence of a Rad50 family catalytic domain, which is similar to the ATP-binding component of an ABC transporter, but is usually not associated with a membrane-spanning component (Hyde *et al*., 1990). HHpred searches suggested that ORF 10 encodes an ATP-binding protein, as it returned hits for short regions of high-scoring homology to various ATP-binding proteins. Short regions of lower-scoring hits from HHpred searches were returned for metal-dependent phosphoesterases, which was in agreement with the annotation for the CGSP14 homolog. However, due to combined alignment coverage representing only 50% of ORF 10, it was impossible to predict its function. ORFs 11 to 14 returned high-scoring hits for a number of Tn*5252* transposase-related ORFs. The highest-scoring hit returned for ORF 11 was the hypothetical protein, *SPP_1056* of P1031, which showed 100% sequence identity across the full-length ORF. The D39 hit for ORF 11 was annotated as orf 10 protein of the conjugative transposon, Tn*5252*. The lengths of hits returned for ORF 11 were consistent, as most were approximately 119 amino acids in length, suggesting that the full-length gene is present in strain 1861. BLASTp searches of ORFs 12 and 13 returned high-scoring hits for Tn*5252*-related relaxase, but in contrast to ORF 11, were of inconsistent lengths. For example, a single ORF in strains 23-BS72 (serotype 23, ST37) and CDC3059-06 (serotype 19A, ST199) was homologous to the combined sequence of ORFs 12 and 13 in 1861, which was in contrast to D39 (*SPD_0937* and *SPD_0938*) and 3-BS71 that were found to possess the two-ORF version, as found in strain 1861. Many high-scoring hits returned for ORFs 12 and 13 were from *Streptococcus suis*, which in some cases returned higher scores than other pneumococcal strains, suggesting that horizontal transfer of Tn*5252*-related genes has probably occurred frequently between the two species. High-scoring hits returned from BLASTp searches identified ORFs 14 and 15 as two fragments of a putative transcriptional regulator from the Rgg/GadR/MutR family. Following pair-wise alignment between the combined nucleotide sequence of ORFs 14 and 15 with the homologous G54 ORF, *SPG_0976*, a single 'G' insertion at position 13,112 (*t* – *g* PCR sequence, strain 1861) was identified as the cause of the frame-shift and fragmentation. The 'G' at position 13,112 was verified by re-sequencing, and was found to be real. The insertion divided the gene into the 333-bp ORF 14 and the 522-bp ORF 15, probably rendering the encoded protein non-functional. The fragmented version of this gene appeared to be quite unusual in *S. pneumoniae*, being present in only P1031 (*SPP_1060*), despite the full-length gene being present in most strains with completed genomes, with the exception of Hungary 19A (serotype 19A, ST168), Taiwan 19F (serotype 19F, ST236) and TIGR4. It was interesting to note that an identical point mutation was identified in the Rgg/GadR/MutR family transcriptional regulator, *ropB*, of M1T1 *Streptococcus pyogenes*, which has been shown to be a positive regulator of the virulence-associated protease, SpeB (Hollands, *et al*., 2008). The point mutation was shown to be responsible for loss of SpeB expression, which could be rescued by replacing the truncated *ropB* with the full-length gene. Therefore, the products of ORFs 14 and 15 in strain 1861 are probably also non-functional. BLASTp searches of ORF 16 returned high-scoring hits for 3-hydroxyisobutyrate dehydrogenase (3HIBDH) of 11-BS70 and the NAD-binding domain of 6-phosphogluconate dehydrogenase of P1031 (*SPP_1062*) and G54 (*SPG_0977*), suggesting some discrepancy between strains in the annotation of these ORFs in the KEGG and NCBI databases. However, the highest-scoring hits returned from the HHpred search engine were primarily annotated as 3HIBDH, suggesting that this was the most likely enzyme encoded by ORF 16. 3HIBDH (EC 1.1.1.31) catalyses the rate-limiting step in the degradation of branched-chain amino acids (BCAAs), by catalysing the conversion of 3-hydroxyisobutyrate into methylmanoate (Robinson & Coon, 1957). Interestingly, the substrate of 3HIBDH has been suggested to be an 'inter-organ' metabolite (Letto *et al*., 1986). BLASTp searches of ORFs 17 to 20 returned hits for hypothetical proteins of strains 11-BS70 (serotype 11, ST62), MLV-016 (serotype 11A, ST62) and P1031 without returning significant-scoring hits to any other ORFs in either the KEGG or NCBI databases. HHpred was unable to return high-scoring hits for ORF 17, but moderate scores were returned for hits to the nucleotide-binding domain of a nucleotidyl transferase in *H. influenzae*. ORF 18 returned high-scoring hits for prephenate dehydratase (PDT), following HHpred searches. PDT (EC 4.2.1.51) is known be an important regulatory enzyme in phenylalanine biosynthesis, converting prephenate to phenylpyruvate (Cotton & Gibson, 1965). HHpred searches were unable to attribute a putative function to either of ORFs 19 and 20. High-scoring hits returned from BLASTp and HHpred searches of ORF 21 were for UDP-glucose 4-epimerase (*galE*) (EC 5.1.3.2), *SPP_1067* of P1031, which catalyses the interconversion of UDP-glucose and UDP-galactose (Wilson & Hogness, 1969). ORF 22 returned high-scoring hits to biotin carboxylases, in particular the putative biotin carboxylase, *SPP_1068* of P1031. Biotin carboxylase (EC 6.4.1.2) is an important metabolic enzyme, as it catalyses the first committed step in fatty acid biosynthesis (reviewed in Cronan [2002]). BLASTp and HHpred searches returned high-scoring hits for ORFs 23 and 24 to a fragmented transporter of the major facilitator superfamily. In contrast, *SPD_0950* of D39 was encoded by sequence equivalent to both ORFs 23 and 24 of 1861. However, the same fragmented version of the transporter was also found in G54 (*SPG_0984* and *SPG_0985*) and 11-BS70, and was annotated as a pseudogene in P1031 (*SPP_1069*).

Pair-wise alignment of ORFs 23 and 24 with *SPD_0950*, identified the single nucleotide substitution at position 20,455 to ‘T’, that was responsible for the introduction of a premature stop codon, leading to fragmentation of the gene. Whilst some hits suggested that ORFs 23 and 24 could encode the permease component of an ABC transporter, the lack of other ABC transporter components suggested that ORFs 23 and 24 were more likely to encode a transporter from the major facilitator superfamily (reviewed in Pao *et al*. [1998]). High-scoring HHpred hits suggested that ORFs 23 and 24 could have a role in either the import of lactose or glycerol-3-phosphate, due to homology with these transporters in *E coli*. A high-scoring hit was also returned for EMRD of *E coli*, the multidrug resistance efflux pump. It was unclear whether the fragmentation of the transporter into ORFs 23 and 24 would still allow a functional protein to be produced. ORF 25 returned high-scoring hits to the cell division protein, FtsW, a homolog of *SP_1067* in TIGR4, which flanks the 3’ end of PPI-1.

3.4 PPI-1 variable region sequence comparisons

3.4.1 Comparison of PPI-1 between serotype 1 strains

The ACT was utilised to characterise the extent and relative location of homologous and non-homologous regions across the PPI-1 variable region between strains possessing different versions of the island, as described in Section 2.6.1. The comparison was performed between strains 1 and 1861, which were representative of the two groups of strains sequenced in Section 3.3.1. Following alignment of *t – g* sequence between strains 1 and 1861 (Figure 3.6), 4 regions of homologous sequence, two regions of 1861 sequence absent from strain 1 and one region of sequence divergence were identified between the two strains, as summarised in Tables 3.4 and 3.5. The position of the regions of homologous sequence, sequence divergence and

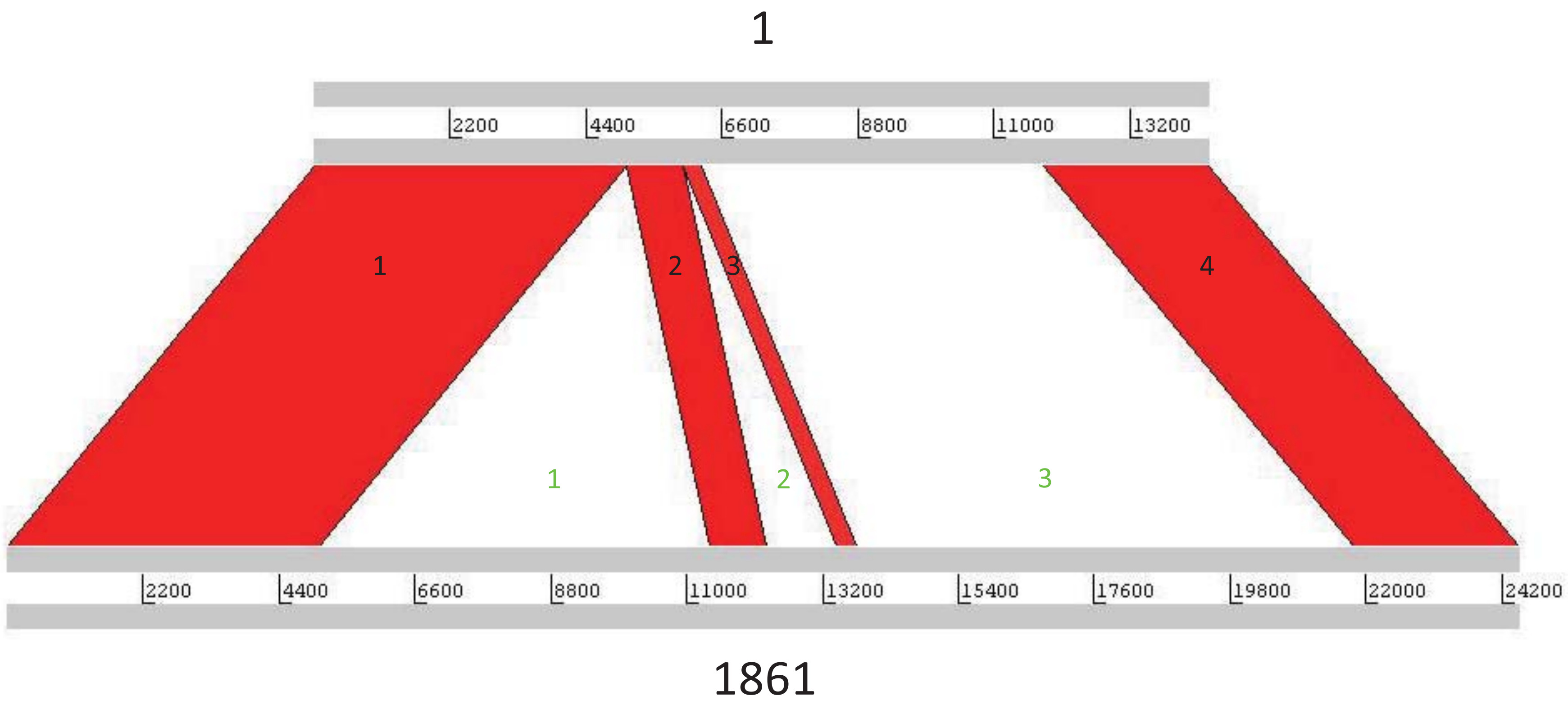


Figure 3.6 Regions of sequence homology, deleted sequence and sequence divergence between t – g sequence of strains 1 and 1861
Alignment between *t* – *g* sequence of strains 1 and 1861 was undertaken using the ACT as described in section 2.6.1. Sequence sharing at least 80% identity over at least 40 bp between the two strains is indicated in red. Regions of homologous sequence are numbered in black and regions of non-homologous sequence in green (Tables 3.4 & 3.5).

deletions relative to ORFs identified in Section 3.3.2, were determined with reference to the ORF positions in Tables 3.2 and 3.3.

Table 3.4 Homologous regions identified in Figure 3.6

Region	Coordinates (1)	Coordinates (1861)	Length	Identity
1	1 – 5,060	1 – 5,060	5.1 kb	99%
2	5,061 – 5,956	11,387 – 12,282	900 bp	95%
3	5,957 – 6,268	13,440 – 13,753	300 bp	90%
4	11,810 – 14,482	21,790 – 24,468	2.7 kb	98%

The coordinates of homologous regions were determined using the ACT generated alignment in Figure 3.6

Table 3.5 Non-homologous regions identified in Figure 3.6

Region	Coordinates (1)	Coordinates (1861)	Length
1	5,060 – 5,061	5,061 – 11,386	6.3 kb (1861)
2	5,956 – 5,957	12,283 – 13,439	1.5 kb (1861)
3	6,269 – 11,810	13,754 – 21,789	5.5 kb (1) 8 kb (1861)

The coordinates of non-homologous regions were determined using the ACT generated alignment in Figure 3.6

As discussed in Section 3.1, sequence flanking the PPI-1 variable region was expected to be shared between strains 1 and 1861, as shown in Figure 3.6 (homologous regions 1 and 4). Homologous region 1, which includes sequence upstream of the PPI-1 variable region, encodes ORFs 1 and up to 748 bp into ORF 6 of 1861 and up to 300 bp downstream of the ORF 5 stop codon in strain 1 (Table 3.4). Homologous region 4 was located downstream of the PPI-1 variable region, which encodes FtsW, ORF 25 of 1861 and ORF 12 of strain 1. Immediately downstream of homologous region 1, 6.3-kb of sequence (region 1 of non-homologous sequence [Figure 3.6]), was found to be present in only 1861, indicating that the region had either been deleted between position 5,060 and 5,061 in strain 1 or had been inserted into this position in strain 1861. This region was found to encode the last 654 bp of ORF 6 and 532 bp of ORF 12 of strain 1861. It seemed most likely that the region had been lost from strain 1 as the deletion had resulted in truncation of *nplT* (ORF 5) and the loss of a number of ORFs from the region of the Tn*5252*-related sequence that were present in 1861. Homologous regions 2 and 3 (Figure 3.6) are two small regions, both encoding ORF 6 in strain 1, which extends from 151 bp before the end of region 2 to 202 bp into region 3. Given that ORF 6 of strain 1 extended across two regions of homologous sequence, it was not surprising that this ORF was not present in 1861. Instead region 2 of the homologous sequence in 1861 encoded the last 510 bp of ORF 12 and the first 151 bp of ORF 13 and region 3 encoded the last 186 bp of ORF 15. 1.5 kb of sequence present in only 1861 (non-homologous region 2) was identified between homologous regions 2 and 3 (Figure 3.6). Non-homologous region 2 includes ORFs 14 and up to 335 bp into ORF 15, which encode the fragmented putative Rgg/GadR/MutR family transcriptional regulator. Again, it seemed most likely that the putative transcriptional regulator had been lost from strain 1 between positions 5,956 and 5,957, rather than inserted into 1861, as the last 186 bp of ORF 15 sequence was retained in strain 1. Between regions 3 and 4 of homologous sequence was region 3 of non-homologous sequence, which was found to contain a large region of sequence divergence between the two strains; 5.5 kb in strain 1 and 8 kb in strain 1861. In 1861 this region encoded ORFs 16 to 24 and in strain 1 the region encoded ORFs 7 to 11. The region of sequence divergence was found to encode a series of genes of seemingly unrelated functions, which indicated that the two regions were probably acquired from different sources.

3.4.2 The PPI-1 variable region in a variety of *S. pneumoniae* strains and serotypes

In order to further characterise the PPI-1 variable region, it was decided to compare the organisation of the region between a larger number of strains from serotypes other than serotype 1. As discussed in Section 3.1, the region has previously been analysed in a number of strains including TIGR4, D39, ATCC 700669, and G54 (Brown *et al*., 2004; Croucher *et al*., 2009). Since a larger number of *S. pneumoniae* genomes are now publicly available, it was possible to perform detailed sequence comparisons *in silico* between the serotype 1 strains sequenced in Section 3.3.1 and a selection of other strains. The PPI-1 variable region of a number of strains was aligned with PPI-1 of strains 1 and 1861, and selection of those representing strains that were highly homologous, moderately homologous and largely non-homologous to PPI-1 of strains 1 and 1861 are shown in Figure 3.6. Such homology ranged from a small number of individual nucleotide differences, to a combination of homologous and non-homologous regions, to only a small number of short regions of homology. Having identified a pattern of sequence similarity and dissimilarity of the PPI-1 variable region, a diagnostic/taxonomic approach using PCR and restriction digest patterns was used to survey a much larger selection of strains.

3.4.2.1 Alignment between the PPI-1 variable region in strains 1861 and 1 against a selection of *S. pneumoniae* genome sequences using the ACT

The ACT was used to align the *t* – *g* sequence of ATCC 700669 (serotype 23F, ST81), P1031 (serotype 1, ST303), INV104B (serotype 1, ST227), D39 (serotype 2, ST595), Hungary 19A (serotype 19A, ST63) and G54 (serotype 19F, ST63). Figure 3.7 illustrates regions of homology between strain 1 and the aligned strains. Figure 3.7a shows that the PPI-1 variable region in the serotype 1 strain, INV104B, shares the greatest sequence identity with strain 1. The high degree of conservation between strain 1 and INV104B was expected given that the STs of both strains are within lineage A (Section 3.2.2). Strain ATCC 700669 (Figure 3.7b) was the next most homologous strain following INV104B. Of particular note was that ATCC 700669 possessed the region encoding the lantibiotic mersacidin system identified in Section 3.3.2.1, which was located within a 7.2 kb region of homologous sequence (6,269 – 14,482 [strain 1]), sharing 99% sequence identity between the two strains. Two regions of ATCC 700669 sequence missing from strain 1 were identified in Figure 3.7b. It was interesting to note that these two regions of missing sequence were highly similar to non-homologous

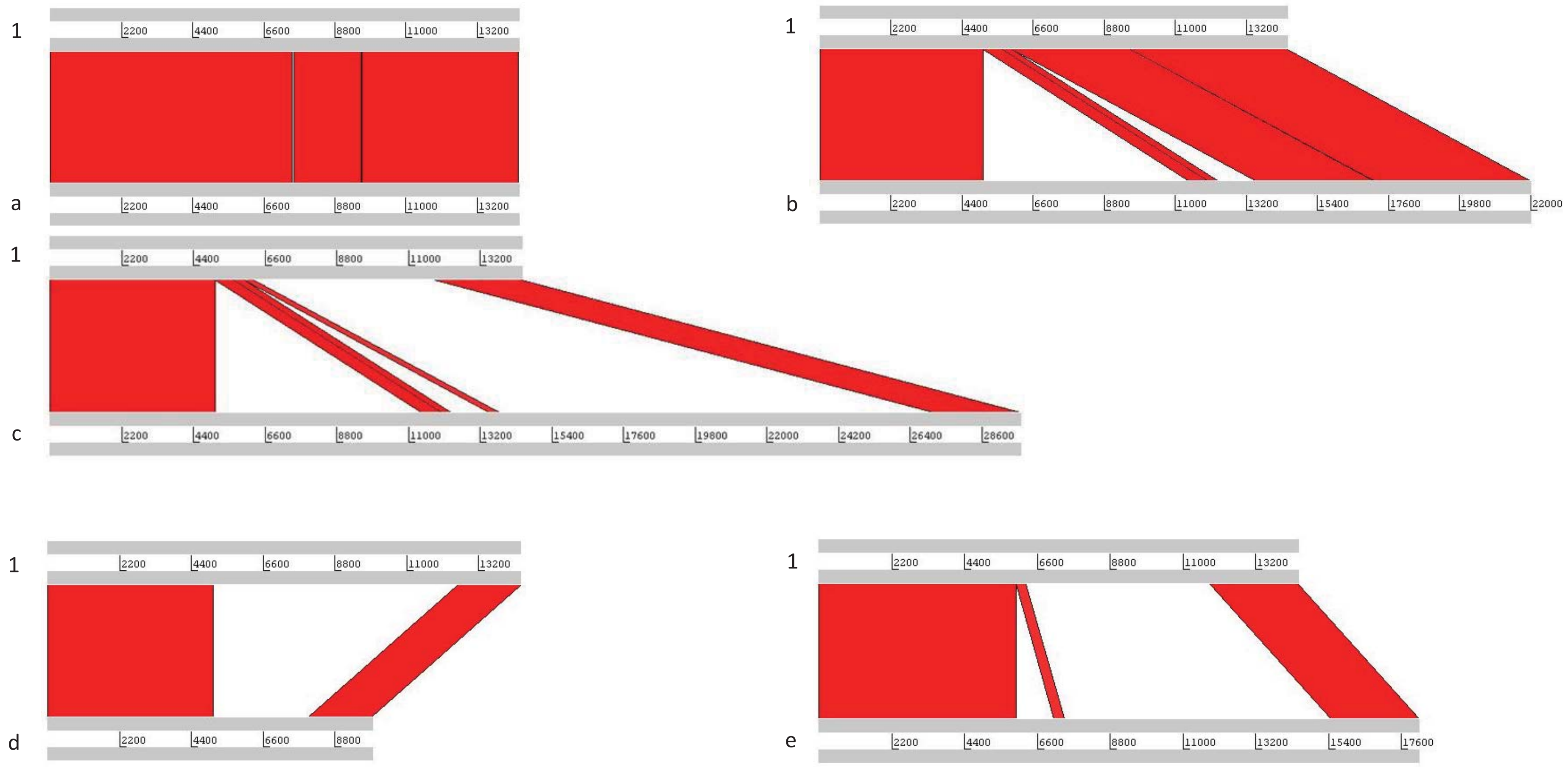


Figure 3.7 ACT alignments between the PPI-1 variable region of strain 1 and strains INV104B, ATCC 700669, D39, Hungary 19A and G54
Alignments were generated using the ACT between *t* – *g* sequence of strains 1861 (Section 3.3.1) and the *t* – *g* of strains INV104B (a), ATCC 700669 (b), D39 (c), Hungary 19A (d), G54 (e), as described in Section 2.6.1. Red indicates alignment of at least 80% identity over greater than 40bp. *t* – *g* sequence of strains ATCC 700669, D39, Hungary 19A and G54 were obtained from the KEGG database (Section 2.6.2) and INV104B sequence was obtained from http://www.sanger.ac.uk/Projects/S_pneumoniae.

regions 1 and 2 in Figure 3.6. Therefore, the primary differences identified between ATCC 700669 and strain 1 were that strain 1 lacked *pezAT* and the putative Rgg/GadR/MutR family transcriptional regulator and that strain 1 possessed a truncated *nplT* and a more degenerated Tn*5252*-related region. On the other hand, D39 was found to be quite different to strain 1 (Figure 3.7c), with only four relatively short regions of homologous sequence shared in the two strains. The homologous regions included two that flanked the PPI-1 variable region (1 – 5,060 [99% identity] and 11,809 – 14,482 [99% identity]) and two that flanked the putative Rgg/GadR/MutR family transcriptional regulator encoded by D39 (5,061 – 5,956 [95% identity] and 5,957 – 6,268 [91% identity], [strain 1]). Interestingly, the homologous regions that flanked the Rgg/GadR/MutR family transcriptional regulator corresponded to regions 2 and 3 in the strain 1 alignment with strain 1861 in Figure 3.6. Similar to ATCC 700669, D39 also possesses full-length *nplT*, the putative Rgg/GadR/MutR family transcriptional regulator and *pezAT*, which were all absent from strain 1. Between positions 6,269 and 11,810, strain 1 possesses the 5.5-kb region identified in Section 3.4.1, which contains the genes associated with a lantibiotic mersacidin transport system not present in D39. Instead, D39 possesses an alternative 13-kb region encoding a number of putative metabolic enzymes. The comparison between strain 1 and Hungary 19A strain (Figure 3.7d), showed that the PPI-1 variable region was very small in Hungary 19A, only 4.9-kb in total length. However, similar to alignments performed with ATCC 700669, D39 and 1861, the first region of homologous sequence ended at position 5,060. As a result Hungary 19A was found to possess a full-length version of *nplT* (*SPH_1147*) that was 579 amino acids in length. Hungary 19A lacked the 5.5-kb region containing the lantibiotic mersacidin transport system present in strain 1, and instead encoded a short 3.6-kb region containing the end of *SPH_1147* and ORFs *SPH_1148 – SPH_1151*. Like strain 1, Hungary 19A lacked *pezAT* and the putative Rgg/GadR/MutR family transcriptional regulator. G54 was found to possess three regions of sequence homologous to strain 1 (Figure 3.7d). Of particular interest was the first region of homologous sequence, which was 896 bp longer (1 – 5,956 [strain 1]) than the first region of homologous sequence in the ATCC 700669, D39 and 1861 alignments. Within the first region of homologous sequence, G54 contains the ORFs *SPG_0967 – SPG_0974*, which includes a short *nplT* (*SPG_0972*). G54 *nplT* was 100 amino acids longer than strain 1, but still only 375 amino acids in length, compared to the more common length of between 580 to 600 amino acids that was found in most other pneumococcal strains. Although lacking *pezAT*, as does strain 1, G54 possesses the putative Rgg/GadR/MutR family transcriptional regulator, which appeared to be lost from strain 1 between positions 5,956 and 5,957. A small 313-bp region (5,957 – 6,268 [strain 1]) of homologous sequence was found following the suspected deletion of the putative transcriptional regulator. Similar to the D39 (Figure 3.7c) and 1861 (Figure 3.6) alignments, the sequence found in strain 1 between 6,269 and 11,810 was not present in G54, which instead possessed an alternative 8-kb region of sequence containing ORFs *SPG_0977 – SPG_0985* (Figure 3.7e).

Figure 3.8 shows the alignments that were performed between the PPI-1 variable region of 1861 and strains P1031, ATCC 700669, D39, G54 and Hungary 19A. Strain P1031 was found to possess sequence identity of at least 99% across the full length alignment (Figure 3.8a). The high degree of conservation between P1031 and 1861 was not surprising given that P1031 (ST303) is of the same lineage as 1861 (Section 3.2.2). Four regions of homologous sequence were identified between 1861 and ATCC 700669 (Figure 3.8b), which included an initial 10-kb sequence with 98% sequence identity. In 1861, the first region encoded ORFs 1 – 9 (from primer '*t*'), which corresponded to ORFs *SPN23F_09610 – SPN23F_09720* and showed that ATCC 700669 possessed *pezAT* (as discussed in Section 3.1). However, unlike 1861, the companion putative

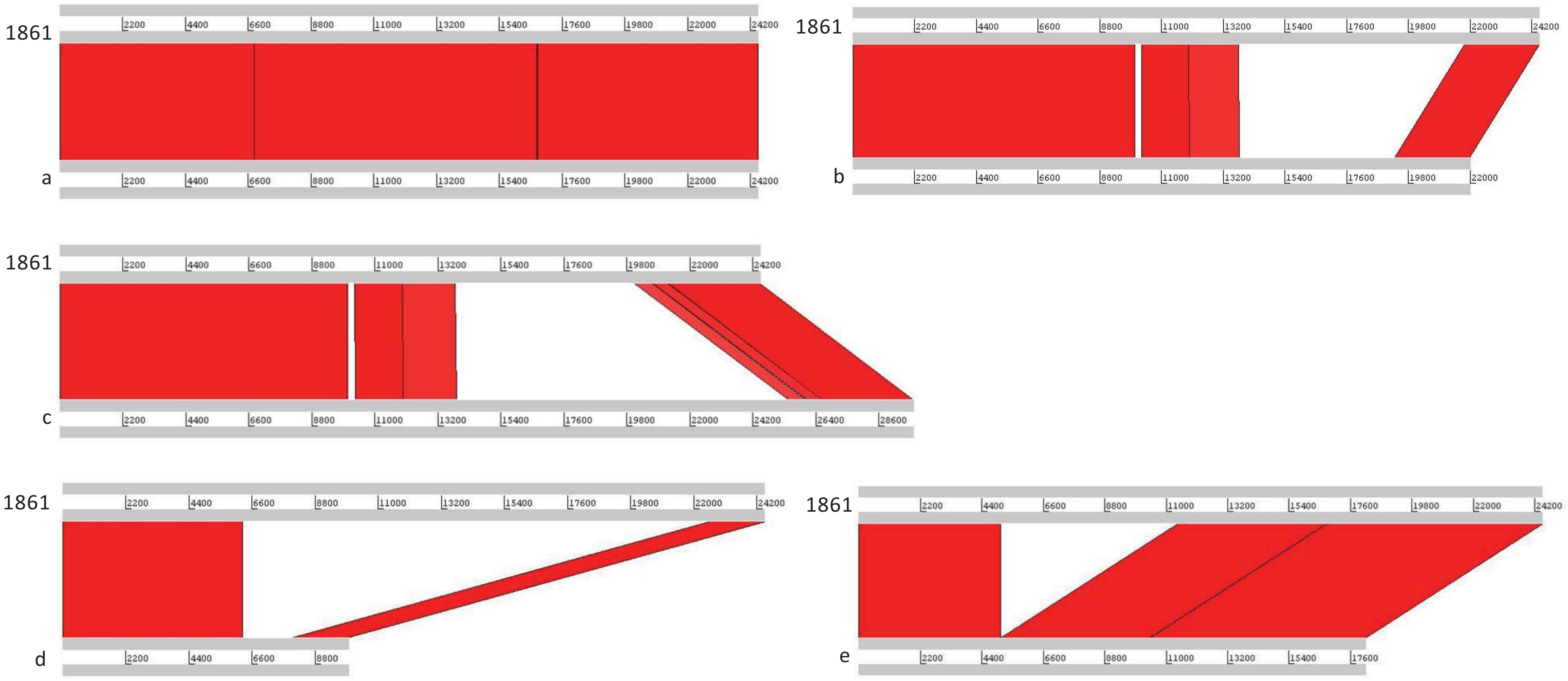


Figure 3.8 ACT alignments between the PPI-1 variable region of 1861 and strains P1031, ATCC 700669, D39, Hungary 19A and G54

Alignments were generated using the ACT between $t - g$ sequence of strains 1861 (Section 3.3.1) and the $t - g$ of strains P1031(a), ATCC 700669(b), D39(c), Hungary 19A(d), G54(e), as described in Section 2.6.1. Red indicates alignment of at least 80% identity over greater than 40bp. $t - g$ sequence of strains P1031, ATCC 700669, D39, Hungary 19A and G54 were obtained from the KEGG database (Section 2.6.2) .

phosphoesterase (ORF 10 in 1861) was annotated as a pseudogene (*SPN23F_09720*) in ATCC 700669 due to a point mutation, similar to *SP_1052* & *SP_1053* in TIGR4 that were annotated as two separate genes. Two regions of homologous sequence, (94% and 92% sequence identity), shared between ATCC 700669 and 1861 were identified within the region of Tn*5252*-related sequence between positions 10,297 and 13,754 in 1861. The region extended from 163 bp into ORF 11 to downstream of the stop codon of ORF 15. The equivalent region in ATCC 700669 encoded ORFs *SPN23F_09731*, *SPN23F_09740*, *SPN23F_09750*, *SPN23F_09780* and *SPN23F_09790*, which includes the putative Rgg/GadR/MutR family transcriptional regulator (*SPN23F_09790*). A 5.5-kb region of divergent sequence was present downstream of *SPN23F_09790*, which was found to include the lantibiotic merscaidin transport system also found in strain 1 (Section 3.3.2.1). The equivalent region of divergent sequence in 1861 was 8-kb in size and encoded ORFs 16 to 24. The final region of homologous sequence was found to encode FtsW. The alignment with D39 (Figure 3.8c), showed a similar pattern of homologous sequence to that of the ATCC 700669 alignment. Therefore, as expected, D39 encoded full-length *nplT* (*SPD_0927*), *pezAT* (*SPD_0930* & *SPD_0931*), the putative Rgg/GadR/MutR family transcriptional regulator (*SPD_0939*) and ORFs *SPD_0934* to *SPD_0938* within the region of Tn*5252*-related sequence. Downstream of *SPD_0939*, D39 contained 11.6-kb of divergent sequence, including ORFs *SPD_0940* to *SPD_0949*, which was not found in either 1861 or strain 1, indicating that D39 possessed a unique set of genes within this region of the PPI-1 variable region. Unlike the ATCC 700669 alignment, D39 shared an extended region of sequence at the 3’ end of the PPI-1 variable region from position 20,102 to the end of the island, which was 1.7-kb longer than ATCC 700669 and encoded the major facilitator family protein *SPD_0950*, homologous to the same, but fragmented, 1861 gene designated ORFs 23 and 24. When aligned with Hungary 19A (Figure 3.8d), 1861 shared an initial 6.3-kb region of homologous sequence that was 1.2 kb longer than the equivalent region in strain 1, and was mostly due to the presence of full-length *nplT* (*SPH_1147*) in Hungary 19A. However, downstream of *SPH_1147* was 1.7-kb of divergent sequence in Hungary 19A, which includes ORFs *SPH_1148* to *SPH_1151*. These ORFs were not found in either 1861 or strain 1. The entire PPI-1 variable region of G54 was present strain 1861 (Figure 3.8e). However, a 6.3-kb region of 1861 sequence was absent from G54. This region in 1861 encodes the last 655 bp of ORF 6 to the first 534 bp of ORF 12, meaning that G54 lacks *pezAT* and possesses a shorter *nplT* (*SPG_0972*), as was also shown in the alignment with strain 1 (Figure 3.7d). Downstream of *pezAT*, the PPI-1 variable region of G54 shared at least 98% sequence identity with 1861, which includes the putative Rgg/GadR/MutR family transcriptional regulator and ORFs homologous to the 1861 ORFs 16 to 25.

As discussed in Section 3.1, a 7-bp repeat sequence had previously been found to flank differences in sequence between TIGR4 and strain 142. However, despite attempts to identify the same 7-bp repeat or alternative repeats using bioinformatic software, no relationship could be identified between the boundaries of deletions or insertions (Figures 3.6, 3.7 & 3.8) and specific repeat sequences.

3.4.2.2 Survey of PPI-1 in a selection of *S. pneumoniae* clinical isolates

As discussed in Section 3.1, previous comparisons between the PPI-1 variable regions of different pneumococcal strains have been undertaken using bioinformatic analysis of sequenced genomes or using PCR and Southern blotting. In order to compare the PPI-1 variable region from a larger selection of strains where the genome sequence was not available, the PPI-1 variable region was amplified from the strains listed in Table 3.6, using primers *a* and *c*, as described in Section 2.9.6, with an extension time of 25 min. Successfully amplified products were subsequently purified

Table 3.6 *S. pneumoniae* strains screened for the PPI-1 variable region

Strain	**Serotype (ST– if available)**	**PPI-1 variable region amplification (+ / -)**	***pezA/T* (+ / -)**
63	18	+	+
94	18C	+	+
160	23F (ST81)	+	+
WCH211	11 (ST3020)	+	+
3773	15B (ST199)	+	+
WU2	3 (ST378)	+	+
4104	19A (ST199)	+	+
1	1 (ST304)	+	-
G54	19F	+	-
3518	11A (ST62)	+	-
2663	11A (ST3019)	+	-
MSHR5	11 (ST62)	+	-
1861	1 (ST3079)	+	+
TIGR4	4 (ST205)	+	+
WCH43	4 (ST205)	+	+
WCH16	6A	+	-
WCH206	3 (ST180)	-	+
73	5	+	+
49	5	+	+
171	19A	+	+
71	5	+	+
141	16	+	-
MSHR17	3 (ST458)	+	+
MSHR1	11A (ST3021)	+	+
D39	2 (ST545)	-	+
EF3030	19F	-	+
164	7C	-	+
140	16	-	+
153	9V	-	+
163	35F	-	-
67	23	-	-

DNA was extracted from the strains listed above as described in Section 2.9.2 and used to amplify the PPI-1 variable region using primers *a* and *c* (Section 2.9.6). *PezAT* detection was performed by PCR (Section 2.9.6) using primers *dm* and *j*.

(Section 2.9.7) and then subjected to *Eco*RI-HF (NEB) digestion (Section 2.9.4). The restriction patterns produced from the PCR product could be used to predict the content of the PPI-1 variable region by comparison to restriction patterns predicted from *in silico* digestion of the same region in strains for which genomic sequences were publicly available. The strains listed in Table 3.6 were chosen to represent a variety of serotypes and strains, including recent clinical isolates and older laboratory strains. Primers *a* and *c* were chosen for amplification of the PPI-1 variable region as the binding sites were found to be conserved in the genomes of strains that had been sequenced so far, with the exception of CDC1087-00 (serotype 7F, ST191) and CGSP14. Strain CDC1087-00 was found to lack the primer *a* binding site, whereas the *c* binding site was within a region that had undergone an inversion in CGSP14. Of the other 20 pneumococcal strains for which at least draft genome sequences were available at the time of this work, successful amplification was predicted to be possible in 15 strains. The strains where amplification was predicted to be very difficult included SP195 (serotype 9V, ST156), D39, CDC1873-00, OXC141 (serotype 3, ST180) and 3-BS71, as products of greater than 20 kb would need to be produced. Products of such sizes are not reliably amplified using the amplification systems described in Section 2.9.6. In addition to predicting the PPI-1 variable region content by restriction patterns, the presence of *pezAT* was also confirmed by PCR. The suitability of primers *dm – j* used to detect *pezAT* was ascertained using BLASTn searches of the NCBI database, which confirmed 100% sequence identity between the primers and their recognition sites in all pneumococcal strains containing the target genes, and not in any of the pneumococcal strains that lacked the system. Table 3.6 shows whether PPI-1 could be amplified and whether *pezAT* was detected.

a – c amplification was not achieved from strains WCH211, D39, EF3030, 164, 140, 153, 163 and 67. Of these strains, strains WCH211, D39, EF3030, 140 and 164 were positive for *pezAT*. Presence of *pezAT* and shared ST between strains WCH211 and OXC141 (McAllister *et al.*, unpublished) suggested that the content of PPI-1 was shared between these strains. Furthermore, since the region was 30-kb in size in OXC141, this would explain the inability to amplify the target region in strain WCH211. As explained above, the target region in the *pezAT* positive strains, SP195 and CGSP14, would have also been impossible to amplify using primers *a* and *c* due to excessive expected product size. Amplification using primers *a* and *c* would also have been difficult in the *pezAT* negative strain, CDC1087-00, but in this case, due to the absence of the primer *a* binding site. Therefore, it seemed likely that when a strain was positive for p*ezAT*, but negative for *a – c*, the target was too large to amplify rather than lacking the target region completely. Since strains 67, 153 and 163 were negative for *pezAT* and negative for *a – c*, it was predicted that these strains possessed an unusual configuration of PPI-1, such as lacking the primer *a* binding site as in CDC1087-00. Alternatively, a configuration not present in any of the genomes published to date could be present in these strains. Figure 3.9 shows the restriction patterns produced following *Eco*RI-HF-digestion of successful *a – c* PCR products.

Table 3.7 shows predicted *Eco*RI restriction product band sizes of a hypothetical *a – c* product from strains for which genomic sequences were publicly available. These predicted restriction patterns were compared with the restriction patterns of strains in Figure 3.9 to predict the content of PPI-1 in these clinical isolates.

Strains 1 and 1861 were found to produce unique restriction patterns, not shared with any of the other strains in Figure 3.9. However, as expected, the restriction pattern produced from strain 1 was the same as that predicted for INV104B (Table 3.7) (3,499 bp, 2,343 bp, 1,888 bp, 1,385 bp and 288 bp). Similarly, strain 1861 produced restriction patterns the same as that predicted for P1031 (7,787bp, 4,816bp, 3,337 bp, 2,503 bp, 568 bp, 288 bp and 86 bp). Strains 141, MSHR17, WCH16 and MSHR1 also

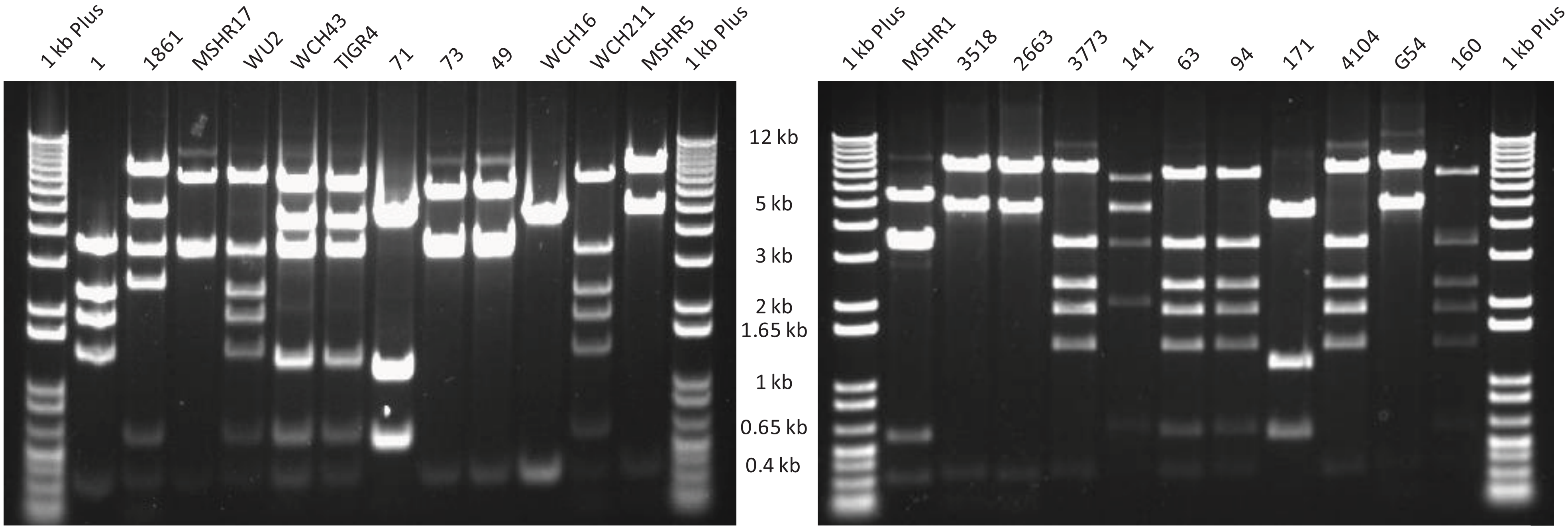


Figure 3.9 *Eco*RI-HF digestion of *a* – *c* sequence from a number of *S. pneumoniae* strains representing a selection of serotypes
Primers *a* and *c* were used to amplify the PPI-1 variable region in the indicated strains (Table 3.6) as described in Section 2.9.6. Successfully amplified products were purified (Section 2.9.7), digested using *Eco*RI-HF (NEB) as described in Section 2.9.4 and analysed by agarose gel electrophoresis, as described in Section 2.9.1. The 1 kb Plus DNA size marker was used to estimate fragment sizes (Section 2.9.1).

Table 3.7 *Eco* RI digestion of *a - c* in KEGG genomes

Strain	Serotype	Number of fragments	Fragment sizes
1861	1 (ST3079)	7	7,788bp, 4,816bp, 3,339bp, 2,503bp, 568bp, 290bp, 86bp
ATCC 700669	23F (ST2)	7	7,102bp, 3,341bp, 2,342bp, 1,888bp, 1,383bp, 568bp, 290bp
1	1 (ST304)	5	3,499bp, 2,343bp, 1,888bp, 1,385bp, 290bp
70585	5 (ST289)	4	6,002bp, 3,593bp, 3,341bp, 290bp
D39	2 (ST595)	11	5,961bp, 4,050bp, 3,341bp, 3,231bp, 1,402bp, 943bp, 800bp, 778bp, 568bp, 346bp, 288bp
G54	19F (ST63)	4	7,868bp, 4,816bp, 290bp, 86bp
Hungary-19A	19A (ST168)	2	4,573bp, 288bp
JJA	14 (ST66)	7	7,109bp, 3,341bp, 2,342bp, 1,888bp, 1,383bp, 568bp, 288bp
P1031	1 (ST303)	7	7,787bp, 4,816bp, 3,337bp, 2,503bp, 568bp, 288bp, 86bp
Taiwan-19F	19F (ST236)	2	4,572bp, 288bp
TIGR4	4 (ST205)	6	6,588bp, 4,374bp, 3,341bp, 1,282bp, 568bp, 289bp
CDC0288-04	12F (ST220)	6	7,677bp, 2,342bp, 2,007bp, 1,888bp, 1,385bp, 288bp
CDC1873-00	6A (ST376)	7	10,502bp, 3,702bp, 3,442bp, 3,341bp, 568bp, 288bp, 175bp
CDC3059-06	19A (ST199)	5	7,644bp, 3,341bp, 2,342bp, 1,888bp, 1,383bp, 288bp
INV104B	1 (ST227)	5	3,499bp, 2,343bp, 1,888bp, 1,385bp, 288bp
MLV-016	11A (ST62)	4	7,868bp, 4,773bp, 288bp, 86bp
OXC141	3 (ST180)	15	7,652bp, 6,162bp, 5,457bp, 3,341bp, 3,224bp, 2,054bp, 1,848bp, 1,050bp, 815bp, 603bp, 568bp, 375bp, 302bp, 288bp, 252bp
3-BS71	3 (ST180)	15	7,651bp, 6,161bp, 5,464bp, 3,341bp, 3,224bp, 2,054bp, 1,848bp, 1,050bp, 815bp, 603bp, 568bp, 375bp, 302bp, 288bp, 252bp
11-BS70	11 (ST62)	5	7,868bp, 4,897bp, 793bp, 288bp, 86bp
18-BS74	6 (new)	2	4,572bp, 288bp
23-BS72	23 (ST37)	10	6,658bp, 3,476bp, 2,587bp, 2,343bp, 2,007bp, 1,888bp, 1,350bp, 568bp, 288bp, 170bp
SP195	9V (ST156)	15	10,861bp, 10,502bp, 5,514bp, 3,676bp, 3,341bp, 2,969bp, 2,427bp, 1,911bp, 1,629bp, 730bp, 672bp, 574bp, 288bp, 229bp, 171bp

The number and sizes of fragments following *Eco* RI digestion was predicted using DNAMAN (Section 2.6.1) produced band patterns not shared with any of the other strains that were screened. However, WCH16 was found to produce a similar restriction pattern to that predicted for strains Hungary 19A, Taiwan 19F and 18-BS74. After examining the reference bands in Table 3.7, strains 141, MSHR17 and MSHR1 could not be grouped with other strains. The remaining strains could each be grouped with at least one other sequenced strain that shared the same restriction patterns. It was interesting to note that strains WU2, WCH211, 3773, 63, 94, 171, 4104 and 160 all produced the same 7-fragment pattern as that predicted for ATCC 700669 (7,102 bp, 3,341 bp, 2,342 bp, 1,888 bp, 1,383 bp, 568 bp and 290 bp). Therefore, the PPI-1 variable region of ATCC 700669 was the most common configuration in the selection of laboratory strains that were screened. Another group of strains included MSHR5, 3518 and 2663, which shared a 4-fragment pattern, and was the same as that predicted for G54 (7,868 bp, 4,816 bp, 290 bp and 86 bp). The predicted restriction patterns for strains MLV-016 and 11-BS70 were also the same as that for G54, WCH43 and TIGR4 (6,588 bp, 4,374 bp, 3,341 bp, 1,282 bp, 568 bp and 289 bp). This was not surprising given that both WCH43 and TIGR4 were ST205. Strains 73 and 49 formed a group with both strains possessing the same 3-fragment pattern (6,000 bp, 3,500 bp and 300 bp). Initially none of the predicted restriction patterns in Table 3.7 could be matched to those generated from strains 73 and 49. However, closer analysis suggested that the 4-fragment pattern predicted for strain 70585 included two fragments differing in size by only 200 bp (3,593 bp and 3,341 bp), which would be difficult to resolve on the 0.8% agarose gel used in Figure 3.9. Therefore, given that the 3,500-bp fragment for strains 73 and 49 was probably a doublet due to its greater intensity of GelRed^{TM} fluoresence than the 6-kb fragment, it was predicted that the PPI-1 variable region of strains 73 and 49 was the same as for strain 70585. Additional restriction analysis using *Bam*HI and *Hin*dIII confirmed that strains 73 and 49 shared the PPI-1 variable region present in strain 70585 (data not shown). The final group included strains 171 and 71, which both produced a 3-fragment pattern estimated at 4,500 bp, 1,000 bp and 600 bp. The restriction patterns of strains 171 and 71 did not appear to correlate with any of the predicted band patterns in Table 3.7, and therefore did not allow for the prediction of the content of the PPI-1 variable region in these strains. However, *pezAT* was detected in both strains 71 and 171.

3.5 Discussion

Brueggemann and Spratt (2003) identified three lineages of STs from an extensive selection of serotype 1 strains that showed distinct patterns of geographic distribution (Section 1.5.2.1). As discussed in Section 1.5.2.1, a number of outbreaks of pneumonia and IPD (Leimkugel *et al*., 2005; Dagan *et* al., 2000; Gratten *et al*., 1993; Mercat *et al*., 1991; Gupta *et al*., 2008; Mehiri-Zghal *et al*., 2009) have been associated with strains of single or closely related STs of serotype 1, which has highlighted the involvement of clonal properties in invasive potential. In addition, the ability of the pneumococcus to acquire genetic material from the environment (Section 1.3.4) may enable horizontally acquired genomic islands to contribute to heightened virulence, thus promoting the efficient transmissibility and rapid disease progression associated with outbreaks of IPD. The acquisition of virulence-promoting genetic elements may provide a short-term selective advantage within a distinct geographic location. However, genetic elements contributing to heightened virulence may not provide a significant long-term advantage, but could still persist due to forced maintenance by TA systems, such as *pezAT*. Therefore, genomic islands, such as PPI-1, that are characterised by extensive variability between clones become interesting loci for the investigation of genetic elements that could confer enhanced IPD potential on single or closely related clones.

3.5.1 Association between ST and content of the PPI-1 variable region

As discussed above, clones of serotype 1 have frequently been associated with outbreaks of IPD, possibly influenced by the content of the various variable genetic islands throughout the pneumococcal genome (Embry *et al*., 2007; Obert *et al*., 2006). Therefore, identifying any potential association between ST and the content of the PPI-1 variable region was important for the serotype 1 clinical isolates under investigation in this study. The STs of the non-invasive strains 1, 2 and 4 and the intermediately virulent strains 3415 and 5482 had previously been found to be members of lineage A. However, prior to commencement of this work the STs of the highly virulent strains 1861 and 4496 were unknown. Given that the organisation of the PPI-1 variable region was known to vary between the lineage A strains and strains 1861 and 4496 (Section 1.6.3), it was important to determine whether there was a correlation between serotype 1 lineage and the content of PPI-1. The STs of strains 1861 and 4496 were determined by MLST and found to be novel, being allocated ST3079 and ST3018, respectively (Section 3.2.1). Further sequence analysis (Section 3.2.2) enabled the identification of previously defined STs that were most closely related to ST3079 and ST3018, based on the assumption that minimum genetic drift would have occurred between the most closely-related STs. The most closely related STs to the STs of 1861 and 4496 were ST217 and ST615, respectively. Therefore, strains 1861 and 4496 were found to be members of lineage B and C, respectively. As a result, the non-invasive and intermediately virulent strains appeared to cluster within lineage A, separate from the two highly virulent strains, supporting previous work showing that the genotype influences virulence (Section 1.5.2). The conservation of the PPI-1 variable region between 1861 and 4496, but not with the lineage A strains, indicated that either one or both of the two versions of the region were acquired at a time between the divergence of lineage A clones and lineages B and C, and the divergence of lineage B and C clones from each other (Figure 3.3). Interestingly, the highly virulent strain 1861 was a single-locus variant of ST217 and a double-locus variant of ST303, which have previously been described as hypervirulent in humans (Antonio *et al*., 2008). Therefore, the heightened virulence of strain 1861 (and possibly 4496) in mice is likely to also be reflected in humans.

3.5.2 Sequencing and annotation of the PPI-1 variable region

As TA systems such as *pezAT* have been suggested to play a role in the maintenance of mobile genetic elements (Szekeres *et al*., 2007) and in the case of strains 1861 and 4496, is associated with heightened virulence (Section 1.6.3), it was decided to sequence and annotate the region surrounding *pezAT* in strains 1861 and 4496 and to compare this to the sequence and annotation of the PPI-1 variable region in the lineage A strains, which had previously been shown to be *pezAT* negative (Section 1.6.3). Sequencing confirmed that the PPI-1 variable region of strains 1861 and 4496 was considerably larger than that of the lineage A strains. Having completely sequenced the PPI-1 variable region in the strains of this study, ORFs within the region of each strain could be identified and their putative functions predicted by use of the NCBI and KEGG databases, and the HHpred search engine (Section 3.3.2). Figure 3.10 summarises the configuration of ORFs predicted for the PPI-1 variable region and flanking sequence of the lineage A strains and strains 1861 and 4496. Within the region, 12 ORFs were identified in the lineage A strains and 25 ORFs in strains 1861 and 4496. The two versions of the sequence were aligned using the ACT, which identified four regions of homologous sequence, two regions of absent sequence and one region of divergent sequence.

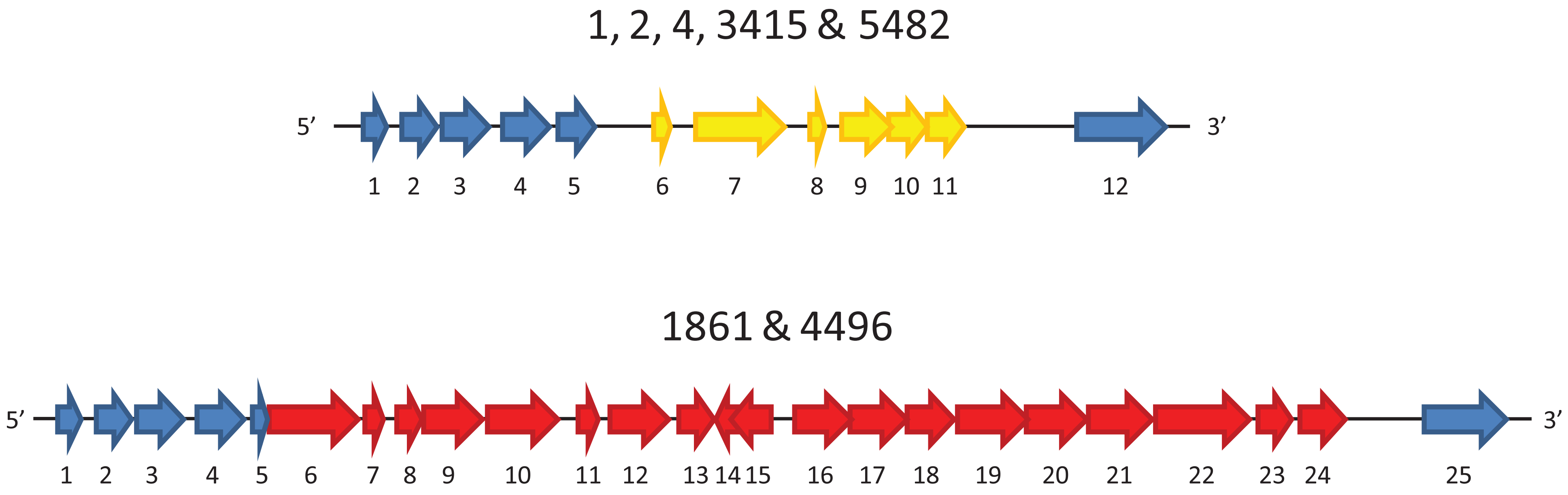


Figure 3.10 Summarised organisation of predicted ORFs in the PPI-1 variable region and flanking regions of the lineage A strains and strains 1861 and 4496

Arrows represent approximate relative sizes and position of ORFs predicted in Section 3.3.2. ORFs are numbered as in Tables 3.2 & 3.3. Blue ORFs are shared by both groups of strains, yellow ORFs are present in only the lineage A strains and red ORFs are present in only strains 1861 and 4496.

3.5.2.1 Sequence shared between both versions of the PPI-1 variable region

As expected, the regions flanking the PPI-1 variable region (homologous regions 1 and 4, Figure 3.6), encoded ORFs that were highly conserved between the two groups of strains. Four complete ORFs were predicted within the homologous region 1 (ORFs 1 – 4, Figure 3.10). Of the 4 ORFs, a function could not be predicted for ORFs 1 and 2, but ORFs 3 and 4 were found to encode a putative ADP-ribosylglycohydrolase and a diacylglycerol kinase catalytic domain protein, respectively (Tables 3.2 & 3.3). As noted in Section 3.3.2.1, ADP-ribosylglycohydrolase is involved in ADP-ribosylation, which involves the reversible post-translational modification of target proteins to regulate their activities (Ludden *et al*., 1994). ADP-ribosylation is catalysed by ADP-ribosyltransferase and the modification is removed by ADP-ribosylglycohydrolase. ADP-ribosylation has been implicated in the regulation of a number of processes such as nitrogenase activity and modification of glutamine synthetise in *Rhodospirillum rubrum* and sporulation in *B. subtilis* (Masepohl *et al*., 1993; Woehle *et al*., 1990; Huh *et al*., 1996). It was not clear which proteins were targeted by the ADP-ribosylglycohydrolase and an ADP-ribosyltransferase was not identified in the sequenced portion of PPI-1. However, the virulence-associated *bvr* locus in *Listeria monocytogenes* has also been found to encode an ADP-ribosylglycohydrolase without an ADP-ribosyltransferase (Brehm *et al*., 1999). Perhaps the ADP-ribosyltransferase is present elsewhere on the chromosome and as such could act in trans. As also noted in Section 3.3.2.1, diacylglycerol kinase is involved in the turnover of membrane phospholipids (reviewed in Sakane *et al*., 2007). However, it is unclear what role such an enzyme would have in PPI-1. The 3’ end of homologous region 1 was within the *nplT* encoded by both groups of strains (ORF 5 in strain 1 and ORF 6 in strain 1861). However, *nplT* was found to be truncated in the lineage A strains, probably leading to lack of function due to the loss of most functional domains. By contrast, the total length of the fragmented *nplT* in strains 1861 and 4496 (ORFs 5 and 6) was comparable to that of other pneumococcal strains and the functional enzyme in *B. stearothermophilus*, for which the crystal structure has been determined (Hondoh *et al*., 2003). Interestingly, a frame-shift mutation in ORF 5/6 had led to fragmentation of neopullulanase into ORFs 5 and 6. Whilst the enzyme was fragmented in strains 1861 and 4496, the fragmentation occurred at a position between the N-domain and A domain, leaving the N-domain complete within ORF 5 and the remaining domains intact within ORF 6. Therefore, three theoretical possibilities exist for the functionality of *nplT* in strains 1861 and 4496. Firstly, the fragmentation could lead to synthesis of a non-functional protein, which would then eliminate a role for neopullulanase in the virulence of strains 1861 and 4496. Secondly, the N-domain has been shown to be responsible for the tightened specificity of neopullulanase when compared to other members of the α-amylase superfamily (Hondoh *et al*., 2003), and as such an intact α-amylase is produced without an N-domain, which could increase the range of potential substrates for the enzyme in strains 1861 and 4496. Thirdly, a functional neopullulanase could exist in strains 1861 and 4496, if the N-domain of ORF 5 was able to maintain its function by interacting with the remaining domains of ORF 6 via peptide complementation. Expression analysis of *nplT* in 1861 and 4496 will be further investigated in Chapter 4.

Remnants of the conjugative transposon Tn*5252*, were observed in both groups of strains (ORF 6 in strain 1 and ORFs 11 – 13 in 1861) within homologous regions 3 and 4 of Figure 3.6. Whilst both groups of strains possessed remnants of Tn*5252*, the extent to which the region had degraded at the sequence level between the two groups, combined with the presence of non-homologous region 2 of Figure 3.6 in only strains 1861 and 4496, led to significant differences in annotation. However, in both cases the

Tn*5252* conjugative transposon was incomplete, which suggested that the region was unlikely to have any real function in either group of strains.

A single ORF was predicted within homologous region 4 of Figure 3.6, which was found to be the cell division protein FtsW. FtsW is thought to be involved in cell division by stabilising the FtsZ-ring (Ishino *et al*., 1989).

3.5.2.2 PPI-1 variable region present in only the lineage A strains

The primary feature of the PPI-1 variable region in strain 1 was the presence of part of a mersacidin lantibiotic biosynthesis and export system (Section 3.3.2.1). Lantibiotics are a group of post-translationally modified bacteriocins, which includes the bacteriocin, mersacidin (reviewed in Willey *et al*. [2007]). Mersacidin functions to inhibit the transglycosylation step in peptidoglycan synthesis and has been of particular interest as a future antimicrobial due its activity against methicillin-resistant *S. aureus* in a mouse model of infection (Kruszewska *et al*., 2004). In *B. subtilis* the lantibiotic mersacidin system includes the mersacidin structural gene (*mrsA*), as well as genes encoding the post-translational modification enzymes (*mrsM* and *mrsD*), a dual-function endopeptidase-transporter gene (*mrsT*), a group B ABC transporter involved in mersacidin immunity (*mrsF*, *mrsG* and *mrsE*), a two-component regulatory system important for regulating transcription of the immunity genes (*mrsR2* and *mrsK2*) and a regulatory gene with some sequence similarity to response regulators (*mrsR1*), which was required for transcription of *mrsA*, *mrsM*, *mrsD* and *mrsT*. A companion sensor histidine kinase gene has not been identified for *mrsR1* in *B. subtilis* (Altena *et al*., 2000). The strain 1 ORFs identified in Section 3.3.2.1 were homologous to *mrsT* (ORFs 7 and 8), *mrsF* (ORF 9), *mrsG* (ORF 10) and *mrsE* (ORF 11). Therefore, the strain 1 PPI-1 variable region lacked the mersacidin structural gene, most of the modification enzymes and the two-component regulatory systems. However, the possession of the complete immunity system suggested that strain 1 may be resistant to mersacidin without producing the bacteriocin itself. Even though strain 1 lacks homologs of *mrsR2* and *mrsK2*, it has previously been shown that mersacidin can autoinduce the expression of its own immunity system (Schmitz *et al*., 2006). This suggests that despite the lack of immunity regulatory genes in strain 1, the system could still be expressed in the presence of mersacidin. Chapter 4 will address the expression of ORFs 9 to 11 *in vitro*, showing whether or not the mersacidin immunity genes are actively transcribed. The arrangement of lantibiotic genes seen in strain 1, with the exception of the truncation of the *mrsT* homolog and the lack of an *mrsM* fragment, has been previously identified in strain ATCC 700669 (Croucher *et al*., 2009). To date, these systems have been identified in a number of soil dwelling organisms such as *B. subtilis*. However, bioinformatic analyses have identified a complete mersacidin biosynthesis and export system in the pneumococcal strain, 23-BS72 (Croucher *et al*., 2009). Therefore, if strains of the pneumococcus do indeed exhibit mersacidin resistance, it is difficult at this point to know what these strains are competing with, other than with other strains of *S. pneumoniae*.

3.5.2.3 PPI-1 variable region present only in strains 1861 and 4496

The first region of sequence present in only strain 1861 was the non-homologous region 1 (Figure 3.6), which encodes *pezAT* (ORFs 8 and 9) and ORF 10 (Figure 3.10). As had been previously established, *pezAT* was present in strains 1861 and 4496, but absent from the lineage A strains (Section 1.6.3). However, unlike what had been observed in TIGR4 (*SP_1052* and *SP_1053*) and D39 (*SPD_0932* and *SPD_0933*), a single ORF (ORF 10) was present in the equivalent sequence in 1861, indicating that fragmentation of the gene had occurred in strains such as TIGR4 and D39. A putative function for ORF 10 could not be predicted from bioinformatic analyses, due to a lack of significant homology representing more than half the length of the ORF. However, it appeared that ORF 10 was an ATP-binding protein (Section

3.3.2.2). Brown *et al*. (2004) suggested that the reduction in virulence observed following mutagenesis of *pezT* was primarily due to *pezT*, rather than through polar effects on *SP_1052* (homologous to ORF 10). However, given that Section 3.3.2.2 showed that ORFs *SP_1052* and *SP_1053* are probably a fragmented gene, *SP_1052* is possibly non-functional in the strain used by Brown *et al*. (2004). Performing comparisons between *in vivo* expression profiles of *pezAT* in different niches of the mouse may shed some light on what role, if any, this system has in pathogenesis (Chapter 4).

A second region of sequence present in only strains 1861 and 4496 (non-homologous region 2 [Figure3.6]) was predicted to encode a fragmented transcriptional regulator from the Rgg/GadR/MutR family. As described in Section 3.3.2.2, a single nucleotide insertion had led to fragmentation and most likely leads to inactivation of the regulator. An identical point mutation had previously been identified in *ropB*, which is a positive regulator of the virulence-associated protease, *speB* in *S. pyogenes* (Hollands *et al*., 2008). However, it is interesting that the *speB* negative phenotype can be selected following subcutaneous passage. This raised the possibility that *ropB* and the equivalent transcriptional regulator in strains 1861 and 4496 might be phase variable. However, in *S. pyogenes*, the *ropB* truncation was never recovered following subcutaneous passage and that the *speB* negative expression phenotype was due to phase variation of the *covR/S* regulator. Therefore, it seems unlikely that the regulator in PPI-1 would be phase variable and as such it is unlikely that ORFs 14 and 15 have a role in the virulence of strains 1861 and 4496.

The strain 1861 version of non-homologous region 3 (Figure 3.6) was found to encode a cluster of genes (ORFs 16 – 24, Figure 3.10), which were not present in strain 1 (Section 3.4.1). Annotation of this region predicted putative functions for ORFs 16, 18, 21, 22, 23 and 24. These genes encoded a putative 3HIBDH (ORF 16), PDT (ORF

18), GalE (ORF 21), biotin carboxylase (ORF 22) and a fragmented transporter of the major facilitator superfamily (ORFs 23 and 24). Putative functions could not be predicted for ORFs 19 and 20, and ORF 17 only returned moderate-scoring hits for the nucleotide-binding domain of a nucleotidyl transferase. HHpred searches of ORFs 23 and 24 as a single sequence query returned high-scoring hits to importers of lactose and glycerol-3-phosphate. However, it was unclear whether the fragmented gene found in strain 1861 and 4496 could be functional, as evidence of peptide complementation rescuing the function of a fragmented major facilitator has not been previously reported. The products encoded in ORFs 18, 21 and 22 have well-characterised housekeeping functions (Wilson & Hogness, 1969; Cotton & Gibson, 1965; Cronan, 2002). It is important to note that the versions of these genes in PPI-1 are in addition to similar enzymes located elsewhere in the genome. Therefore, ORFs 18, 21 and 22 might be under alternative regulatory control to that of the common housekeeping genes. From the bioinformatic analyses of Section 3.3.2.2, it was difficult to predict a mechanism by which ORFs 16 – 24 could function to enhance virulence, making mutagenesis and expression analysis of these genes particularly important (Chapter 4).

3.5.3 Generalised organisation of PPI-1

The PPI-1 variable region of strains 1 and 1861 was aligned with that of INV104B and P1031 respectively, which highlighted the high degree of homology exhibited between strains of the same lineage (Section 3.2.2). Both strains 1 and 1861 were aligned with strains ATCC 700669, Hungary 19A, and G54. The alignments in Figures 3.7 and 3.8 revealed a pattern of independent components within the PPI-1 variable region that is shown in Figure 3.11. The pattern included three regions of homologous sequence (overlap A, B and C), the *pezAT* region and an accessory region. Overlap A and C were the most highly conserved and consisted of ORFs found to be shared by all strains sequenced to date, with the exception of CDC1087-00. Overlap A

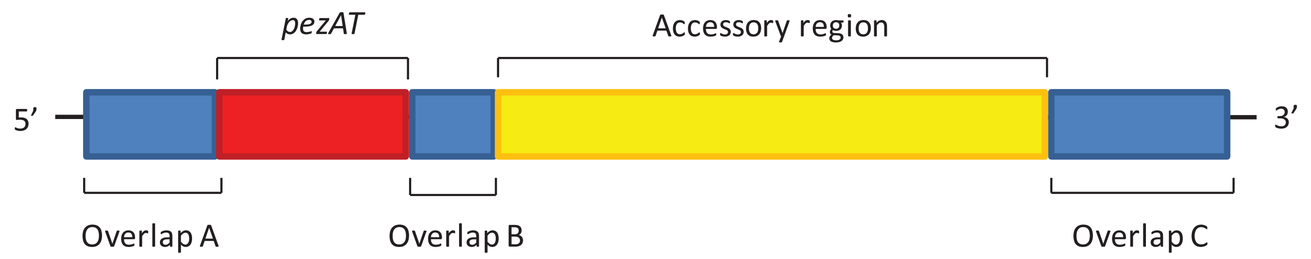


Figure 3.11 General structure of PPI-1 variable region
The generalised structure of the PPI-1 variable region indicates the three regions of homologous sequence (blue) the *pezAT* region (red) and the accessory region (yellow). The presence or absence of the *pezAT* region and the content of the accessory region, has been observed to vary between strains. Overlap B can vary in length, most likely due to recombination occurring at the region’s boundaries.

includes sequence from the start of the PPI-1 variable region to within *nplT* and corresponds approximately to homologous region 1 in Figure 3.6. Overlap C can vary in length and includes the 3' end of the PPI-1 variable region, which is included in homologous region 4 of Figure 3.6. Overlap C was not found to encode any significant ORFs. Overlap B tended to vary in length between strains and was found to be absent in a few strains, such as Hungary 19A. In all cases where overlap B sequence was present, homology was found to exist between strains, as the variation that was observed was a result of genetic drift and recombination at the region's boundaries, rather than variation due to replacement events. Overlap B was found to consist primarily of Tn*5252*-related sequence and the putative Rgg/GadR/MutR family transcriptional regulator. Whilst the Tn*5252* ORFs within overlap B of most strains are unlikely to encode a functional conjugative transposase, the ubiquitous nature of transposase sequence throughout many bacterial genomes may promote the sort of frequent recombination required to produce variation characteristic of the PPI-1 variable region. The overlap B region of strain 1 was unusual when compared to other pneumococcal genomes, as it lacked the putative transcriptional regulator. Since the 300 bp sequence of overlap B downstream of the transcriptional regulator deletion site (equivalent to homologous region 4 [Figure 3.6]) was present in the lineage A strains, the loss of the transcriptional regulator was likely to have occurred independently of the acquisition of the lineage A accessory region.

The *pezAT* region of the PPI-1 variable region consists of the remaining sequence of full-length *nplT*, *pezAT* and the putative phosphoesterase. The *pezAT* region appeared to be either present or absent in the pneumococcal strains that were analysed. The absence of the *pezAT* region usually resulted in direct contact between overlap A and B leading to a truncated *nplT*, with the exception of strains Hungary 19A and Taiwan 19F. This was usually followed by Tn*5252*-related sequence, which was observed in strains such as strain 1 and G54. The accessory region was the region of greatest variation between strains. Unlike overlap B, the variability of the accessory region suggested that it has been a hotspot for recombination, frequently leading to replacement of existing sequence. Often the accessory region encodes a set of genes of a specific function, such as the lantibiotic immunity genes present in strains 1 and ATCC 700669. In a number of genomes the accessory region of PPI-1 has been found to encode genes for antibiotic resistance, such as chloramphenicol acetyltransferase in CDC1873-00. Given the common structure of the PPI-1 variable region between pneumococci, it is predicted that the role of the region for a given strain is determined by the content of the accessory region.

As indicated in Section 3.4.2.1, several attempts were made to identify various repeat sequences potentially responsible for the component structure of the PPI-1 variable region. Whilst many repeated sequences were identified, no relationship between the boundaries of components of the PPI-1 variable region and repeat sequences was identified. However, finding repeats was probably hampered by the fact that the evolution of the PPI-1 variable region appears to have been mediated by multiple steps of recombination and replacement over time.

3.5.4 Survey of PPI-1 in a selection of clinical isolates

Previously the content of the PPI-1 variable region had been screened by directly targeting specific genes, such as *pezT* by PCR or Southern hybridisation analysis (Brown *et al*., 2004). However, by directly targeting individual genes, there is the risk of reporting false negatives due to point mutations in primer binding sites, in the case of PCR, or false positives due to hybridisation to sufficiently homologous genes located elsewhere on the genome, in the case of Southern Blotting. In Section 3.4.2.2, a more holistic approach was taken to compare the entire PPI-1 variable region between strains, by taking advantage of the conservation of sequence flanking the region in the majority of strains. This was achieved by amplifying the PPI-1 variable region using primers *a* and *c* and digesting the resultant product, if any, with *Eco*RI. Characteristic *Eco*RI restriction patterns were used to identify strains possessing the same version of the PPI-1 variable region. To complement the restriction patterns, PCR was used to confirm the presence of *pezAT* and homologs of ORF 10, which detected only the PPI-1 TA system and not the homologous system located elsewhere on the chromosome. A similar approach was attempted for the Rgg/GadR/MutR family transcriptional regulator, but this gene was found to exhibit sequence variation that made such an approach unreliable. Section 3.4.2.2 showed that many strains did produce identical restriction patterns allowing grouping of strains sharing the same version of the PPI-1 variable region. In addition, Table 3.8 summarises the grouping of the screened strains, including those where band patterns were predicted *in silico*, according to accessory region content. For those strains where sequence data was available, it was possible to show that some possessed homologous accessory regions, but produced different restriction patterns due to differences in the presence of the *pezAT* region. Alignments from Figure 3.7 showed that the accessory region of the serotype 1 lineage A strains was homologous to that of ATCC 700669, allowing them to be placed in group 1. Likewise the alignment in Figure 3.8 between strain 1861 and G54 showed extensive homology between their accessory regions allowing strains 1861 and 4496 to be placed in group 2. In addition, the accessory region of the *pezAT*-positive strain MSHR17 was found to be homologous to that of Hungry 19A (McAllister, unpublished), allowing strain MSHR17 to be placed in group 4.

Table 3.9 shows the strains that could not be grouped as either their restriction patterns were unique, or because the PPI-1 variable region could not be amplified. Excluding those strains where *a* – *c* amplification was unsuccessful, the inability to group strains in Table 3.9 was most likely due to that fact that they either possessed

Table 3.8 Grouping of strains according to accessory region

Group	Strains	Serotype (ST)	*pezAT*
1	ATCC 700669	23F	+
	63	18	+
	94	18C	+
	160	23F	+
	WCH211	11 (ST3020)	+
	3773	15B (ST199)	+
	WU2	3 (ST378)	+
	4104	19A (ST199)	+
	170	19A	+
	152	9N	+
	162	24	+
	1	1 (ST304)	-
	3415	1 (ST227)	-
	INV104B	1 (ST227)	-
2	G54	19F	-
	3518	11A (ST62)	-
	2663	11A (ST3019)	-
	MSHR5	11 (ST62)	-
	MLV-016	11A	-
	11-BS70	11	-
	1861	1 (ST3079)	+
	4496	1 (ST3018)	+
3	TIGR4	4 (ST205)	+
	WCH43	4 (ST205)	+
4	Hungary 19A	19A	-
	Taiwan 19F	19F	-
	WCH16	6A (novel)	-
	18-BS74	6 (novel)	-
	MSHR17	3 (ST458)	+
5	OXC141	3 (ST180)	+
	3-BS71	3 (ST180)	+
	WCH206	3 (ST180)	+
6	70585	5 (ST289)	+
	73	5	+
	49	5	+
7	171	19A	+
	71	5	+

Strains are grouped with other strains sharing the same version of the accessory region of the PPI-1 variable region. Restriction patterns of Figure 3.9 were compared to the patterns predicted in Table 3.7. *PezAT* was detected using primers *dm* and *j* (Table 3.6). Red indicates the serotype 1 strains of this study and blue indicates the strains with publicly available genome sequences.

Table 3.9 Strains that could not be grouped by accessory region

a – c amplification	Strains	Serotype (ST)	*pezAT*
Positive	141	16	-
	MSHR1	11 (ST3021)	+
	JJA	14 (ST66)	+
	CDC0288-04	12F (ST220)	+
	CDC1873-00	6A (ST376)	+
	CDC3059-06	19A (ST199)	+
	SP195	9V (ST156)	+
	23-BS72	23 (ST37)	+
Negative	D39	2 (ST454)	+
	EF3030	19F	+
	164	7C	+
	140	16	+
	153	9V	-
	163	35F	-
	67	23	-

Strains that could not by accessory region in Table 3.8 due to unique restriction patterns or unsuccessful *a* – *c* amplification are listed above. Red indicates the serotype 1 strains of this study and blue indicates the strains with publicly available genome sequences.

unique versions of the accessory region, or if the accessory region was shared with other strains, differences in the presence of *pezAT* may have altered the restriction patterns.

As discussed in Section 3.4.2.2, the inability to amplify the PPI-1 variable region in some strains was probably due to the size of the region being greater than the limits of reliable PCR amplification, especially in those strains that were *pezAT*-positive. However, where *pezAT* could not be detected, it is possible that these strains possess a version of the PPI-1 variable region that may lack at least one of the primer-binding sites, such as in CDC1087-00 and CGSP14. Therefore, whilst it is clear that only sequencing can definitively determine the content of the PPI-1 variable region in all strains, the comparison of restriction patterns between strains was quite successful at grouping the majority of strains according to accessory region content. In addition to grouping strains, the comparison of restriction patterns between tested strains for which there was no sequence data available, with that of a strain from Table 3.7, enabled the rapid identification of the content of the PPI-1 variable region.

In summary, the mersacidin immunity system was the most frequently detected component of the accessory region, and was present in a wide variety of serotypes such as 23F, 19A, 11, 18 and 1. The apparent success of this region supports the hypothesis that the region provides immunity against strains or species producing the lantibiotic bacteriocin, mersacidin. The serotype 1 lineage A strains of this study were unique, as they possessed the mersacidin immunity genes in the absence of *pezAT*, which suggested that in these strains the mersacidin immunity genes could be maintained independently of *pezAT*. The accessory region of strains 1861 and 4496 was also found to be distributed throughout a range of other serotypes including serotype 11 and 19F. However, this accessory region, unlike in strains 1861 and 4496, was most commonly found in the absence of the *pezAT* region (Table 3.8). This indicated that these genes could also be maintained in the absence of *pezAT*. Therefore, it seems unlikely that *pezAT* is required for maintaining the accessory region of at least groups 1, 2 and 4 (Table 3.8).

The bioinformatic data of Chapter 3 have identified an association between the content of the PPI-1 variable region and invasive potential in serotype 1 isolates. In addition, these versions of the PPI-1 variable region were identified in many other strains of *S. pneumoniae* that represented many different serotypes. In addition, the PPI-1 variable region was found to exhibit an ordered structure made up of a number of components that either varied in their presence or in their content. Functional characterisation is required to demonstrate the role potentially played by the PPI-1 variable region in virulence, and this will be addressed in Chapter 4.

Chapter 4 – Functional Characterisation of Pneumococcal Pathogenicity Island-1

4.1 Introduction

Previous work showed that at the DNA level, differences in the size of the PPI-1 variable region and the possession of *pezAT* were associated with invasive potential (Section 1.6.3). More specifically, of the non-invasive, intermediately virulent and highly virulent serotype 1 isolates that were analysed, the highly virulent strains possessed a larger version of the PPI-1 variable region that was positive for *pezAT*, whereas the non-invasive and intermediately virulent strains possessed a smaller version of the region that lacked *pezAT* (Section 1.6.3). In Section 3.3.2, sequencing and bioinformatic analyses were used to identify ORFs within the PPI-1 variable region and predict their function based on homology with previously annotated genes. The lineage A strains were found to possess a version of the region that encoded a truncated *nplT*, lacked *pezAT* and encoded a putative mersacidin lantibiotic immunity system. However, the mersacidin structural gene and modification enzymes were lacking in these strains (Section 3.3.2.1). In contrast, strains 1861 and 4496 harboured a fragmented *nplT*, the *pezAT* locus and a region of putative metabolic enzymes and hypothetical proteins that were flanked at the 3' end by a transporter of the major facilitator superfamily. Whilst the putative functions were predicted for some of the individual genes in the region, it was unclear what role the region as a whole might play in pathogenesis. Brown *et al*. (2004) showed that deletion of *pezT* from TIGR4 led to a reduction in competitive fitness compared to the wild-type following both i.n. and i.p. challenge. However, whilst this gene alone may have had an impact on the virulence of TIGR4, the role played by the entire PPI-1 variable region was not investigated. In particular, considering that strains 1861 and 4496 exhibit heightened invasiveness, the importance of the entire variable region to the virulence of these strains should be investigated.

In light of the sequence analysis conducted in Chapter 3, Chapter 4 aimed to perform functional characterisation of the PPI-1 variable region of the lineage A strains and the highly invasive strains. In particular, attention was paid to the transcriptional structure of the PPI-1 variable region and the niche-specific expression patterns of the region in the highly virulent strains. Subsequently, mutagenesis of the PPI-1 variable region was used to assess the region's contribution to competitive fitness *in vivo*.

4.2 Transcription of the PPI-1 variable region *in vitro*

Having completed the characterisation of the PPI-1 variable region at the DNA sequence level (Chapter 3), it was possible to predict the transcriptional structure of the region in the lineage A strains and strains 1861 and 4496. The putative transcriptome for PPI-1 would be used to help select the genes for differential gene expression analysis. PCR was used to amplify regions spanning consecutive genes from cDNA template, which if successful would show that the targeted genes were co-transcribed. However, initially individual genes were checked to ensure that they are in fact transcribed *in vitro*. In addition, mRNA was extracted from both exponential and stationary phase cultures in case expression of the selected genes was favoured in one of these conditions. cDNA was generated from the relevant strain under both conditions (Section 2.12.4) from RNA that was extracted and DNase-treated from mid-log phase and stationary-phase SB cultures, as described in Section 2.12.1. For each cDNA sample, amplification was also attempted using primers J237 and J276, which are located within the capsule locus, and was used as a positive control for detection of cDNA. A positive control was performed for each primer set using genomic DNA from the relevant strain as template.

Table 4.1 shows that no expression of ORF 5 was detected, which suggested that the truncated *nplT* of ORF 5 was either not expressed under the *in vitro* conditions that were used, or that the ORF was a no longer functional remnant of a previously active gene. Expression of ORFs 6, 7, 8, 9, 10 and 11 was successfully detected, which showed that the immunity system of strain 1 remains transcriptionally active despite the absence of the remainder of the mersacidin lantibiotic biosynthesis system (Section 3.3.2.1). Exponential and stationary growth conditions did not affect detection of expression in strain 1. Table 4.2 shows that successful amplification was achieved between consecutive ORFs from ORF 7 to ORF 11, which indicated that these ORFs were transcribed on the same mRNA transcript. Therefore, promoter prediction software was used (Section 2.6.1) to identify putative promoters upstream of ORF 7 that might regulate the expression of the operon consisting of ORFs 7 – 11. Potential promoters were considered based on their score and position relative to the start codon of the first gene of the predicted operon. The putative promoter with the highest probability score (score 0.99) that was predicted upstream of ORF 7 was between positions 6,567 – 6,612 (*t* – *g*, strain 1 PPI-1 sequence [Section 3.3.1]), with a putative transcription start site at position 6,603. In addition, terminator prediction software (Section 2.6.1) was used to identify putative Rho-independent terminators within the strain 1 sequence that contain an inverted repeat with ΔG< -11. Whilst these searches would be unable to detect more discrete sequences associated with Rho-dependent terminators, it was possible to predict stem-loop structures that could terminate transcription. A Rho-independent terminator was identified between positions 12,181 – 12,229 (ΔG= -23.3), which is downstream of the ORF 11 stop codon and may function to terminate the transcription of the putative ORF 7 – 11 operon. Figure 4.1 summarises the transcriptional structure of the PPI-1 variable region in strain 1 based on the PCR results of Table 4.3 and the results of promoter prediction and terminator prediction searches. Table 4.3 shows that only the

Table 4.1 Amplification of individual PPI-1 variable region strain 1 ORFs from cDNA

ORFs	Primers*	A_{600} 0.15	A_{600} 0.25
5	*s/cd*	-	-
6	*v/eg*	+	+
7	*aj/ak*	+	+
8	*el/ef*	+	+
9	*ek/y*	+	+
10	*cf/p*	+	+
11	*eh/n*	+	+

*Primers are described in Table 2.3.
'+' PCR product of expected size; '-' no PCR product.
A_{600} 0.15 = Mid-log phase; A_{600} 0.25 = Stationary phase.

Table 4.2 Amplification between consecutive PPI-1 variable region ORFs from strain 1 cDNA

ORFs	Primers*	PCR (- / +)
6 – 7	*ce/dj*	-
7 – 8	*aj/aa*	+
8 – 9	*cf/y*	+
9 – 10	*cg/p*	+
10 – 11	*ch/n*	+

*Primers are described in Table 2.3.
'+' PCR product of expected size; '-' no PCR product.
PCR was performed using cDNA template derived from stationary phase culture (A_{600} 0.25)

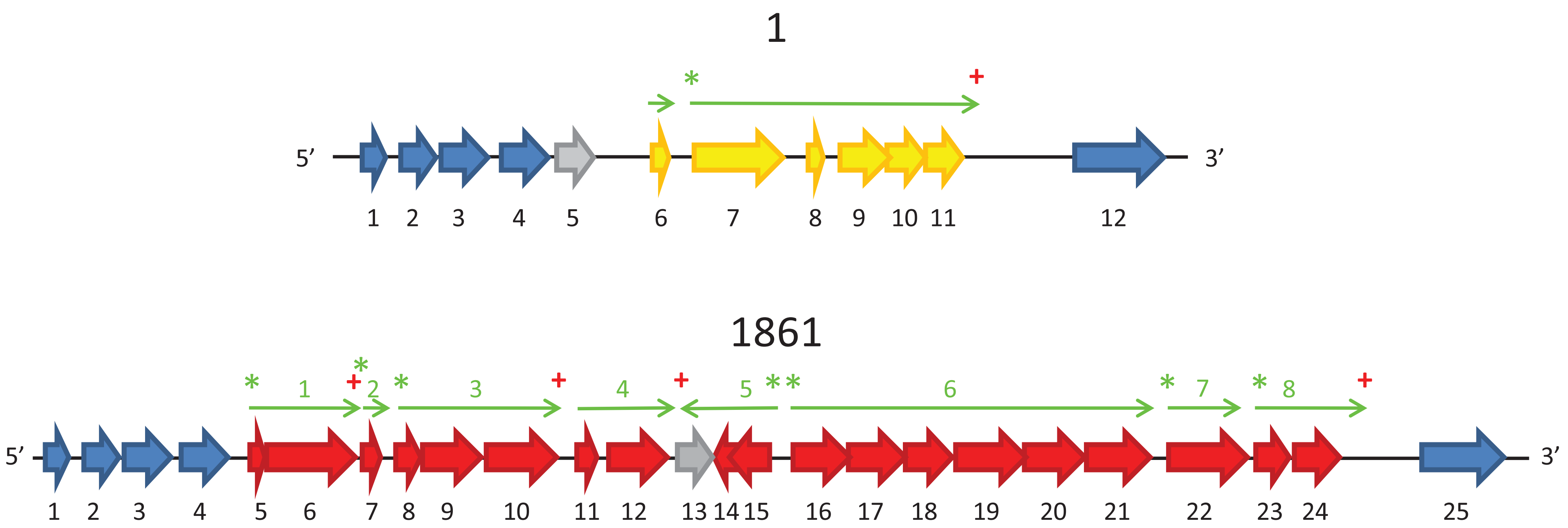


Figure 4.1 Transcriptional structure of the PPI-1 variable region in strains 1 and 1861

PCR was used to detect the *in vitro* expression of individual ORFs in the PPI-1 variable region of strains 1 and 1861. ORFs for which detection was not attempted are coloured blue, ORFs for which a transcript was not detected are coloured grey and those for which a transcript was detected are coloured yellow in strain 1 and red in strain 1861. Green arrows represent the position of mRNAs transcribed from the PPI-1 variable region of both strains 1861 and 4496, which was predicted by PCR amplification of regions spanning ORF junctions from cDNA template (Tables 4.2 and 4.4). Green asterisks indicate the position of putative promoter sequences and red '+' indicates the position of putative transcription terminators that were identified as described in Section 2.6.1. Green numbers indicate the mRNA transcript number as referred to in Section 4.2.

Table 4.3 Amplification of individual PPI-1 variable region ORFs from strain 1861 cDNA

ORFs	Primers*	A_{600} 0.17	A_{600} 0.25
5	*s/ed*	+	+
6	*dk/dl*	+	+
7	*ci/ap*	+	+
8	*dm/dn*	+	+
9	*cj/ei*	+	+
10	*at/au*	+	+
11	*ck/cl*	+	-
12	*bc/cm*	+	-
13	*v/cn*	-	-
14	*co/cp*	+	-
15	*dq/dr*	+	-
16	*ds/du*	+	-
17	*bo/bp*	+	-
18	*bq/cq*	+	-
19	*cr/bk*	+	-
20	*cs/bj*	+	-
21	*ct/bh*	+	-
22	*dx/dy*	+	-
23	*dz/ea*	+	-
24	*ej/ek*	+	-

*Primers are described in Table 2.3.
'+' PCR product of expected size; '-' no PCR product.
A_{600} 0.17 = Mid-log phase; A_{600} 0.25 = Stationary phase.

expression of ORFs 5 – 10 was detected from both exponential and stationary phase cultures. Expression of ORFs 11 and 12 and ORFs 14 – 24 was detected only in exponentially grown strain 1861. A similar experiment conducted using cDNA from THY cultures supported the differential expression between growth phases observed for SB (data not shown). Therefore, the growth phase-sensitive differential expression observed in SB was not a media-specific phenomenon. ORF 13 amplification was not detected under either condition shown in Table 4.3, which may indicate that the gene was not expressed under the conditions used. However, given that the results of BLASTp searches in Section 3.3.2.2 suggested that ORF 13 was the 3' missing fragment of the Tn*5252* relaxase encoded by ORF 12, it seemed unlikely that ORF 13 would have been expressed independently of ORF 12. Perhaps the mRNA of ORF 13 was especially prone to degradation when compared to ORF 12, particularly given its likely position at the 3' end of the mRNA transcript and that only a weak band was observed for ORF 12 expression. Alternatively, the presence of a putative stem-loop structure within ORF 13 (12,433 – 12,401 [ΔG= -14.7]), could have led to the early termination or significantly reduced transcription of ORF 13.

In view of the above findings, cDNA derived from strain 1861 in the exponential phase was used as the template for amplification of regions spanning consecutive ORFs. Amplification between consecutive ORFs was only attempted when expression of both ORFs was detected (Table 4.3). The results of amplification spanning ORF junctions (Table 4.4) led to the prediction that 8 different mRNAs were transcribed from the PPI-1 variable region of strain 1861 (Figure 4.1).

ORFs 5 and 6, which together encode *nplT*, exist on the same transcript and a putative promoter was identified upstream of ORF 5 at position 3,884 – 3,929 (*t – g*, strain 1861 PPI-1 sequence [Section 3.3.1]) with a score of 0.96 and a transcription start site at position 3,920. In addition, a putative Rho-independent terminator was identified

Table 4.4 Amplification of consecutive PPI-1 variable region ORFs from strain 1861 cDNA

ORFs	Primers*	PCR (- / +)
5 – 6	*s/cu*	+
6 – 7	*dk/ap*	-
7 – 8	*cv/cw*	-
8 – 9	*dm/cx*	+
9 – 10	*cj/j*	+
10 – 11	*at/cy*	-
11 – 12	*cz/cm*	+
14 – 15	*co/da*	+
15 – 16	*dq/db*	-
16 – 17	*bn/dc*	+
17 – 18	*dd/cq*	+
18 – 19	*bq/bm*	+
19 – 20	*cr/bj*	+
20 – 21	*cs/bh*	+
21 – 22	*dg/bf*	-
22 – 23	*dh/bb*	-
23 – 24	*di/aw*	+

*Primers are described in Table 2.3.
PCR was performed using cDNA template derived from stationary phase culture.

downstream of ORF 6 between positions 5,981 and 6,020 (ΔG= -13.4). However, whilst both ORFs 5 – 6 appear to be expressed, it is not clear whether ORF 6 is indeed translated. The absence of an obvious Shine-Dalgarno (SDg) sequence upstream of the ORF 6 start codon did not support the translation of ORF 6. However, it is also true that the lack of an obvious SDg sequence could reflect the prevalence of atypical ribosome-binding sites (RBS) within the *S. pneumoniae* genome (reviewed in Lacks [1997]), thus making it difficult to confirm whether ORF 6 has a RBS. Therefore, ORF 6 could not be ruled out as a protein-encoding ORF. In addition, it is not clear whether separate ORF 5- and ORF 6-encoded peptides can produce a functional enzyme, as it might be possible that the ORF 5-encoded peptide can complement the activity of the ORF 6 product, similar to *lacZ*-α, and *lacZ*-ω in *E. coli*. Alternatively, a non-functional peptide might be translated from ORF 5, leaving ORF 6 to carry out its function without the N-domain (Section 3.3.2.2).

The second mRNA encodes the hypothetical protein of ORF 7, which has a putative promoter from 5,846 – 5,891 (score 0.98) and a transcription start site at 5,882. However, a putative transcription terminator was not identified downstream of ORF 7, which indicates that a more discrete terminator exists between ORFs 7 and 8 that prevents amplification spanning this ORF junction.

The third mRNA transcript of the region encodes PezAT and a hypothetical protein (ORF 10), which was expected from previous work (Brown *et al*., 2004; Khoo *et al*., 2007). The *pezAT* promoter that was identified by Khoo *et al*. (2007) was also found in strain 1861 from position 6,831 – 6,876 with a transcription start site of 6,867.

The fourth mRNA transcript includes the Tn*5252*-associated ORFs, 11 and 12, and has a putative promoter from 9,977 and 10,022 (score 0.94) with a transcription start site at position 10,013. Whilst a putative transcription terminator exists within ORF 13 (see above), its position on the reverse strand suggests that it terminates the transcription of the fifth mRNA transcript. However, the presence of the stem loop associated with this transcription terminator could inadvertently retard the transcription of ORF 13.

The fifth mRNA transcript encodes the putative Rgg/GadR/MutR family transcriptional regulator of ORFs 14 and 15. However, despite the likely inactivation of the transcriptional regulator encoded by ORFs 14 and 15 due to the fragmentation discussed in Section 3.3.2.2, active transcription of the gene was shown to be maintained in strain 1861 *in vitro*. A putative promoter was identified upstream of ORF 15 from position 13,700 – 13,655 with a score of 0.99 and a transcription start site at 13,664. A putative transcription terminator was identified downstream of ORF 14 from 12,433 – 12,401 (ΔG= -14.7), which may terminate the transcription of the ORF 15 – 14 mRNA transcript.

The sixth mRNA transcript includes ORFs 16 – 21, which encode a number of metabolic enzymes and proteins of unknown function. A putative promoter was identified upstream of ORF 16 from position 13,852 – 13,901 with a score of 0.97 and a transcription start site at 13,892. However, a putative transcription terminator that could explain the lack of amplification spanning the junction of ORFs 21 and 22 was not identified. Therefore, a more discrete transcription terminator, such as some Rho-dependent terminators, is likely to exist between the two ORFs.

The biotin carboxylase encoded by ORF 22 was the only gene encoded on the seventh mRNA transcript of the region. A putative promoter was also identified upstream of ORF 22 from position 18,948 – 18,997 with a score of 1.00 and a transcription start site at position 18,988. The position of this putative promoter suggested that the ORF 22 start codon was probably located at position 19,050 rather than at position 18,804 (Table 3.3). This alternative start site was also shared by the annotation of the PPI-1-associated biotin carboxylase in 11-BS70 (BLASTx, NCBI). In addition a putative SDg sequence was identified upstream of the putative ORF 22 start codon at position 19,050. The final mRNA transcript in the PPI-1 variable region of strain 1861, includes ORFs 23 and 24, which were predicted to encode a fragmented transporter of the major facilitator superfamily (Section 3.3.2.2). A putative promoter was identified upstream of ORF 23 from position 20,426 – 20,471 with a score of 0.99 and a transcription start site at position 20,462. Again, as was found for ORFs 5 and 6, and ORFs 14 and 15, ORFs 23 and 24 were both expressed *in vitro* despite being fragmented by a premature stop codon. However, whilst a SDg sequence was not identified upstream of the ORF 24 start codon, an atypical RBS may exist to allow the translation of ORF 24.

In summary, the ORFs of the PPI-1 variable region in strain 1861 were shown to be transcribed on 7 separate mRNA transcripts. Therefore, it was now possible to select a group of genes for *in vivo* expression analysis that could reflect the expression of other genes on the same transcript.

4.3 *In vivo* expression analysis of the PPI-1 variable region of strains 1861 and 4496

4.3.1 Pathogenesis of strains 1, 1861 and 4496 using an intranasal model of infection

Having conducted a preliminary analysis of the expression of the PPI-1 variable region *in vitro*, the next step was to assess the expression *in vivo* in order to establish whether or not certain genes within this region of strains 1861 and 4496 were preferentially expressed in particular niches of the mouse. *In vivo* transcriptional activity data may provide some indication as to a possible role for these genes in virulence. However, in order to perform *in vivo* gene expression comparisons, the ability of strains

1, 1861 and 4496 to colonise the nasopharynx and invade and survive in the blood and lungs was assessed using an intranasal mouse model of infection, described in Section 2.13.2. In addition to determining CFU in various niches, infected tissue samples were stored as sources of RNA to be used in the *in vivo* expression analysis of Section 4.3.2. Briefly, 30 CD-1 mice per invasive strain and 20 CD-1 mice in the strain 1 group were challenged whilst under anaesthesia with approximately 10^7 CFU of the relevant strains, as described in Section 2.13.2. As stated above, an important part of this animal experiment was to obtain infected tissue samples for gene expression analysis of the PPI-1 variable region in strains 1861 and 4496. However, the group of strain 1-challenged mice was used as a negative control for IPD. Following challenge of the mice, the actual doses given were determined retrospectively and are indicated in Table 4.5.

Table 4.5 Actual challenge dose used for i.n. challenge of mice

Strain	Actual challenge dose
1	2.7×10^6
1861	9.6×10^6
4496	9.8×10^6

At 48 h and 96 h post-challenge, mice were selected from each group for enumeration of pneumococci and for extraction of bacterial RNA from nasal wash, nasopharyngeal tissue, blood and lung samples, as described in Section 2.13.4. The 48 h time point was chosen as previous data showed that terminal infection of mice challenged with either strain 1861 or 4496 occurred within approximately 60 h post-challenge and mice surviving beyond this time did not succumb to infection (Harvey, 2006). In addition, the 96 h time point was chosen to reflect the situation in the target tissues of mice that had not succumbed to infection.

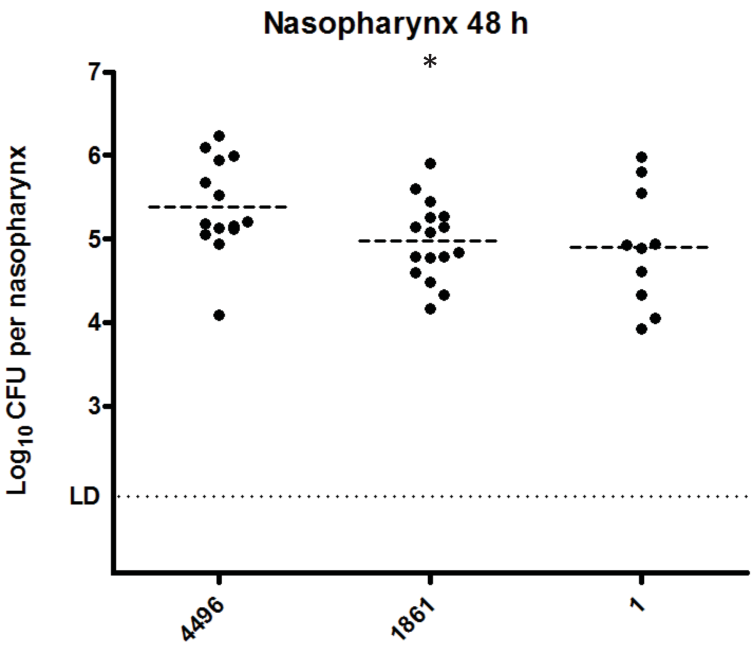
At 48 h, 16 and 14 mice that had been challenged with strains 1861 and 4496, respectively, were humanely killed, as described in Section 2.13.3. Of the 14 (1861) and 16 (4496) mice remaining in each group, 5 (1861) and 3 (4496) mice did not survive to

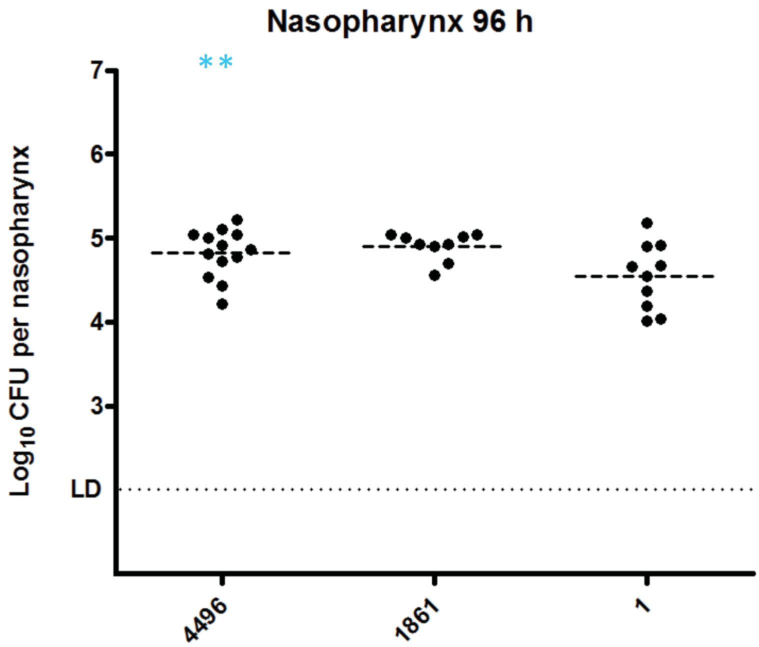
96 h post-challenge. All remaining 9 (1861) and 13 (4496) mice challenged with either of the two highly invasive strains were humanely killed at 96 h post-challenge for analysis as undertaken at the 48 h time point. As all mice challenged with strain 1 appeared healthy at both time points, 10 mice were randomly selected for bacterial counts from nasal wash, nasal tissue, blood and lung samples at 48 h and 96 h post-challenge. Bacterial counts of all three strains were undertaken on BA supplemented with gentamicin, as described in Section 2.13.3.

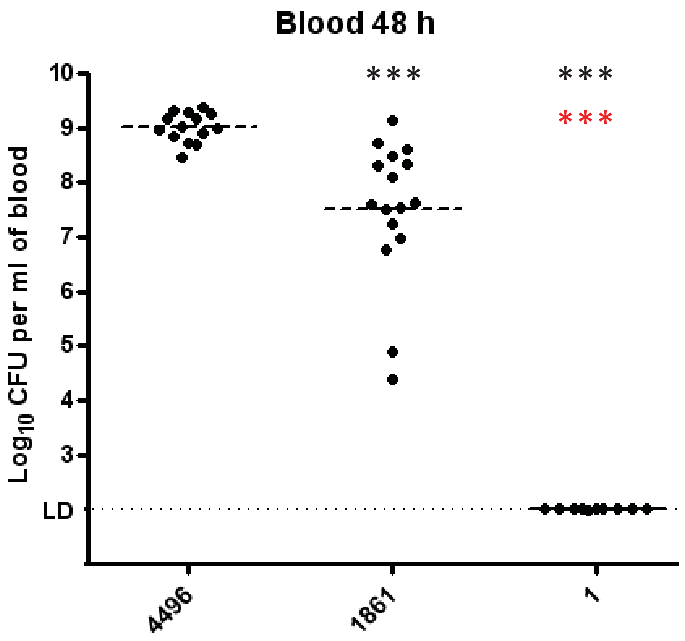
In all three groups, numbers of pneumococci from nasal wash and nasopharyngeal tissue samples of individual mice were combined to determine the total number of pneumococci within the nasopharynx (Figure 4.2). At 48 h there was no significant difference between colonisation of the nasopharynx between strain 1 and either strain 1861 or 4496. However, a significantly greater number of pneumococci were detected in the nasopharynx of 4496-infected mice than 1861-infected mice at this time point (P<0.05). A significant difference in colonisation was not detected between any strains at 96 h (Figure 4.2), which suggested that when strains 1861 and 4496 failed to cause IPD, their ability to colonise the nasopharynx was not different from that of strain 1. In addition, neither colonisation by strain 1 nor strain 1861 was different between 48 h and 96 h (P>0.05). In contrast, significantly greater colonisation by strain 4496 was detected at 48 h when compared to 96 h (P<0.01). Figure 4.2 shows that there were significantly more pneumococci in the blood of 4496-infected than 1861-infected mice at 48 h (P<0.001). In addition, both strains 1861 and 4496 were detected in significantly greater numbers in the blood than strain 1 (P<0.001), as the number of strain 1 CFU in the blood was below the limit of detection in all mice. No significant difference was detected between the numbers of strain 1861 and strain 4496 pneumococci in the lungs. However, the number of strain 1 pneumococci in the lungs was significantly lower than both strains 1861 and 4496 (P<0.001), and below the limit

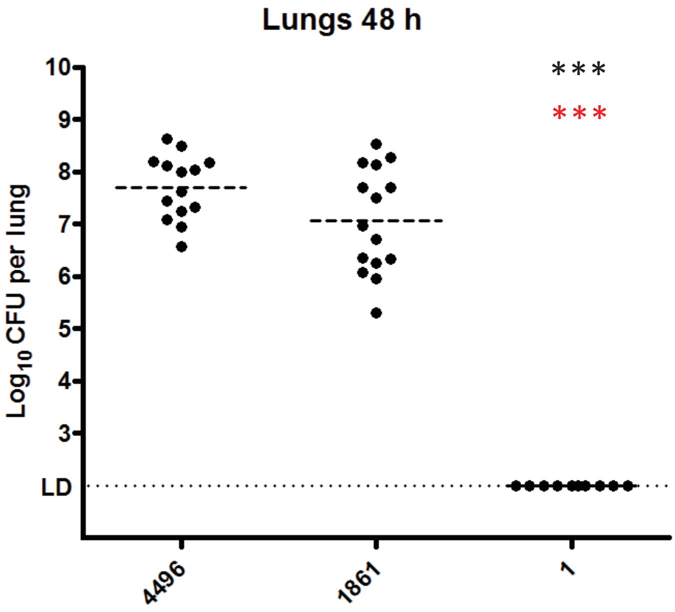
Figure 4.2 Number of pneumococci detected in the nasopharynx, blood and lungs following pneumonia sepsis challenge

Groups of 30 mice for each invasive strain (strains 1861 and 4496) and a group of 20 mice for strain 1 were challenged via the i.n. route whilst under anaesthesia with ~10^7 CFU of the indicated opaque-phase strain, as described in Section 2.13.2. The $\log_{10}$ of the mean number of pneumococci detected in each niche of each mouse from duplicate platings is plotted as a single spot and broken lines in each group indicate the geometric mean at 48 h and 96 h (nasopharynx only) post-challenge, which were determined as described in Section 2.13.3. Statistical differences were analysed by two-tailed unpaired *t*-test on log-transformed values (*, *P*<0.05; **, *P*<0.01; ***, *P*<0.001). Black asterisks are compared with 4496, red with 1861 and blue between time points of the same strain. The Limit of detection (LD) is indicated as a broken line at 1×10^2 CFU/nasopharynx, 1×10^2 CFU/ml blood and 2×10^2 CFU/lung.


Nasopharynx 48 h
*
Log$_{10}$ CFU per nasopharynx
7
6
5
4
3
LD
4496
1861
1


Nasopharynx 96 h
**
Log$_{10}$ CFU per nasopharynx
7
6
5
4
3
LD
4496
1861
1


Blood 48 h
Log10 CFU per ml of blood
10
9
8
7
6
5
4
3
LD
***
***
***
4496
1861
1


Lungs 48 h
Log10 CFU per lung
10
9
8
7
6
5
4
3
LD
***
***
4496
1861
1

of detection in all mice. In summary, some variation was detected in the ability of strains 1861 and 4496 to survive in the blood and the nasopharynx. However, this variation was comparatively minor given that strain 1 was not detected in the blood or lungs of any mice at any time point, and provided a clear distinction between the invasive potential of strains 1861 and 4496 compared to strain 1.

4.3.2 Differential expression of select PPI-1 variable region genes in different niches of the mouse

Having performed the virulence analysis and collected the tissue samples in Section 4.3.1, the *in vivo* expression analysis of the PPI-1 variable region in 1861 and 4496 could now be performed. Since strain 1 was only detected in the nasopharynx, the analysis of niche-specific expression patterns of the PPI-1 variable region in this strain would not be possible and so was not included in this section. ORFs 6, 8, 10, 15, 16, 19, 22 and 23 were chosen for expression analysis as these genes represented all transcripts of the region (Section 4.2), with the exception of ORF 7 and the Tn*5252* region. RNA was extracted from nasal wash, blood and lung samples obtained in Section 4.3.1, as described in Sections 2.12.1 and 2.13.4. Nasopharyngeal tissue samples were frequently contaminated with other species of bacteria and so were not suitable for gene expression analysis. Extracted RNA was DNase treated and checked for contaminating DNA by RT-PCR, as described in Section 2.12.1. RNA samples were pooled from groups of four mice of the 48 h time point following DNase treatment. Pooled lung-derived DNase-treated RNA underwent enrichment for prokaryotic RNA, as described in Section 2.12.2. As nasal wash and blood samples were likely to contain much smaller quantities of eukaryotic RNA, enrichment was not performed on these samples. The quantity of prokaryotic RNA present in *in vivo* samples was quite low, hence RNA amplification was performed on all samples following DNase treatment and enrichment (if required), as described in Section 2.12.3. Where one round of amplification generated insufficient quantities of RNA for the intended downstream applications, a second round was performed. Amplification was performed from each niche for each strain, producing a total of six samples. The expression of the selected ORFs was then compared between each niche of the same strain by real-time RT-PCR, as described in Sections 2.12.6 and 2.12.7. Reactions were performed in technical triplicates and additional duplicate reactions were performed for each primer set in the absence of template, to quantify background amplification. The primers used to detect expression of the target genes are shown in Table 4.6. For each reaction, 150 – 300 ng of target template RNA was used. The exact quantity of template varied between samples due to differences in the quantity of contaminating eukaryotic RNA and also due to differences in the success of RNA amplification for samples derived from different niches. Differences in the amount of RNA between samples were normalised using the quantity of 16S rRNA detected in each sample. The fold difference in the relative amount of target transcript between two niches was determined using the ∆Ct method as described in Section 2.12.7. Differences in relative expression of less than 2.0 were deemed to be insignificant in this study. Statistical significance between the relative amounts of target mRNA in different niches was calculated by comparing the mean amount of target mRNA relative to 16S rRNA in each niche using the two-tailed unpaired *t*-test, where $P<0.05$ was considered statistically significant. A number of gene transcripts, particularly in the nasal wash samples, were not detected at cycles prior to that of the no-template controls and so expression was considered to be below the limit of detection. Therefore, the no-template control Ct values were used to quantify the least possible difference in relative expression that could be detected in this experiment. In addition, melt-curve analysis was conducted to ensure that only a single product was amplified in each reaction and to ensure that the melt temperature of each amplified product was the same as that amplified from *in vitro* broth culture RNA for each strain, the product sizes of which had been confirmed by agarose gel electrophoresis (Section 2.9.1). Where melt-curve analysis detected only the presence of a product with a significantly different melt temperature to that of the target, non-specific amplification was assumed to have occurred, which implied that the desired target was below the limit of detection. Figure 4.3 shows the expression of individual PPI-1 variable region genes relative to 16S rRNA of strains 1861 and 4496 from the nasal wash, blood and lungs of infected mice. Significant differences were detected in the expression of individual genes between niches and are shown in Tables 4.7, 4.8 and 4.9.

Table 4.6 Primers used for real-time RT-PCR amplification of select PPI-1 variable region genes in strains 1861 and 4496

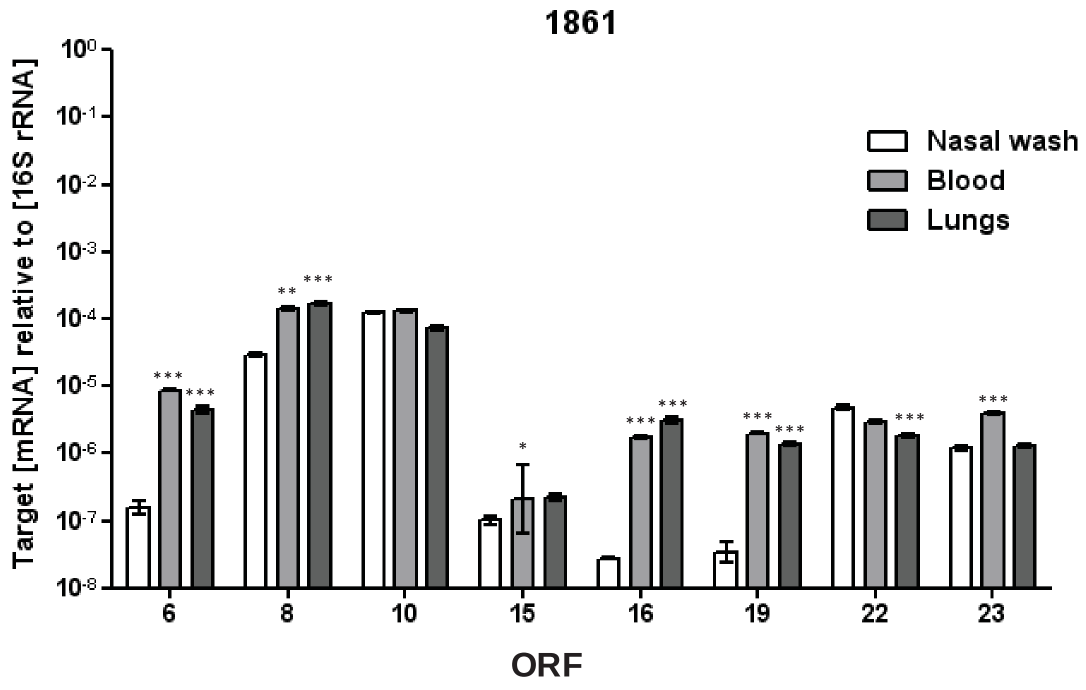
ORF	Primers
16S	RH16SF$_{(3)}$/ RH16SR$_{(3)}$
6	*dk/dl*
8	dm/dn
10	do/dp
15	dq/dr
16	ds/du
19	dv/dw
22	dx/dy
23	dz/ea

*Primers are described in Table 2.3

Table 4.7 shows that in both strains 1861 and 4496 the relative expression of ORFs 6, 8, 15, 16, 19 and 23 was significantly greater in the blood than at the nasopharyngeal surface. Of particular interest was the inability to detect transcripts for ORFs 6, 16 and 19 from the nasal wash-derived RNA of either strain. In addition, the relative expression of ORFs 10 and 22 was greater in the blood than the nasal wash of strain 4496-infected mice, but not in strain 1861-infected mice. A lack of consistency in the relative expression of ORFs 10 and 22 between strains may have reflected strain-specific differences, or differences in expression that were independent of the niche. In addition, it was puzzling that the changes in expression of ORFs 8 and 10 were different

Figure 4.3: Expression of PPI-1 genes of strains 1861 and 4496 in the nasopharynx, lungs and blood compared to 16S rRNA

The amount of target mRNA relative to 16S rRNA in the nasal wash, blood and lung samples of 1861-infected and 4496-infected mice was determined by real time RT-PCR (Section 2.11.6). Error bars indicate the standard deviation of triplicate reactions for each gene per niche. Statistical significance between the relative expression of individual genes in different niches was determined by unpaired *t*-test (*, $P<0.05$; **, $P<0.01$; ***, $P<0.001$). Black asterisks indicate comparison with nasal wash and red with blood.


1861
Target [mRNA] relative to [16S rRNA]
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
Nasal wash
Blood
Lungs
6
8
10
15
16
19
22
23
ORF

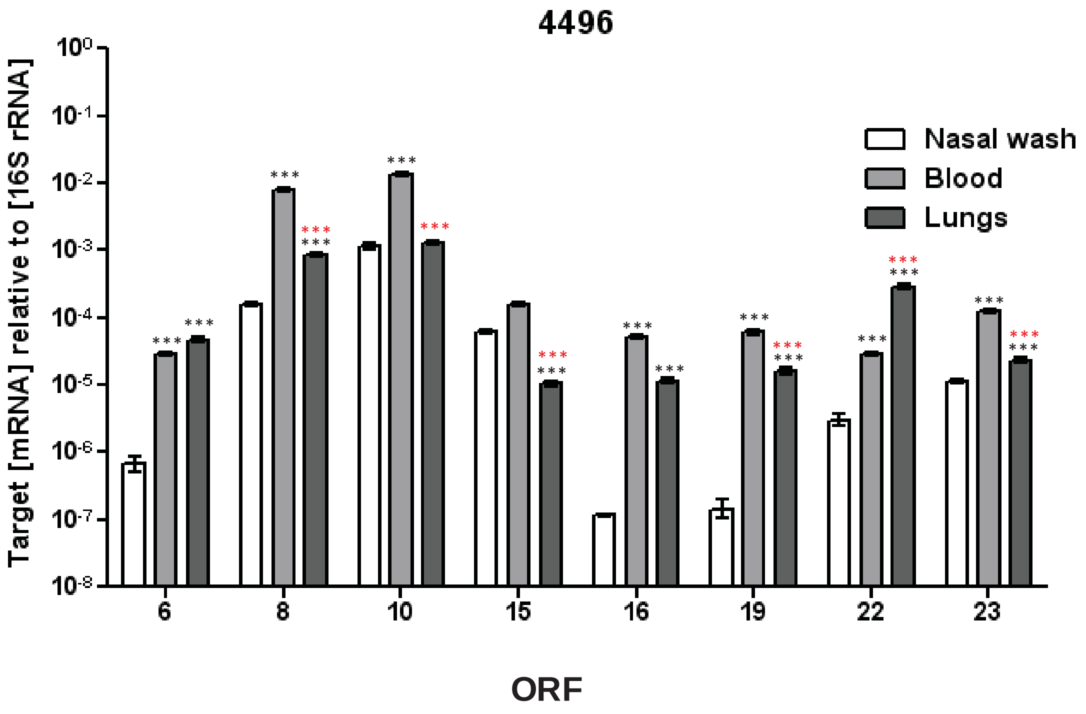

4496
Target [mRNA] relative to [16S rRNA]
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
Nasal wash
Blood
Lungs
6
8
10
15
16
19
22
23
ORF

Table 4.7 Relative expression of PPI-1 genes in the blood versus the nasal wash

ORF	1861		4496	
	Fold change	Direction	Fold change	Direction
6	>55.72*^{c}	up	>42.62*^{c}	up
8	4.75^{b}	up	48.95^{c}	up
10	1.06^{ns}	up	11.71^{c}	up
15	2.06^{a}	up	2.51^{a}	up
16	>64.59*^{c}	up	>432.53*^{c}	up
19	>58.49*^{c}	up	>414.91*^{c}	up
22	1.63^{ns}	down	97.01^{c}	up
23	3.27^{c}	up	10.95^{c}	up

*Indicates where the target was not detected from nasal wash-derived RNA

Results of statistical analysis: ns, not significant (includes values <2); *a*, *P*<0.05; b, *P*<0.01; *c, P*<0.001.

Table 4.8 Relative expression of PPI-1 genes in the lungs versus the nasal wash

ORF	1861		4496	
	Fold change	Direction	Fold change	Direction
6	>28.38*^{c}	up	>69.07*^{c}	up
8	5.79^{c}	up	5.43^{c}	up
10	1.70^{ns}	down	1.14^{ns}	up
15	2.20^{ns}	up	5.76^{c}	down
16	>115.89*^{c}	up	>97.46*^{c}	up
19	>40.79*^{c}	up	>111.95*^{c}	up
22	2.61^{c}	down	9.78^{c}	up
23	1.08^{ns}	up	2.04^{c}	up

*Indicates where the target was not detected from nasal wash-derived RNA

Results of statistical analysis: ns, not significant (includes values <2); *a*, *P*<0.05; *b, P*<0.01; *c, P*<0.001.

Table 4.9 Relative expression of PPI-1 genes in the blood versus the lungs

ORF	1861		4496	
	Fold change	Direction	Fold change	Direction
6	1.96^{ns}	up	1.62^{ns}	down
8	1.22^{ns}	down	9.02^{c}	up
10	1.80^{ns}	up	10.29^{c}	up
15	1.07^{ns}	down	14.49^{c}	up
16	1.79^{ns}	down	4.44^{c}	up
19	1.43^{ns}	up	3.71^{c}	up
22	1.60^{ns}	up	9.92^{c}	up
23	3.03^{c}	up	5.36^{c}	up

Results of statistical analysis: ns, not significant (includes values <2); *a*, *P*<0.05; b, *P*<0.01; *c, P*<0.001.

from each other given that they are co-transcribed (Figure 4.1). However, it has since been discovered that the serotype 1 strain P1031 (ST303) possesses a second chromosomal TA system that is homologous to *pezAT* (*SPP_1188* and *SPP_1189*) and would be expected to bind the primers *dn* and *dm* (Table 4.6). In contrast, this second TA system lacked a homolog of ORF 10 and had very different promoter sequence from *pezAT*. Therefore, it is possible that a second TA system could be present in strains 1861 and 4496 that could interfere with the quantitation of *pezAT* expression. In addition, the absence of potential stem-loop structures between ORF 8 and ORF 10 suggests that ORF 10 expression probably reflects the expression of the *pezAT* operon.

Significantly greater expression of ORFs 6, 8, 16 and 19 was detected in the lungs than at the nasopharyngeal surface for both strain 1861- and 4496-infected mice (Table 4.8). Differential expression of ORF 10 was not detected between the two niches for either strain, which indicates that expression of *pezAT* probably does not change between the nasopharyngeal surface and the lungs of infected mice. Differential expression of ORF 23 was not detected between the lungs and nasopharyngeal surface of strain 1861-infected mice, whereas a small but significant increase in expression was detected in the lungs when compared to the nasopharyngeal surface of 4496-infected mice. The expression pattern of ORF 15 differed between the two strains. No significant difference in expression was detected between the nasopharyngeal surface and lungs of 1861-infected mice, whereas expression of ORF 15 was significantly lower at the nasopharyngeal surface of 4496-infected mice. Similarly, significantly increased expression of ORF 22 was detected at the nasopharyngeal surface for 1861-infected mice when compared to the lungs, whereas the opposite was true for 4496-infected mice. Therefore, expression of ORFs 15 and 22 did not appear to reflect niche-specific responses that were consistent between strains suggesting that a combination of strain-specific and other niche-independent factors might be responsible for these differences in expression.

With the exception of ORF 23, there was no detectable difference in the expression of the PPI-1 variable region genes in 1861 between the blood and the lungs (Table 4.9). In contrast, the expression of ORFs 8, 10, 15, 16, 19 and 22 was greater in the blood than the lungs of 4496-infected mice. These differences perhaps reflected strain-specific requirements for the PPI-1 variable region genes in the blood and lungs rather than niche-specific expression. Sequence alignments between the PPI-1 variable region of strains 1861 and 4496 were analysed for differences within promoter sequences that could potentially alter the regulation of expression, but none were identified. Whilst somewhat speculative, a possible alternative cause could relate to the significant difference that was observed in the number of pneumococci in the blood of strain 1861-infected compared to 4496-infected mice (Figure 4.2). For example, differences in the severity of bacteraemia could cause microenvironmental alterations that impact on gene expression of the PPI-1 variable region. However, such claims require experimental evidence. In contrast, consistently greater expression of ORF 23 was detected in the blood than the lungs, which suggested a possible greater requirement for the product of this gene in the former niche.

In summary, it was clear that the expression of ORFs 6, 16 and 19 was consistently greater in the lungs and blood compared to the nasopharynx, and that the expression of ORF 23 was greater in the blood than both the lungs and nasopharyngeal surface. In contrast, the expression of ORFs 10 and 22 in different niches was rarely consistent between strains, making it difficult to conclude that these genes exhibit niche-specific expression. In addition, due to the potential presence of a second TA system homologous to ORF 8, the expression of ORF 10 might have better reflected the expression of the PPI-1 *pezAT* locus. Therefore, it was not clear whether *pezAT* exhibits niche-specific expression. As only strains 1861 and 4496 were able to invade and survive in blood (Section 4.3.1), the preferential expression of ORFs 6, 16 and 19 by strains 1861 and 4496 in the lungs and the blood of infected mice supports the notion that the expression of the PPI-1 variable region may at least partly contribute to the invasive potential of these strains.

4.4 The effect of mutagenesis of PPI-1 in D39 on virulence in the mouse

Due to the inability to transform the serotype 1 strains used in this study (Chapter 6), the effect of mutagenesis of the PPI-1 variable region genes on pathogenicity was assessed in the readily transformable serotype 2 laboratory strain, D39 (ST454). Whilst it is possible that PPI-1-independent differences between the serotype 1 isolates and D39 may interfere with the results of mutagenesis of the PPI-1 variable region, D39 was deemed suitable as it contains a version of the PPI-1 variable region and is a well-characterised strain in both the literature and within our laboratory. Therefore, we assessed whether mutagenesis of the PPI-1 variable region could result in detectable differences in the competitive fitness of D39 within certain niches of the mouse. The impact of the strain 1 and 1861 versions of the PPI-1 variable region on the virulence of D39 was also assessed.

4.4.1 Construction of PPI-1 mutants in D39

As discussed in detail in Chapter 3, the PPI-1 variable region in the lineage A strains, the highly virulent strains (1861 & 4496) and D39 is quite different in terms of genetic composition. Therefore, derivatives of D39 in which the endogenous PPI-1 variable region was replaced with the version of the region from strain 1861 and strain 1 D39^{1861} and D39^{1}, respectively were constructed. An additional mutant was constructed that completely removed the PPI-1 variable region of D39 (D39ΔPPI-1). The primers used during the construction of the mutants in this section are shown in Table 2.3.

D39ΔPPI-1 was constructed by deleting the PPI-1 variable region of D39 and replacing it with a chloramphenicol acetyltransferase gene (*cml*^{R}), as shown in Figure 4.4. As reviewed in Van Melderen and Saavedra DeBas (2009), deleting a chromosomal TA system such as *pezAT* may be lethal, due to the rapid degradation of cytoplasmic PezA, allowing the active and stable toxin, PezT, to cause cell death. Therefore, *pezAT* needed to be removed in two stages. In the first stage, a D39 ΔPezT mutant was generated by deletion of the PPI-1 variable region sequence from the start codon of *pezT* (*SPD_0931*) to a position within non-coding sequence corresponding to the 5' end of homologous region 5 (Figure 3.6) and replacement with an erythromycin resistance cassette (*erm*^{R}). In order to increase the frequency of recombination, 3 – 4 kb of sequence flanking the region to be deleted was amplified by PCR using the primers *a* - and *aa* for the 5' product and primers *ec* and *g* for the 3' product, using D39 genomic DNA as the template. The *erm*^{R} gene was amplified from pVA831 using primers J214 and J215. Primer *aa* was designed to encode sequence complementary to J214 and primer *ec* was designed to encode sequence complementary to J215. Amplification and purification of the three individual products was followed by overlap extension PCR, using the primers *a* and *g* to join the products. D39 competent cells (prepared as described in Section 2.8.1) were transformed with the overlap extension PCR product,

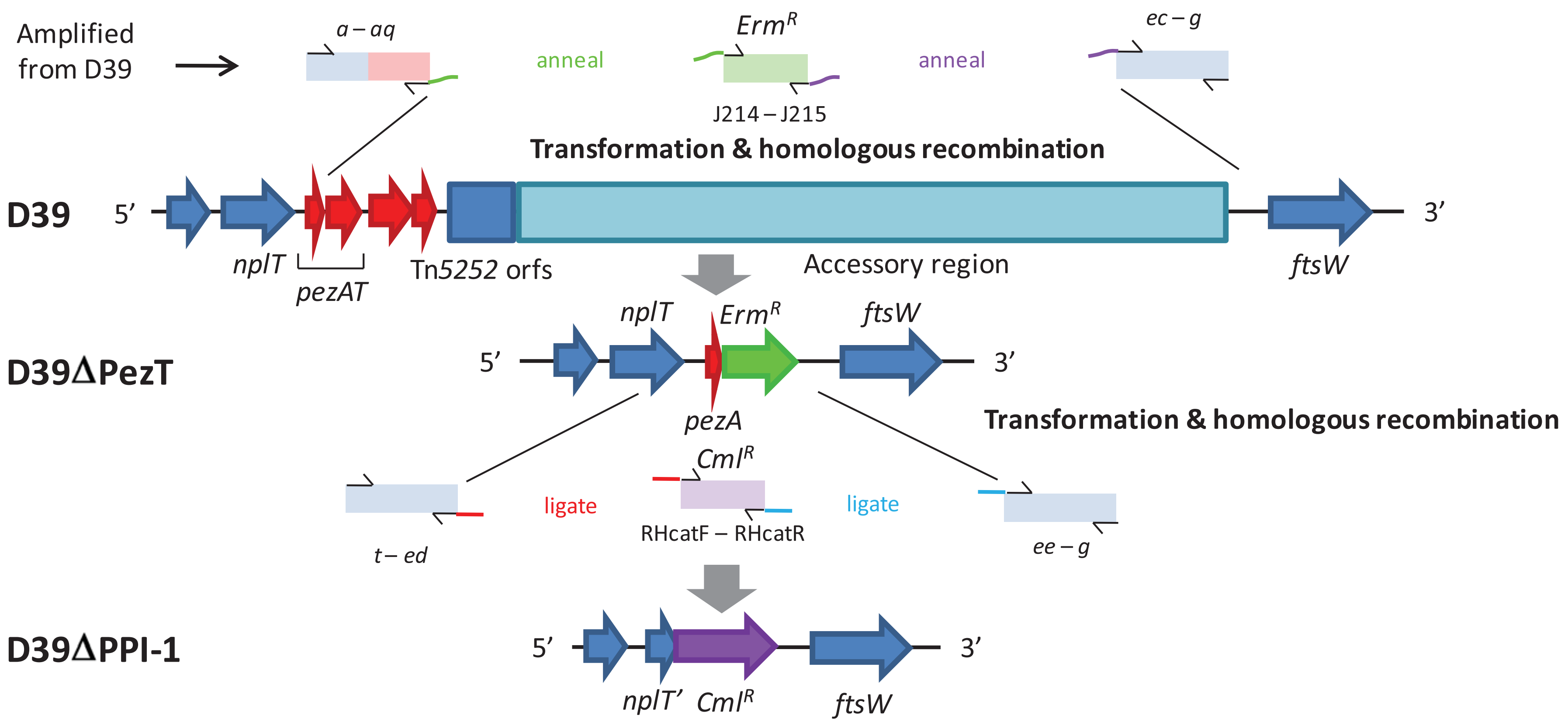


Figure 4.4 Construction of the D39 PPI-1 variable region deletion mutant, D39ΔPPI-1
D39ΔPezT was constructed by transformation of D39 competent cells (Section 2.8.1) with the overlap extension PCR product of *a* – *aq*, *ec* – *g* (from D39 DNA) and *erm*^{R} (pVA891). Selection was undertaken on BA supplemented with erm. Erm^R colonies were checked for the deletion of the D39 PPI-1 variable region from *pezT* to upstream of *ftsW* by PCR and targeted sequencing, as described in Section 2.9.8. D39ΔPPI-1 was made by transforming D39ΔPezT competent cells (Section 2.8.2) with the ligation product (Section 2.9.5) of digested *t* – ed, *ee* – *g* and *cml*^{R} PCR products (Section 2.9.4). Digestion was carried out with *Eag*I (red), *Xho*I (blue) or both as indicated. Selection was carried out on BA supplemented with cml. Cml^R colonies were checked for the complete deletion of the PPI-1 variable region in D39 by PCR and targeted sequencing (Section 2.9.8).

as described in Section 2.8.2, and recombinants were selected on BA supplemented with erm. Erm^{R} colonies were screened for replacement of the target region with the *erm^{R}* gene by PCR and sequencing, as described in Section 2.9.8. The second stage of D39ΔPPI-1 construction involved deletion of the D39ΔPezT PPI-1 variable region sequence from a position corresponding to the 3' end of homologous region 1 (Figure 3.6) to the 3' end of the *erm^{R}* gene. PCR products of sequence flanking the region to be deleted were amplified with primers *t* and *ed* for the 5' product and primers *ee* and *g* for the 3' product. The *cml^{R}* gene and promoter, which confer cml resistance to *S. pneumoniae*, were amplified from Rx1 Δply promoter:cml^{R} (Berry, unpublished), using primers RHcatF and RHcatR. Primers *ed* and RHcatF included an *Eag*I restriction site and primers *ee* and RHcatR included a *Xho*I restriction site. These enzymes were chosen as their respective target sequences were not present in the PPI-1 variable region of strains D39, 1 or 1861. The two products of flanking sequence were purified and digested with the relevant enzyme and the *cml^{R}* gene-containing product was digested with both enzymes, as described in Sections 2.9.7 and 2.9.4, respectively. The three purified products of restriction digestion were ligated to form a single product for transformation, as described in Section 2.9.5. The final product of ligation was used to transform competent cells of D39ΔPezT, and recombinants were selected on BA supplemented with cml. The cml^{R} colonies were screened for the desired mutation by PCR and sequencing (Section 2.9.8). The successful D39ΔPPI-1 mutant contained truncated *nplT* (*nplT'*) and replacement of the downstream PPI-1 variable region with the *cml^{R}* gene.

D39^{1861} was constructed by replacing the D39 accessory region with the accessory region of strain 1861 (other regions of PPI-1 are conserved between D39 and 1861 [Figure 3.8c]) (Figure 4.5). In order to replace the accessory region of D39, the spectinomycin resistance gene (*spe^{R}*) was incorporated into the 1861 accessory region.

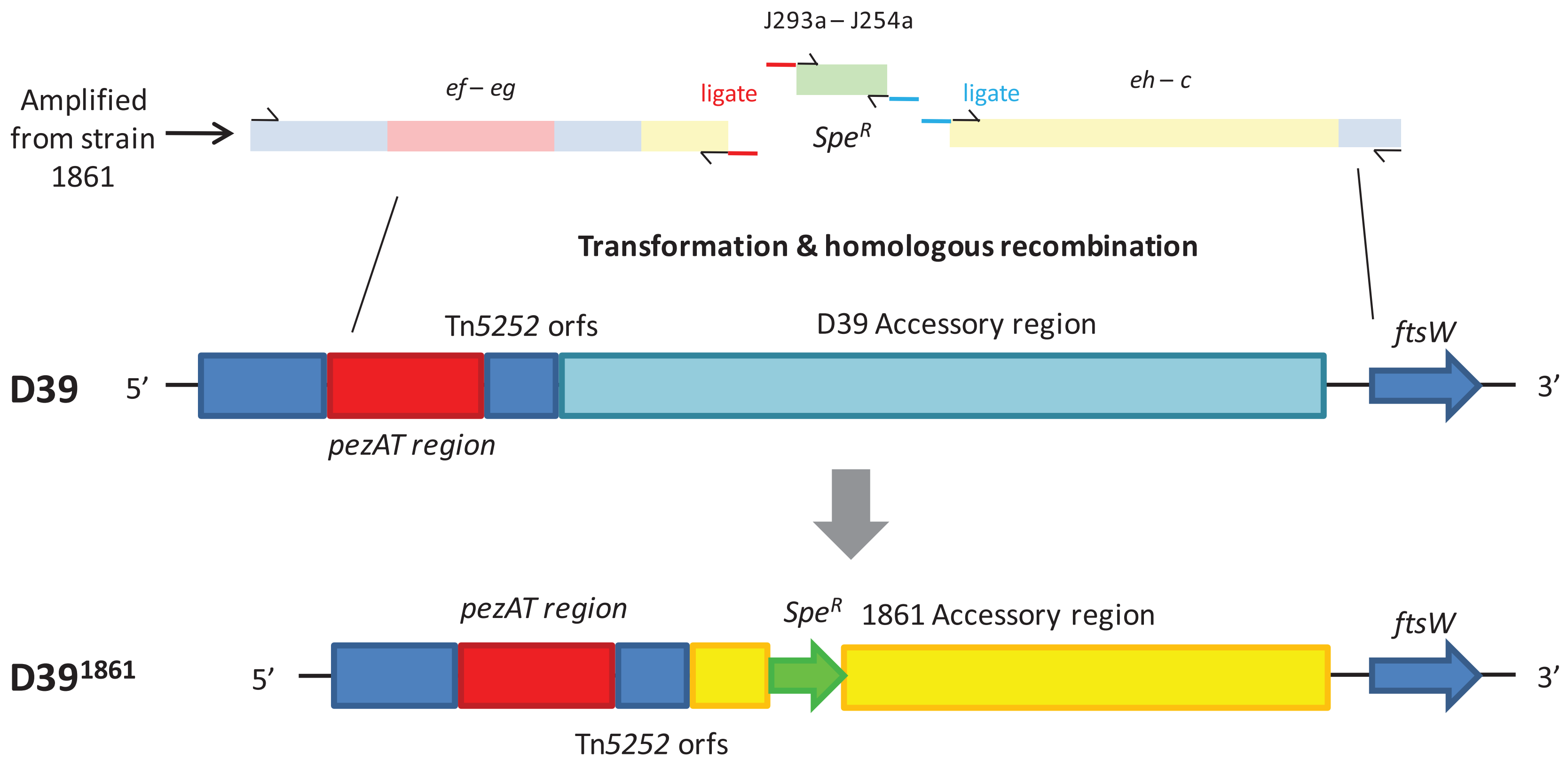


Figure 4.5 Construction of the 1861 accessory region replacement mutant of D39, D39^{1861}
The strain 1861 PPI-1 variable region was amplified in two parts using *ef – eg* and *eh –* c, which were digested by *Eag*I and *Xho*I respectively (Section 2.9.4) and ligated to *Spe^{R}* (digested by both enzymes) (Section 2.9.5). The ligation product was then used to transform wild-type D39 competent cells (Section 2.8.2) and successful Spe^{R} mutants were selected on BA supplemented with Spe. Spe^{R} colonies were checked for the presence of the complete strain 1861 accessory region by PCR and targeted sequencing, as described in Section 2.9.8.

The 1861 accessory region was amplified in two parts using the primers *ef* and *eg* and the primers *eh* and *c*. The *spe*^{R} gene and the *cps2* promoter were amplified using J293a – J254a from D39ΔcpsC:spe^R (Byrne *et al*., unpublished). The reactions *ef* – *eg* and *eh* – *c* were designed to amplify the entire region of sequence unique to 1861 and approximately 3-kb of sequence shared by both 1861 and D39 that flanked the 1861 accessory region. Primers *ea* and J293a included restriction sites for *Eag*I, and primers *eh* and J254a included *Xho*I restriction sites. The insertion site of *spe*^{R} within the PPI-1 variable region of strain 1861 was chosen to ensure that the sequence flanking the antibiotic resistance gene was unique to strain 1861, thus providing selection for the accessory region. Competent D39 cells were transformed with the final ligation product and recombinants were selected on BA supplemented with spe. *Spe*^{R} mutants were screened for the desired mutation by PCR and targeted sequencing. Sequencing showed that homologous recombination had occurred within the Tn*5252* region and upstream of *ftsW*, which confirmed that the entire accessory region from strain 1861 had replaced the equivalent region in D39.

D39^1 was constructed by replacing the PPI-1 variable region of D39 with that of strain 1. The product for making the D39^1 mutant was constructed by amplifying the strain 1 PPI-1 variable region using the primers *af* and *ei* for the 5’ product and the primers *ej* and *g* for the 3’ product (Figure 4.6). Primers *ej* and *ei* included *Eag*I and *Xho*I restriction sites respectively, to allow ligation of these products to a digested *cml*^{R} gene, which was amplified as above. Products *af* – *ei* and *ej* – *g* included sequences unique to the PPI-1 variable region of strain 1 and approximately 3 kb of sequence shared by both strain 1 and D39. The *cml*^{R} gene insertion site within the PPI-1 variable region of strain 1 was chosen to ensure that sequence flanking the site is unique to strain 1 and would thus provide selection for the region during the construction of D39^1. The final ligation product was used in a transformation of D39ΔPezT competent cells (see

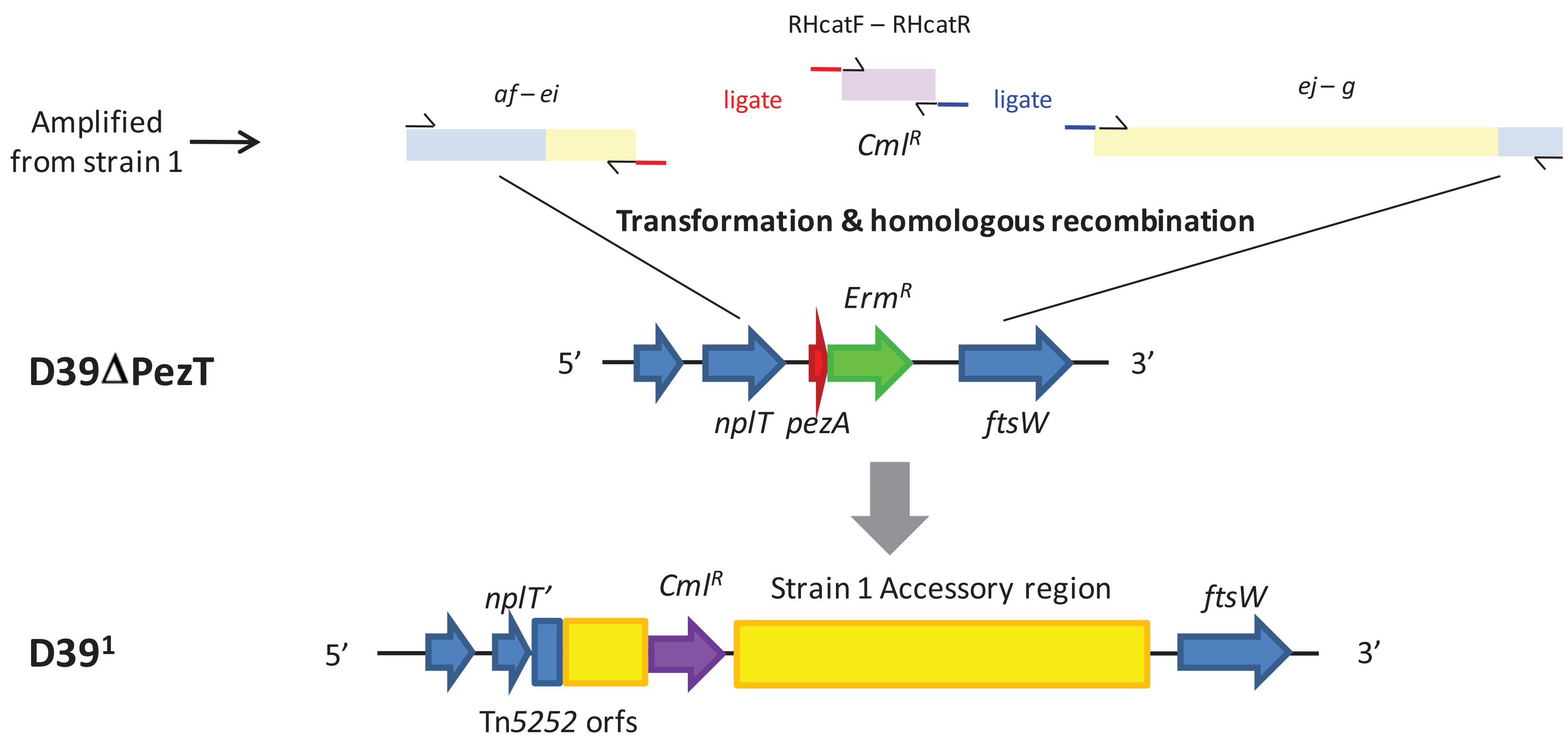


Figure 4.6 Construction of the strain 1 PPI-1 variable region replacement mutant of D39, D39^1

The PPI-1 variable region of strain 1 was amplified in two parts using *af – ei* and *ej –g,* and cml^R was inserted into the region by digesting all three products with a combination of either *Eag*I (red), *Xho*I(blue) or both as indicated (Section 2.9.4), which was followed by ligation (Section 2.9.5). The ligation product was used to transform D39ΔPezT competent cells (Section 2.8.2) and selected on BA supplemented with cml. Cml^R recombinants were checked for the presence of the complete strain 1 PPI-1 variable region by PCR and targeted sequencing (Section 2.9.8).

for D39ΔPPI-1), as *pezT* and the D39 accessory region were already absent and allowed replacement of the remaining D39 PPI-1 variable region components such as *pezA* and *nplT* with the PPI-1 variable region of strain 1 and the *cml*^{R} gene in a single step. Recombinants were selected on BA supplemented with cml. The cml^{R} mutants were screened for the desired mutation by PCR and targeted sequencing (Section 2.9.8).

Following the construction of the mutants, *in vitro* growth comparisons were performed between wild-type D39, D39^{1}, D39^{1861} and D39ΔPPI-1 to ensure that none were at a growth disadvantage. However, there was no consistent difference in the growth rate between any strains.

In summary, D39^{1} and D39^{1861} were constructed to represent the PPI-1 variable regions of the lineage A strains and the highly virulent strains, respectively. In addition, D39ΔPPI-1 was designed to assess the contribution of the D39 wild-type version of the region to virulence. Therefore, the role of the PPI-1 variable region in the invasive potential of the highly virulent strains compared to the lineage A strains could be compared.

4.4.2 Competitive index of PPI-1 variable region mutants in different niches of the mouse

The virulence/fitness of the various mutants of D39 was compared using a competitive mouse intranasal challenge model, as described in Sections 2.13.2 and 2.13.5. Five CD-1 mice per group were challenged whilst under the effects of anaesthesia with approximately 10^{7} CFU, as described in Section 2.13.2. Each challenge dose comprised approximately 5×10^{6} CFU of each competing strain. For each competition, the IR of strains used to challenge each group was determined retrospectively by plating on BA, as described in Section 2.13.5. The numbers of D39ΔPPI-1 and D39^{1} pneumococci were determined following selection on BA supplemented with cml. The number of D39^{1861} pneumococci was determined following selection on BA supplemented with spe.

Duplicate experiments were carried out for each competition and the IR values of each competition are shown in Tables 4.11 and 4.12. At 24 h and 48 h post-challenge the CI of each competition in each nice of each mouse was determined, as described in Section 2.13.5. CIs were calculated from nasal wash, nasal tissue, blood and lung samples.

Table 4.11 Input ratios for experiment 1

Competition	IR*
D39 v D39ΔPPI-1	1.4
D39 v D39^{1}	2.9
D39 v D39^{1861}	2.4
D39^{1861} v D39^{1}	1.4

*Indicates the ratio of the first strain to the second strain

Table 4.12 Input ratios for experiment 2

Competition	IR*
D39 v D39ΔPPI-1	2.9
D39 v D39^{1}	1.5
D39 v D39^{1861}	1.1
D39^{1861} v D39^{1}	0.9

*Indicates the ratio of the first strain to the second strain

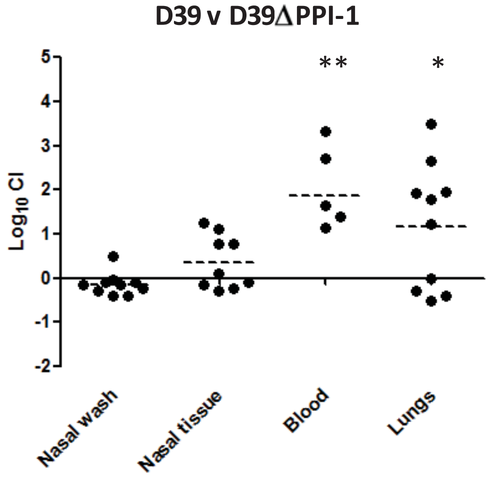
Figure 4.7 shows the state of the competitions at 24 h post-challenge, which highlighted any differences in the ability of the tested strains to establish their presence in a particular niche shortly after challenge. Figure 4.7 shows that wild-type D39 was 100-fold more fit than D39ΔPPI-1 in the blood (P<0.01) and 16-fold more fit in the lungs (P<0.05). In contrast, a difference in the fitness of the wild-type and D39ΔPPI-1 in the nasopharynx was not detected. Therefore, it appeared that the PPI-1 variable region of wild-type D39 was required for wild-type fitness in the lungs and blood, but not the nasopharynx, following the first 24 h of infection.

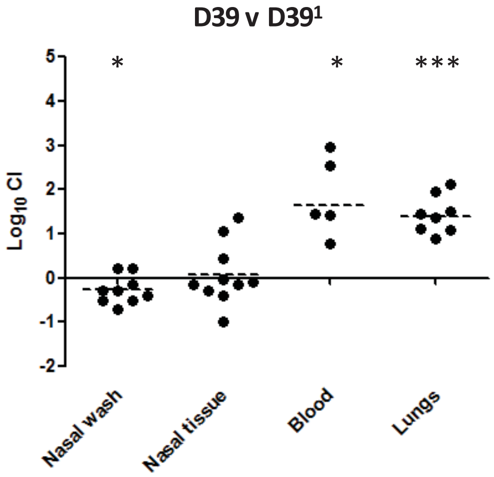
The fitness of the D39^{1} mutant was 67-fold (P<0.05) and 27-fold (P<0.001) less in the blood and lungs respectively than the wild-type, but was 3-fold greater at the nasopharyngeal surface (P<0.05). However, there was no difference in fitness within the nasopharyngeal tissue. Therefore, replacement of the PPI-1 variable region of wild-

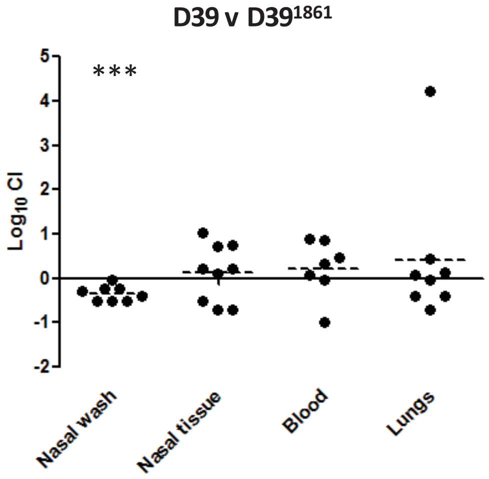
Figure 4.7 Competition between PPI-1 variable region mutants of D39

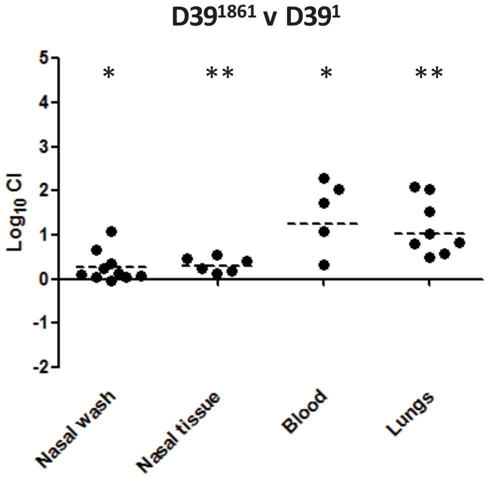
Groups of 10 mice in replicate experiments were challenged using a competitive pneumonia sepsis model with a total inoculum of ~10^7 CFU per competition (~ 5 × 10^6 CFU per strain), as described in Section 2.13.2. All challenge strains were in the opaque phase prior to challenge (Section 2.2). The CI was calculated for each niche of each mouse using the IRs of Tables 4.11 and 4.12, as described in Section 2.13.5. Log-transformed CI values from the 24 h and 48 h time points of two replicate experiments are plotted on a $\log_{10}$ scale, with the geometric mean indicated by a broken line for each niche. Where the number of pneumococci from a particular niche was below the limit of detection that group had <10 data points plotted. CI values describe the ratio of the first strain relative to the second strain, as indicated in the title of each competition. Statistical differences between the log-transformed geometric mean CI and a hypothetical value of 0 (no difference in comparative fitness) in each niche were determined using the one-sample *t*-test (*, *P*<0.05; **, *P*<0.01; ***, *P*<0.001).

24 h

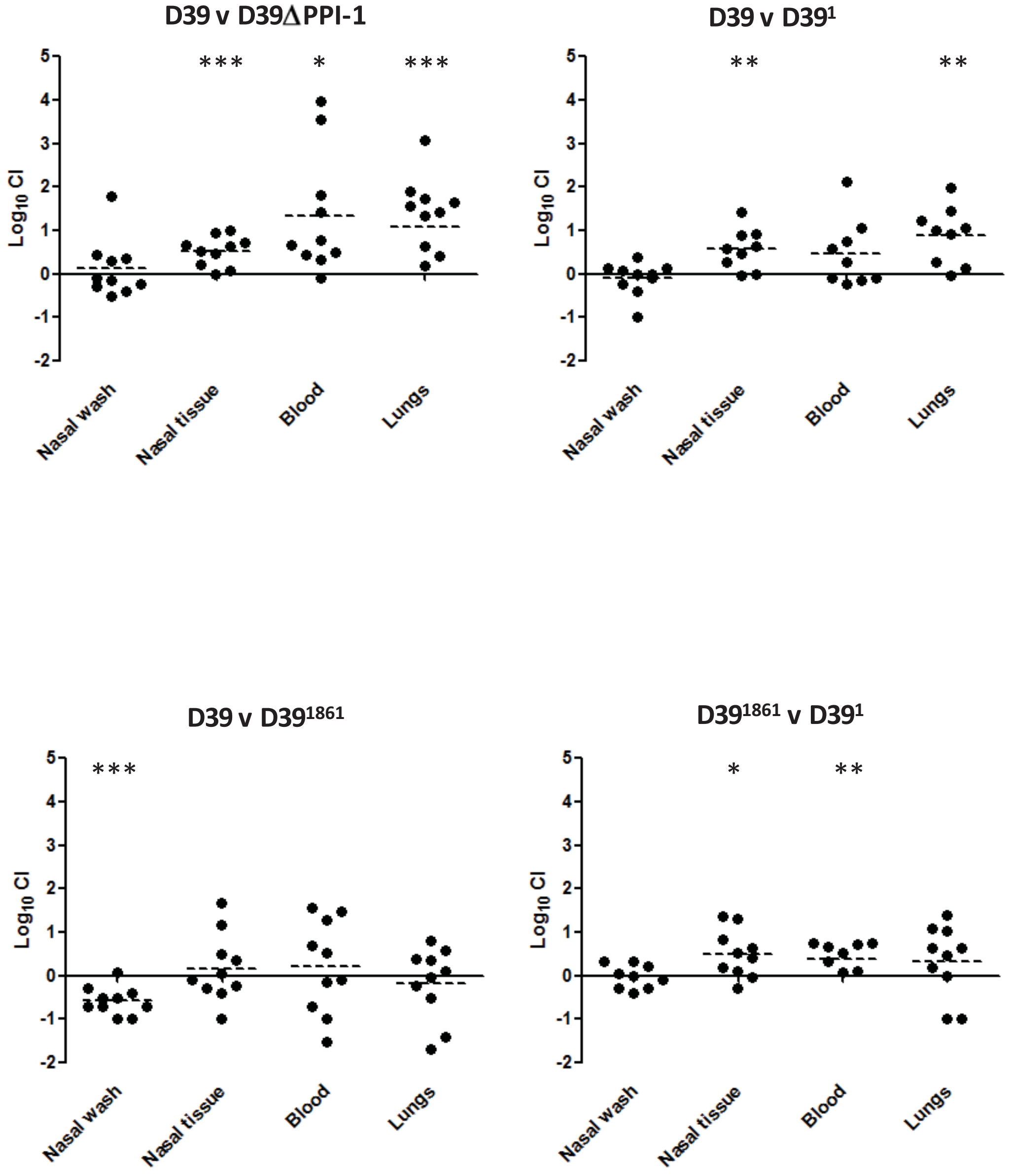

D39 v D39ΔPPI-1


D39 v D39^1


D39 v D39^{1861}


D39^{1861} v D39^1

48 h


D39 v D39ΔPPI-1
*** * ***
Log10 CI
5 4 3 2 1 0 -1 -2
Nasal wash
Nasal tissue
Blood
Lungs
D39 v D39¹
** **
Log10 CI
Nasal wash
Nasal tissue
Blood
Lungs
D39 v D39¹⁸⁶¹
***
Log10 CI
Nasal wash
Nasal tissue
Blood
Lungs
D39¹⁸⁶¹ v D39¹
* **
Log10 CI
Nasal wash
Nasal tissue
Blood
Lungs

type D39 with the strain 1 version of the region led to significant reductions in fitness in the blood and lungs and a slight increase in fitness at the nasopharyngeal surface following 24 h of infection.

In contrast to both the D39ΔPPI-1 and D39^{1} mutants, D39^{1861} was of equal fitness in the blood, lungs and nasopharyngeal tissue to that of wild-type D39. However, D39^{1861} was 2-fold more fit at the nasopharyngeal surface than the wild type (*P*<0.001). It appeared that the strain 1861 version of the PPI-1 variable region was able to maintain wild-type fitness in the blood, lungs and nasopharyngeal tissue and contribute to a small, but significant increase in fitness at the nasopharyngeal surface following 24 h of infection. Therefore, after 24 h of infection, it is either possible that the components of the PPI-1 variable region common to both D39 and D39^{1861} could be required for wild-type fitness of D39, or that additional components present in D39^{1861} may compensate for those lost from the wild type.

When compared to D39^{1}, the D39^{1861} mutant exhibited 2-fold greater fitness at the nasopharyngeal surface (*P*<0.001), 32-fold greater fitness in the blood (*P*<0.05), 16-fold greater fitness in the lungs (*P*<0.01) and 2-fold greater fitness in the nasopharyngeal tissue, which suggested that the strain 1861 version of the PPI-1 variable region could confer a greater fitness to D39 in the nasopharynx, blood and lungs following 24 h of infection, when compared to the strain 1 version of the same region.

Figure 4.7 also shows the progression of the competitions at 48-h post-challenge. Wild-type D39 exhibited 22-fold greater competitive fitness in the blood (*P*<0.05), 25-fold greater fitness in the lungs (*P*<0.001) and 3-fold greater fitness in the nasopharyngeal tissue (*P*<0.001) when compared to the D39ΔPPI-1 mutant. However, a significant difference between the fitness of wild-type D39 and D39ΔPPI-1 was not detected at the nasopharyngeal surface.

Wild-type D39 was 8-fold more competitive in the lungs (*P*<0.01) and 4-fold more competitive in the nasopharyngeal tissue than the D39^{1} mutant (*P*<0.01). However, there was no significant difference in the competitive fitness of wild-type D39 and D39^{1} was not detected at the nasopharyngeal surface or in the blood.

At 48 h post-challenge the D39^{1861} mutant exhibited 4-fold greater fitness at the nasopharyngeal surface than wild-type D39 (*P*<0.001), which was slightly greater than the 2-fold difference observed at 24 h. Similar to the 24 h time point, no difference was observed between the fitness of the wild type and mutant in either the blood, lungs or nasopharyngeal tissue. Therefore, it appears that the PPI-1 variable region of strain 1861 was able to continue to maintain wild-type fitness at 48 h post-challenge in the lungs, blood and nasopharyngeal tissue and increase fitness at the nasopharyngeal surface. When D39^{1861} and D39^{1} were compared at 48 h, a small, but significant difference in fitness was observed in the blood and nasopharyngeal tissue, with a 3-fold greater presence of D39^{1861} in both niches compared to D39^{1}.

At 48 h post-challenge the greatest reduction in fitness was observed in all niches other than the nasopharyngeal surface following the removal of the PPI-1 variable region from wild-type D39. However, this competitive fitness was completely restored by the PPI-1 variable region of strain 1861 and restored in the blood by the strain 1 version of the region. Direct comparison between D39 mutants carrying the strain 1861- and 1-derived versions of the region indicated a small but significantly increased fitness of the former in the blood and nasopharyngeal tissue, but not the lungs. In summary, the PPI-1 variable region of wild-type D39, strain 1 and strain 1861 influence the ability of the pneumococcus to survive in the lungs, blood and nasopharyngeal tissue.

4.5 Discussion

The primary aim of Chapter 4 was to build on the sequence analysis of Chapter 3 and characterise the role of the PPI-1 variable region *in vivo* in particular, to determine whether the region contributes to the differences in invasive potential between the lineage A strains and strains 1861 and 4496.

4.5.1 Transcription of the PPI-1 variable region *in vitro*

Initially the transcriptional structure of the PPI-1 variable region in both groups of strains was investigated using PCR analysis of cDNA derived from *in vitro* broth cultures of strains 1 and 1861. In strain 1, the expression of *nplT* was below the limit of detection, which may indicate that the truncated gene is non-functional *in vivo*. In addition, ORFs 7 – 11 were found to be co-transcribed and predicted to be regulated by a putative promoter identified upstream of ORF 7.

The PPI-1 variable region of strains 1861 and 4496 consisted of 8 different putative transcripts. Predictably, *pezAT* was transcriptionally coupled to ORF 10, as were the Tn*5252*-associated genes to each other. Also, despite fragmentation, *nplT* (ORFs 5 and 6), the Rgg/GadR/MutR transcriptional regulator (ORFs 14 and 15) and the major facilitator transporter (ORFs 23 and 24) were expressed *in vitro*. In some cases, such as *nplT* and the major facilitator, these fragmented genes may still give rise to functional proteins. However, based on previous work (Hollands *et al*., 2008) it is unlikely that the transcriptional regulator would be functional in strains 1861 and 4496. The largest mRNA transcribed from the region included ORFs 16 – 21, which encoded enzymes such as 3HIBDH, PDT and GalE and three hypothetical proteins. In addition, the biotin carboxylase encoded by ORF 22 was transcribed on a separate mRNA transcript from ORFs 16 – 21.

4.5.2 Differential expression of PPI-1 variable region genes *in vivo*

In vivo expression analysis is a valuable tool for highlighting niches in which greater transcription of certain genes is required. Previous work has shown that the preferential expression of many virulence factors within certain niches is largely consistent with their apparent roles in virulence (LeMessurier *et al*., 2006; Mahdi *et al*., 2008). Whilst expression analysis cannot take into account the regulatory roles played by mechanisms independent of mRNA, such as post-translational regulation and changes in the concentration of enzyme substrates, transcriptional patterns provided a basis for the characterisation of the roles of given putative virulence factors in the progression of disease. In Section 4.3.2 it was shown that the expression of *nplT* and ORFs 15, 16, 19 and the major facilitator transporter (ORF 23) was greater in the blood than on the nasopharyngeal surface for both 1861- and 4496-infected mice. Similarly, the expression of *nplT* and ORFs 16 and 19 was greater in the lungs of both 1861- and 4496-infected mice than on the nasopharyngeal surface. Such expression patterns suggest more important roles for these genes for growth and survival in the blood and lungs than on the nasopharyngeal mucosa. In addition, whilst differences in expression of ORF 8 (*pezA*) were detected between niches, if strains 1861 and 4496 possess a second TA system, similar to P1031, the expression of ORF 8 may not be a true reflection of the expression of the *pezAT* locus. Therefore, ORF 10, the expression of which did not consistently vary, was deemed to more accurately reflect the expression of *pezAT*.

Some differences were observed between the two strains, particularly regarding differences in expression between the lungs and blood, where only differences in the expression of the major facilitator were consistent between strains. These discrepancies between strains 1861 and 4496 may have occurred for a number of reasons. It might be possible that these results reflect strain-specific expression requirements. Alternatively, the expression differences between strains might reflect differences that were observed between strains in the number of pneumococci within the blood (Figure 4.2).

In summary, it is clear that the expression of many genes within the PPI-1 variable region, most notably ORFs 6, 16, 19 and 23, exhibited niche-specific expression patterns that were consistent between both strains 1861 and 4496, and strongly suggested that the PPI-1 variable region responds to the changes in the *in vivo* environment encountered by the pneumococcus during disease progression.

4.5.3 Mutagenesis of the PPI-1 variable region in D39 and competitive fitness

Whilst expression analysis supported a role for components of the PPI-1 variable region in the blood and lungs, it was important to confirm these results by investigating the effect that mutating the region would have on virulence. However, as will be discussed in detail in Chapter 6, numerous attempts to genetically manipulate the serotype 1 clinical isolates used in this study were unsuccessful. To circumvent this roadblock, various PPI-1 variable region mutants were constructed in D39. At both 24 h and 48 h it was shown that the PPI-1 variable region found in wild-type D39 was required for wild-type fitness in both the lungs and blood and by 48 h was also required for wild-type fitness in the nasopharyngeal tissue, when compared to the mutant lacking the region altogether (D39ΔPPI-1). Interestingly at both 24 h and 48 h the D39 mutant carrying the strain 1861 version of the region ($D39^{1861}$) exhibited wild-type equivalent fitness in all niches and even a small but significant increase in fitness at the nasopharyngeal surface. The mutant carrying the strain 1 version of the region ($D39^{1}$) was significantly less fit than the wild type in the lungs and blood at 24 h. However, wild-type fitness was reached in lungs when compared to $D39^{1861}$, and the blood when compared to the wild-type at 48 h. Similar to D39ΔPPI-1, $D39^{1}$ exhibited a reduction in fitness in the nasopharyngeal tissue at 48 h compared to the wild-type and $D39^{1861}$.

The ability of the strain 1861 version of the PPI-1 variable region to confer wild-type fitness to D39 supported its role as a region contributing to virulence in strains 1861 and 4496. The intention was to subsequently test the effect of mutating individual components of the PPI-1 variable region in D39^{1861}. However, difficulty in obtaining the desired mutants within the time constraints of this project led to the deferral of such work to a later date. Since Brown *et al*. (2004) found a reduction in fitness *in vivo* from the deletion of only *pezT* it is possible that this gene was solely responsible for the lack of fitness in the blood, lungs and nasopharyngeal tissue in D39ΔPPI-1 and D39^{1}. However, due to the absence of studies comparing the contribution of different versions of the PPI-1 variable region to virulence, it is not possible to rule out a role for genes other than *pezT* on the virulence of strains 1861 and 4496. However, in any case, it was shown that the increased fitness of D39^{1861} compared to D39^{1} was consistent with the invasiveness of strains 1861 and 4496 compared to the lineage A strains.

In summary, the competition experiments performed in Section 4.4.2 strongly suggest a role for the PPI-1 variable region in survival in the blood, lungs and nasopharyngeal tissue, at least in a D39 background. It is unclear whether similar mutations in the serotype 1 isolates themselves would lead to more or less pronounced differences in fitness. However, the findings of the competition experiments supported the findings of Section 4.3.1, which showed greatest expression of *nplT* and ORFs 16 and 19 in the blood and lungs of strain 1861- and 4496-infected mice.

4.5.4 Potential mechanisms of PPI-1 variable region genes of strains 1861 and 4496 in virulence

As discussed above, Sections 4.3.2 and 4.4.2 suggested that *nplT*, *pezAT*, 3HIBDH, ORF 19 and a transporter of the major facilitator superfamily, either collectively or individually, have a role in increased survival of *S. pneumoniae* in the blood, lungs and nasopharyngeal tissue. In addition, PDT and *galE* were also likely to exhibit greater expression in the lungs and the blood than at the surface of the nasopharynx, given their presence on the same transcript as 3HIBDH and ORF 19 (Section 4.2). For the most part, the mechanisms by which the above genes contribute to increased survival in the blood and lungs are not clear. However, there are some interesting links between previously prescribed roles for the products of some genes and the sites of preferential *in vivo* expression identified in this study.

As discussed in Section 3.3.2, *nplT* is an enzyme of the α-amylase superfamily. An interesting parallel is the virulence factor, *amyA*, of Group A streptococci, which has been shown to promote invasive disease by the degradation of host epithelial-associated carbohydrates of the oropharynx (Shelburne *et al*., 2009). This degradation was linked to utilisation of complex host sugars as carbon sources and an increased rate of translocation across the epithelial surface. Therefore, a similar role played by *nplT* could contribute to the greater invasive potential observed for strains 1861 and 4496 relative to the lineage A strains.

In Section 4.4.2, increased survival in the lungs and blood was associated with *pezAT* and a role for *pezAT in vivo* has previously been reported by Brown *et al*. (2004). However, whilst many similar chromosomal TA systems have been identified, the roles of these systems remain controversial (reviewed in Van Melderen & De Bast [2009]). Proposed functions range from promoting stability of non-core regions of the genome to roles in complex regulatory networks, such as responding to environmental stresses. In order to survive environmental stresses TA systems have been postulated to promote dormancy of the bacterium and sometimes cell death in a subpopulation of cells. However, such activity would appear to contradict the observation of rapid progression to IPD that is a feature of *pezAT*-positive strains, such as strains 1861 and 4496 (Sections 1.6.1, 4.3.1, 4.4.2). Therefore, a role for *pezAT* in genomic stability of the PPI-1 variable region might be more plausible in these strains.

The accessory region of the PPI-1 variable region of 1861 contained a number of metabolic enzymes, with unclear roles in virulence. For example, 3HIBDH catalyses the rate-limiting step in the degradative pathway of BCAAs (Robinson & Coon, 1957). Interestingly, BCAAs are among a limited group of substances that can cross the blood-brain barrier and can be used as an energy source. It is interesting that strains 1861 and 4496 should express 3HIBDH highly in the blood, where its substrate (3-hydroxisobutyrate) is present for energy utilisation by neural cells (Murin *et al*., 2008). PDT has been shown to be an important regulatory enzyme in the biosynthesis of phenylalanine (Shiio *et al*., 1976). However, whilst a high ratio of phenylalanine to tyrosine in the blood is characteristic of sepsis (Druml *et al*., 2001), it is unclear whether the expression of PDT in PPI-1 is somehow linked to this phenomenon. *GalE* encoded by ORF 21 is responsible for the reversible conversion of UDP-glucose to UDP-galactose. Interestingly, UDP-sugar substrate availability has been shown to increase capsular polysaccharide chain length, and this could be facilitated by increased expression of *galE* in the blood, providing survival of the pneumococcus (Ventura *et al*., 2006).

As discussed in Section 3.3.2.2, the greatest homology of ORFs 23 and 24 was to an importer of lactose and glycerol-3-phosphate. Greater expression of such a transporter in the blood could indicate increased uptake of a sugar such as lactose, which might be linked to the increased expression of *galE*. However, as only subtle differences in the amino acid sequence of major facilitators can alter substrate specificity and change the direction of transport (Law *et al*., 2009), the precise role of the major facilitator in the virulence of strains 1861 and 4496 is unclear.

4.5.5 Conclusion

Chapter 4 proposed a model of the PPI-1 variable region transcriptome in strains 1 and 1861 and showed that a number of genes including *nplT*, a 3HIBDH, PDT, *galE* and a hypothetical protein exhibit greater expression in both the blood and lungs when compared to expression at the nasopharyngeal surface. In addition, a transporter of the major facilitator superfamily exhibited greater expression in the blood than at the nasopharyngeal surface. Finally, it was shown that a mutant of D39, possessing the PPI-1 variable region of strain 1861 exhibited greater fitness than a mutant of D39 possessing the PPI-1 variable region of strain 1, in the blood, lungs and nasopharyngeal tissue. This increase in fitness could be promoted through mechanisms ranging from degradation of host-derived sugars and amino acids, to increasing the pool of substrates available for capsule biosynthesis. Given the strong association of the PPI-1 variable region of strains 1861 and 4496 with increased fitness in the lungs and blood of mice, and the association of these strains with IPD in humans, further research to determine the mechanisms underlying this increased fitness is warranted.

Chapter 5 – Genomic Differences between Carriage and Invasive Serotype 1 Isolates Identified by Genomic Sequencing

5.1 Introduction

Chapter 4 showed that whilst the PPI-1 variable region of strains 1861 and 4496 contributes to invasive potential when expressed in the D39 background, the region alone may not account for the differences that have been observed in virulence between the non-invasive serotype 1 strains (1 and 2), the intermediately virulent strains (3415 and 5482) and the highly virulent strains (1861 and 4496) (Section 1.6). Therefore, it was decided to complete the genomic comparisons between the six strains in order to identify additional regions that may be associated with a particular virulence phenotype.

As discussed in Section 1.5, previous studies have attempted to attribute the invasive potential of different serotypes and genotypes to ARs within the genome of *S. pneumoniae* (Blomberg *et al*., 2009; Obert *et al*., 2006; Brukner *et al*., 2004). However, such comparisons were made over large groups of strains representing many serotypes and genotypes and made broad assumptions about the invasive potential of certain serotypes. In the case of serotype 1, these studies did not take into account differences in invasive potential between the serotype 1 isolates in both humans and animal models of infection (Section 1.6; Smith-Vaughan *et al*., 2009; Antonio *et al*., 2008; Nunes *et al*., 2008). In addition, the use of CGH for genomic comparisons, whilst an important starting point, cannot detect ARs that include genes not present in the genomes of TIGR4, R6 or G54.

Therefore, in this chapter, CGH was initially undertaken on a selection of non-invasive, intermediately virulent and highly virulent serotype 1 strains to compile a preliminary list of genes that were associated with only the non-invasive strains (1 and 2), with only the invasive strains (3415, 5482, 1861 and 4496) or with only the highly virulent strains (1861 and 4496). Subsequently, it was decided to sequence the genomes of a representative non-invasive strain (strain 1) and a representative highly invasive strain (strain 1861) using next generation genome sequencing technology to perform genomic comparisons that would not be limited to the genetic content of TIGR4 and R6. In addition, regions that were highlighted as variable by sequencing would be confirmed by PCR or direct sequencing in the six strains of all three virulence phenotypes to identify regions associated with a particular virulence phenotype. Once a list of regions that were associated with IPD had been compiled, the expression of a selection of these genes was compared between different niches of the mouse during infection. The analysis of niche-dependent expression might provide some clues as to the role of such genes in pneumococcal pathogenesis.

5.2 Comparison between non-invasive, intermediately virulent and highly virulent serotype 1 isolates by CGH

CGH was used to produce a list of genes associated with the non-invasive strains (strains 1 & 2), invasive strains (3415, 5482, 1861 & 4496) and the highly invasive strains (1861 & 4496), in combinations shown in Table 5.1, and as described in Section 2.11. The detection of hybridisation was used to indicate the presence of a gene within at least one strain on the slide. Each gene on each slide was checked manually using Gene Pix Pro v 6.0, to ensure that detection of hybridisation for a given gene was due to the accumulation of fluorescence at the appropriate position, in duplicate, rather than due to non-specific fluorescence that did not appear to accumulate at the appropriate template site. As mentioned in Section 2.11, the microarray slides used for CGH contained TIGR4 ORFs and additional R6 ORFs, which restricted the detection of genes to those also possessed by either TIGR4 or R6. Lists of genes detected in at least one strain were compiled and compared with the lists from strains of other virulence phenotypes. Genes that were associated with only one or two of the three phenotypes were checked to determine whether one or both strains on the slide possessed the gene. Green or red spots indicated which strain possessed the gene, and yellow spots showed that both strains possessed the gene. Inconclusive detection of genes occurred when fluorescence at a hybridisation spot was only marginally more intense than the surrounding background. Sequence variation between the labelled DNA and the reference template was likely to be the primary cause of such inconclusive results.

Table 5.1 Combinations of strains and dyes used in CGH

Slide	Strain	Alexa Fluor^{TM} Dye
1	1	546 (green)
	2	647 (red)
2	3415	546
	5482	647
3	1861	546
	4496	647

5.2.1 Differences associated with non-invasiveness

As discussed in Section 1.6.1, strains 1 and 2 were isolated from healthy people and are avirulent in mice. Therefore strains 1 and 2 were considered to be non-invasive. *SP_0826*, which encodes a putative ATPase component of an ABC transporter involved in phosphate transport, was the only gene found that was consistently present only in the non-invasive strains and not in either the intermediately virulent or highly virulent strains (Table 5.2). However, the detection of other genes only in the non-invasive strains was inconclusive, possibly due to sequence variation between the genes in strains 1 and 2 when compared to TIGR4. In particular, given the short length of genes *SP_0670, SP_1531* and *SP_1723*, it is possible that these genes are pseudogenes and as a result would be more vulnerable to genetic drift due to a lack of selective pressure. Alternatively, genes such as the putative transposase (*SP_0300*), the PTS system component (*SP_0324*) and the amino acid carrier protein (*SP_0626)* have significant homology to other genes also located within the TIGR4 genome, thus potentially resulting in a false positive result. However, it is unlikely that *SP_0826* would be the only gene present in strains 1 and 2 and absent from the intermediately virulent or highly virulent strains. Rather, other strain 1- and strain 2-specific genes are probably also absent from TIGR4 and R6. Therefore, a method such as genomic sequencing would be required to produce a more complete list of genes associated with only the non-invasive strains.

5.2.2 Differences associated with invasiveness

Genes within a 37-kb region homologous to *SP_1759 – SP_1767* were detected only in a subset of strains capable of causing IPD (Table 5.3). This region is equivalent to AR 34 (Blomberg *et al*., 2009). AR 34 encodes the putative adhesin, PsrP, and the assembly and transport proteins that accompany the protein (*PsrP-secAY2A2*) (Obert *et al*., 2006). Whilst PsrP has been shown to mediate adherence to kertain 10 of lung epithelial cells, the importance of the protein to virulence remains controversial due to conflicting virulence data using PsrP-deficient mutants (Shivshankar *et al*., 2009; Blomberg *et al*., 2009; Orihuela, 2009; Obert *et al*., 2006). Whilst the region appears to be associated with IPD in this study, its absence from the highly virulent strain 1861, suggests that the *PsrP-secAY2A2* region is not required for virulence in this strain.

Table 5.2 Genes associated with non-invasive strains

Gene ID	Annotation	Strain	
		1	2
SP_0300	IS630-Spn1, transposase Orf2	?	?
SP_0324	PTS system, IIC component	?	?
SP_0626	Branched-chain amino acid transport system II carrier protein	?	?
SP_0670	Hypothetical protein	?	?
SP_0826	ATPase component of an ABC-type phosphate transport system	+	+
SP_1531	CsbD-like protein	?	?
SP_1723	Hypothetical protein	?	?

'+' Detected

'?' Detection was inconclusive

Table 5.3 Genes associated with strains capable of causing IPD

Gene ID	Annotation	Strain			
		3415	5482	1861	4496
SP_0629	D-alanyl-D-alanine carboxypeptidase	+	+	+	+
SP_1418	IS1380-Spn1 transposase	+	+	+	+
SP_1503	IS1380-Spn1 transposase	+	+	+	+
SP_1755*	Hypothetical protein	+	+	–	–
SP_1756*	Conserved domain protein	–	–	–	+
SP_1757	Conserved domain protein	?	?	–	+
SP_1758	Glycosyl transferase, group 1	+	+	–	+
SP_1759	Preprotein translocase, SecA subunit	+	+	–	+
SP_1760	Accessory secretory protein Asp3	+	+	–	+
SP_1761	Accessory secretory protein Asp2	+	+	–	+
SP_1762	Accessory secretory protein Asp1	+	+	–	+
SP_1763	Preprotein translocase SecY family protein	+	+	–	+
SP_1764	Glycosyl transferase, family 2	+	+	–	+
SP_1765	Glycosyl transferase, family 8	+	+	–	+
SP_1766	Glycosyl transferase, family 8	+	+	–	+
SP_1767	Glycosyl transferase, family 8	+	+	–	+
SP_1768*	Hypothetical protein	–	–	–	–
SP_1769*	Glycosyl transferase, authentic frameshift	+	+	–	–
SP_1771*	Glycosyl transferase, family 2/glycosyl transferase family	–	–	–	+
SP_1772	PsrP	+	+	–	+
SP_1986	Hypothetical protein	+	+	+	+
SP_2179	IS1380-Spn1 transposase	+	+	+	+

'?' Detection was inconclusive

*Included only for reference, as was not detected in either strain from either the intermediately virulent or highly virulent phenotype

However, it is not clear whether the lack of this region in strains 1 and 2 has contributed to their inability to cause IPD. Figure 5.1 summarises the *Psrp-secAY2A2* region in strains 3415 and 5482 and strain 4496. The majority of genes predicted to be involved in secretion of PsrP are present in all three configurations of the region. The only differences that were detected were in *SP_1755* and *SP_1756*, which are quite small (< 250 bp). Therefore, these ORFs could be pseudogenes, which could be more vulnerable to genetic drift. However, the absence of *SP_1771* and *SP_1768* in strains 3415 and 5482, and of *SP_1768* in strain 4496 might contribute to potential differences in glycosylation of PsrP between strains. Alternatively, such genes might vary in sequence between strains.

In addition, a putative D-alanyl-D-alanine carboxypeptidase (*SP_0629*), three transposases (*SP_1418*, *SP_1503* and *SP_2179*) and a small hypothetical protein (*SP_1986* [315 bp]) were detected in all four strains associated with IPD and not in the non-invasive strains. Some penicillin-binding proteins (PBPs), such as D-alanyl-D-alanine carboxypeptidase (*SP_0629*), have been shown to vary in sequence as a mechanism of resistance against penicillin (reviewed in Zapun *et al*., 2008). Therefore, it is possible that such sequence variation could also lead to reduced hybridisation by labelled DNA of strains 1 and 2 to the TIGR4 template.

In summary, the *PsrP-secAY2A2* region was found to be associated with IPD due to its presence in strains 3415, 5482 and 4496, whilst being absent from the non-invasive strains. However, the absence of this region in strain 1861 indicates that the region is not required by all strains for virulence. In addition, a PBP was found to be associated with only the invasive strains.

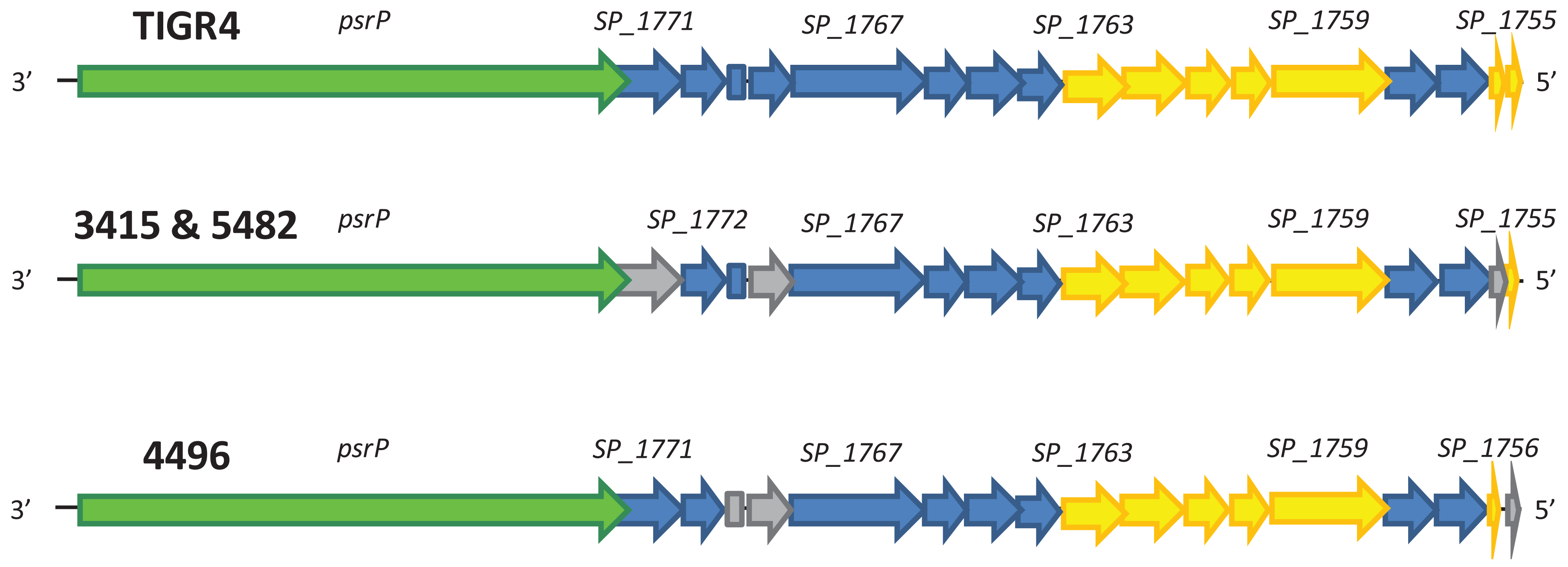


Figure 5.1 Putative organisation of AR 34 in strains 3415, 5482 and 4496, based on that of TIGR4
Diagram of predicted ORFS in TIGR4, strains 3415 and 5482, and strain 4496 derived from the CGH data in Table 5.3. Genes predicted to encode proteins involved in PsrP secretion (yellow), glycosyl transferases (blue) and PsrP (green) are shown above. Genes that were not detected are indicated in grey. Rectangular ORFs represent a truncated gene. TIGR4 ID numbers are used to label a selection of genes.

5.2.3 Differences associated with heightened invasiveness

Three regions of more than one gene were found to be present in only the highly virulent strains and not in either the non-invasive or intermediately virulent strains (Table 5.4). The first region contains genes homologous to the R6 ORFs *SPR_1191 – SPR_1195*, which is part of AR 24 (Blomberg *et al*., 2009; Bruckner *et al*., 2004). The region contains a hypothetical protein (*SPR_1195*) and an ABC transporter, which has been predicted to be involved in the transport of peptides. A comparison between the organisation of the region in R6 and strains 1861 and 4496 is shown in Figure 5.2. Both configurations were similar, with the exception of the absence of *SPR_1192* from the region in strains 1861 and 4496. Perhaps the permease protein encoded by *SPR_1193* is sufficient for the system to function in strains 1861 and 4496. Other ABC transporters have previously been shown to contribute to virulence through a variety of different functions, such as the acquisition of limited nutrients and in antibiotic efflux systems (Section 1.3.4). Therefore, this region cannot be ruled out as having a role in the virulence of strains 1861 and 4496.

The second region associated with the highly virulent strains encodes homologues of *SP_1047 – SP_1053*, and has been designated AR 22 by Blomberg *et al*. (2009). This AR is part of the variable region of PPI-1 (Chapters 3 & 4). In addition, *SPR_0957* and *SPR_0960* are also components of the PPI-1 variable region of strains 1861 and 4496. The detection of PPI-1 variable region components that are absent from the lineage A strains, supports the validity of the use of CGH in genomic comparisons, as it independently confirms the differences in the PPI-1 variable region that were identified in Chapter 3.

The third region associated with the highly virulent strains was a 6.5-kb region encoding genes homologous to *SP_1615 – SP_1621*, which was designated AR 31 by

Table 5.4 Genes associated with heightened virulence

Gene ID	Annotation	Strain	
		1861	4496
SPR_0098	Putative bacteriocin	+	+
SPR_0957	Tn*5252*, relaxase	+	+
SPR_0960	Rgg/GadR/MutR family transcriptional regulator	+	+
SPR_1191	ABC transporter, ATP-binding component – oligopeptide transport	+	+
SPR_1192*	ABC transporter, membrane-spanning permease – oligopeptide transport	–	–
SPR_1193	ABC transporter, membrane-spanning permease – oligopeptide transport	+	+
SPR_1194	ABC transporter, substrate-binding component-oligopeptide transport	+	+
SPR_1195	Hypothetical protein	+	+
SP_0115	Hypothetical protein	+	+
SP_0136	Glycosyl transferase, family 2	+	+
SP_0506	Integrase/recombinase, phage integrase family	+	+
SP_0575	Putative helicase	+	–
SP_1019	Acetyltransferase, GNAT family	+	+
SP_1047	Hypothetical protein	+	+
SP_1048	Hypothetical protein	+	+
SP_1049	Hypothetical protein	+	+
SP_1050	*PezA* – chromosomal toxin-antitoxin system	+	+
SP_1051	*PezT* – chromosomal toxin-antitoxin system	+	+
SP_1052	Putative phosphoesterase	+	+
SP_1053	Conserved domain protein	+	+
SP_1183	Hypothetical protein	+	+
SP_1336	Type II DNA modification methyltransferase	+	+
SP_1615	Transketolase	+	–
SP_1616	Ribulose-phosphate-3-epimerase family protein	+	–
SP_1617	PTS system, IIC component	+	–
SP_1618	PTS system, IIB component	+	–
SP_1619	PTS system, IIA component	+	–
SP_1620	PTS system, nitrogen regulatory, component IIA	+	–
SP_1621	Transcription antiterminator BglG family protein, authentic frameshift	+	–
SP_2046	Putative methyltransferase	+	+
SP_2093	Hypothetical protein	+	–

*Included only for reference, as was not detected in either strain 1861 or 4496

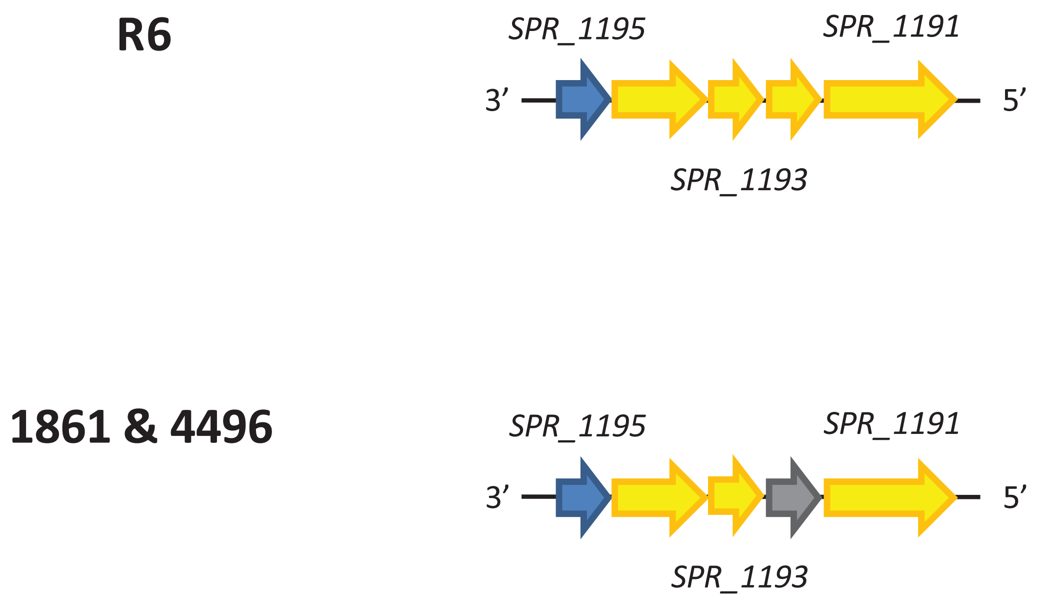


Figure 5.2 Putative organisation of the R6 region *SPR_1191 – SPR_1195*, part of AR 24, in strains 1861 and 4496
Diagram of predicted ORFs in R6 and strains 1861 and 4496 are based on CGH data in Table 5.4. Genes predicted to encode an ABC transporter (yellow) and a hypothetical protein (blue) are shown above. Genes that were not detected are indicated in grey. R6 ID numbers are indicated for a selection of genes. The region has previously been designated AR 24 (Blomberg *et al*., 2009), which includes genes *SPR_1184 – SPR_1198*.

Blomberg *et al*. (2009) (Figure 5.3). The region consists predominantly of a phosphotransferase system (PTS) thought to be involved in the uptake of ribulose, but did not allow *in vitro* growth using ribulose as the sole carbon source (Embry *et al*., 2007). However, due to the complex nature of sugar utilisation by the pneumococcus (Section 1.3.1.3), it is not yet possible to assume that this PTS is not important for survival *in vivo*. Furthermore, important considerations include confirming experimentally that the substrate of this PTS is in fact ribulose and subsequently understanding the role of this substrate in the process of CCR, which was described in Section 1.3.1.3. However, the system is unlikely to be required for virulence in all strains, due to its absence from the genome of strain 4496. The broad array of systems identified within different strains of *S. pneumoniae* highlights the possibility of substantial redundancy between genes required for carbohydrate utilisation (Tettelin *et al*., 2001; Hoskins *et al*., 2001).

In addition, a putative bacteriocin (*SPR_0098*) was only detected in strains 1861 and 4496, and has been reported to be part of AR 2 (Blomberg *et al*, 2009). However, whilst the remainder of AR 2 was absent from strains 3415 and 5482, the region was mostly present in strains 1 and 2, which lacked only *SPR_0098*. It is unlikely that this region would be responsible for the differences that were observed in virulence as the infection models that were used did not directly compare the competitive fitness of the serotype 1 isolates within the same mouse. However, it is possible the region does play a role in intra- or interspecies competition, which could be addressed using a competitive model of infection.

Other non-contiguous genes associated with the highly invasive strains include those encoding a number of transferases, such as acetyltransferases, methyltransferases and glycosyltransferases, which could be involved in the modification of DNA leading

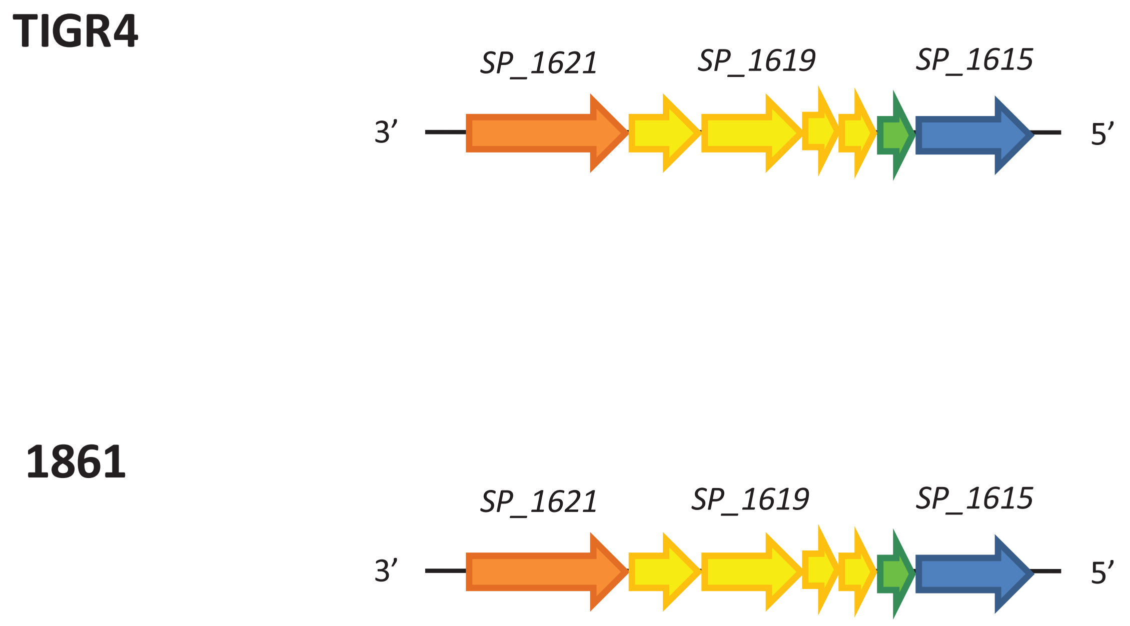


Figure 5.3 Predicted organisation of AR 31 in strains 1861 and 4496, based on that of TIGR4
Diagram of predicted ORFs in TIGR4 and strain 1861 derived from the data shown in Table 5.4. Genes predicted to encode BlgG (orange), the PTS (yellow), ribulose-phosphate-3-epimerase (green) and the transketolase (blue) are shown above. TIGR4 ID numbers are used to label selected genes.

to changes in gene expression, or the modification of surface proteins that could have an impact on immune recognition or ligand specificity. Integrase (*SP_ 0506*) and helicase (*SP_0575*) genes are commonly associated with mobile genetic elements, such as phage and conjugative transposons.

In summary, two regions of more than one gene were found to be present in both highly virulent strains, whilst being absent from the non-invasive and intermediately virulent strains, which included an ABC transporter, predicted to be involved in the transport of peptides and the PPI-1 variable region.

5.3 Genomic sequencing of strains 1 and 1861 using the Illumina® Genome Analyzer *II* System

A number of regions were found to be associated with IPD and heightened invasiveness using CGH (Section 5.2). However, the detection of genes by CGH is limited to those present in the reference genomes, TIGR4 and R6. Therefore, in order to comprehensively compare the genomes of a highly virulent isolate and a non-invasive isolate, the genomes of strains 1 and 1861 were sequenced using the Illumina® Genome analyser *II* system. Chromosomal DNA of colonies deemed to be of the same opacity phase (Section 2.2) was extracted from strains 1 and 1861, as described in Section 2.9.3. Sequencing was carried out by Geneworks in 36-bp sequence reads, which generated 2,278,683 reads from strain 1 and 2,670,211 reads from strain 1861. Sequenced reads were assembled against a reference genome. Recently, the genome sequence of the serotype 1 strain, P1031 (ST303) became available (Genbank accession number CP000920; ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Streptococcus_pneumoniae_P1031/). As P1031 (ST303) is of the same serotype 1 lineage as strain 1861 (Section 3.2), the strain’s genome was predicted to be very similar to strain 1861 and provided a good basis for comparisons with strain 1. Individual sequence reads generated from each strain were assembled against the P1031 genome sequence with a minimum of 80% sequence identity (Section 2.9.9). The assembly of the strain 1 reads generated a 1,947,650-bp consensus sequence with an average read depth of 35.72 reads that aligned against 92.22% of the P1031 genome. The assembly of the strain 1861 reads generated a 2,105,218-bp consensus sequence with an average read depth of 43.66 reads that aligned against 99.68% of the P1031 genome. SNPs between the assembled sequences and the reference sequence were confirmed in the consensus where the variant nucleotide was present in at least 80% of reads at the same position. Unassembled sequences were placed in the ‘boneyard’, which was independently assembled into contigs by *de novo* assembly. Assembly of the boneyard generated 130 contigs of at least 300-bp from strain 1 unassembled sequences and 62 contigs of at least 300-bp from strain 1861 unassembled sequences. Lists of the BLAST results of the boneyard contigs are presented in Tables A.1 and A.2 (Appendix).

Alignments were performed between the genome sequence of P1031 and the consensus sequence generated for strain 1 (Figure 5.4) and strain 1861 (Figure 5.5), as described in Section 2.6.1. The alignments show that the reduced homology shared by strains P1031 and strain 1 included a number of large gaps where strain 1 sequences had not assembled against the P1031 genome. By contrast, the smaller amount of variation between strain 1861 and P1031 was not visible in the whole genome alignment. The primary aim of Section 5.4 was to catalogue the discrepancies that exist between the strain 1 consensus sequence and P1031 and the strain 1861 consensus sequence and P1031.

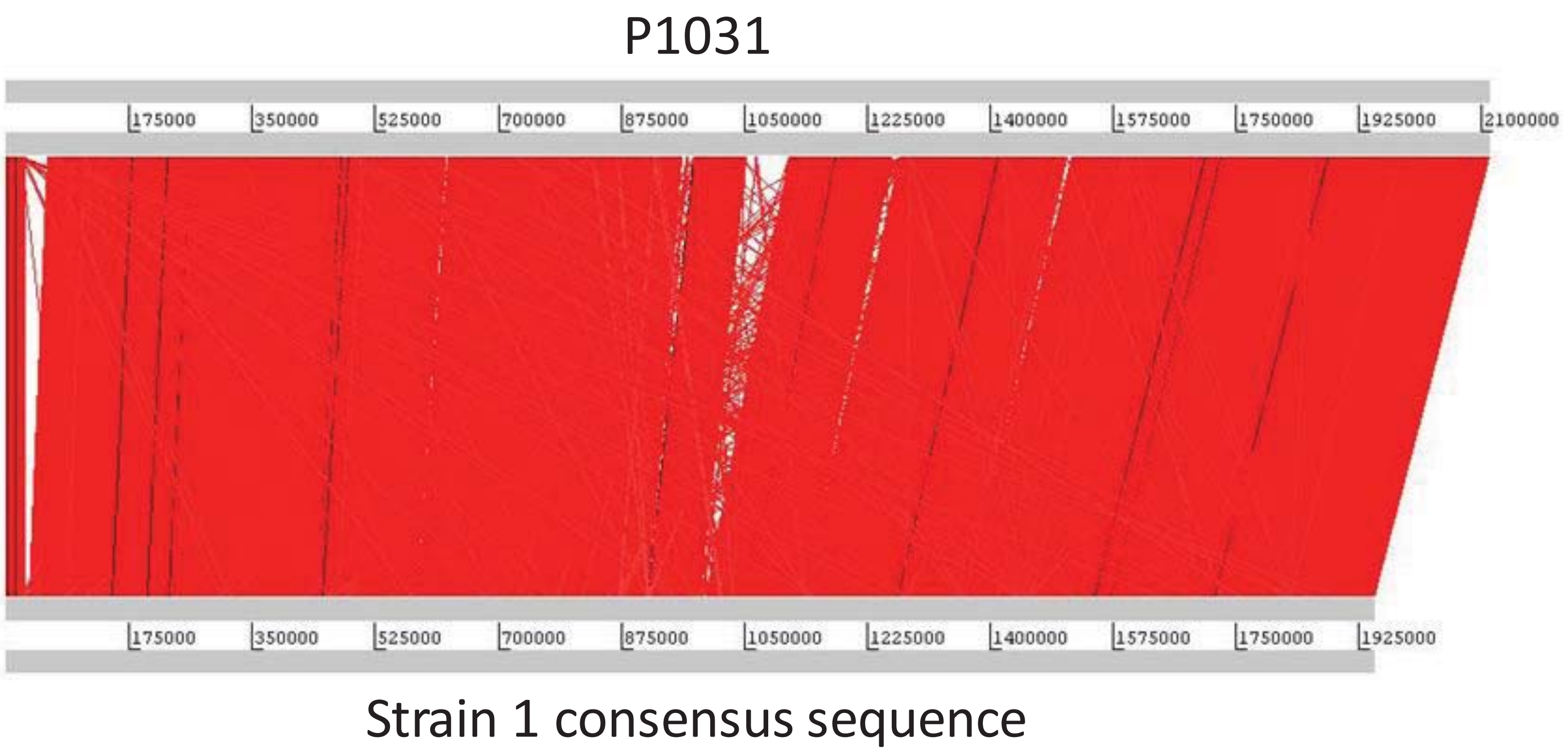


Figure 5.4 Alignment between the strain 1 consensus sequence and the P1031 reference sequence
Alignment between the P1031 genome sequence and the strain 1 consensus sequence was generated using the ACT as described in Section 2.6.1. Sequence sharing at least 80% identity over at least 200 bp between the two strains is indicated in red.

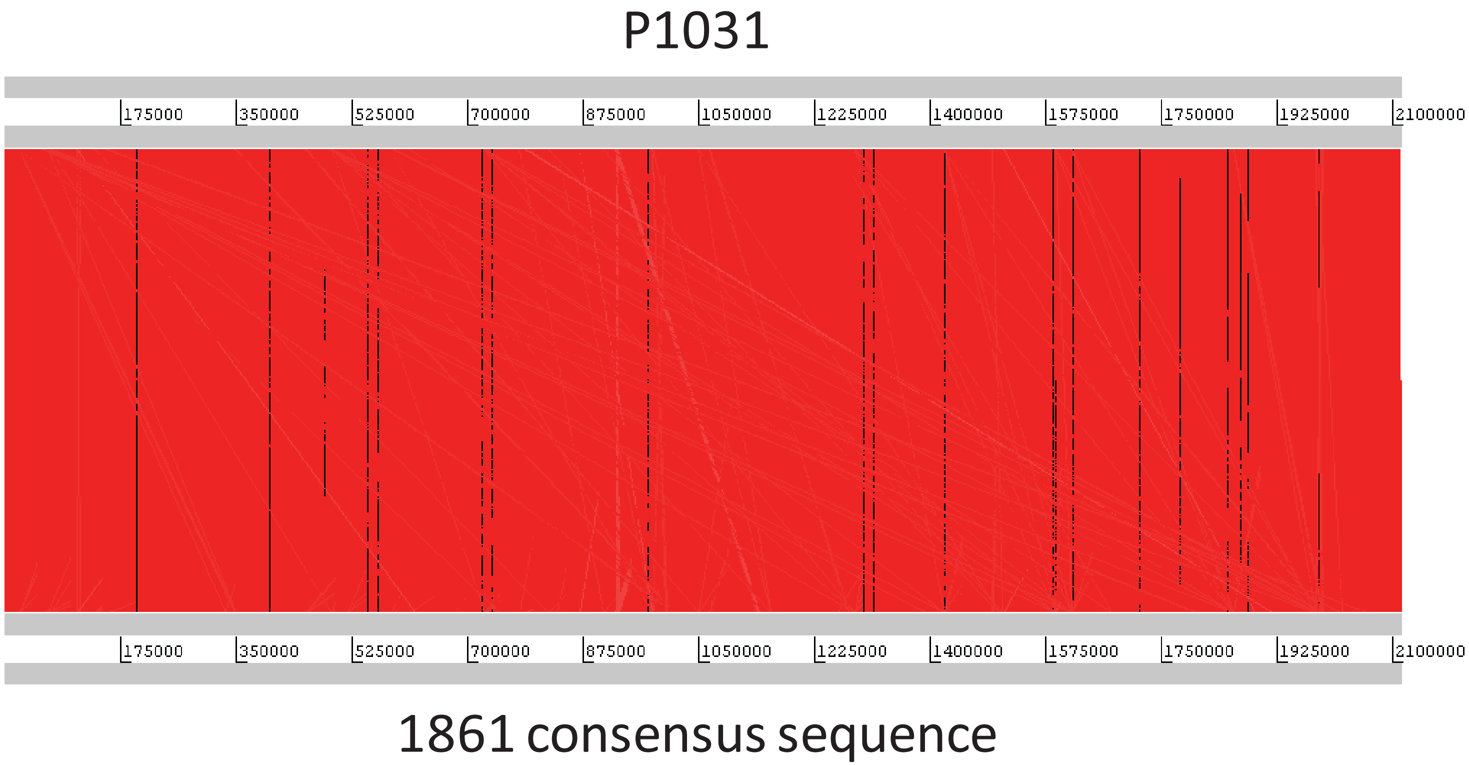


Figure 5.5 Alignment between the strain 1 consensus sequence and the P1031 reference sequence

The alignment between the P1031 genome sequence and the strain 1861 consensus sequence was generated using the ACT as described in Section 2.6.1. Sequence sharing at least 80% identity over at least 200 bp between the two strains is indicated in red.

5.4 Genetic differences identified between strains 1 and 1861 by genomic sequencing

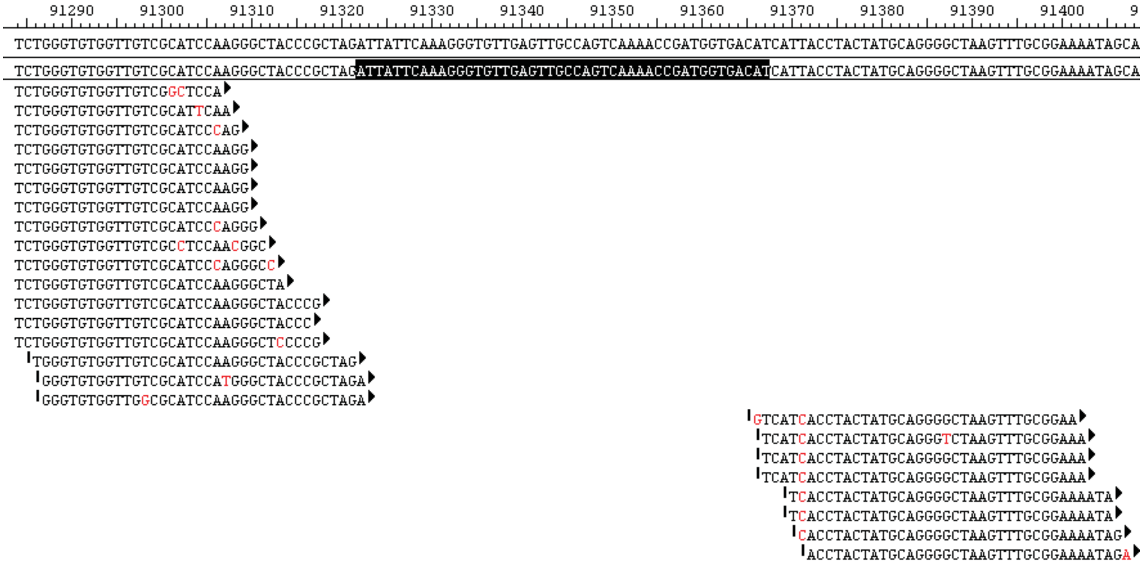
From the assembled sequences generated from both strains 1 and 1861 in Section 5.3, all discrepancies between the P1031 reference sequence and the assembled sequence, excluding SNPs, were recorded in sequence files of at least 150-bp in length. Discrepancies varied from large gaps between assembled sequences, to gaps of <150 bp. Regions of poor assembly caused by localised assembly errors were also included in the lists of discrepancies and are presumed to have occurred due to sequence variation between the reference and the sequenced genome at that location (Figure 5.6). In addition, series of small consecutive deletions were recorded as a single discrepancy. Translation overviews were generated for sequencing gaps of greater than 1 kb, using DNAMAN (Section 2.6.1) and predicted ORFs were subjected to BLASTp searches of the NCBI and KEGG databases. In addition to cataloguing the putative ORF descriptions from BLAST searches, the distribution of the identified genes amongst the genomes of TIGR4 (serotype 4), D39 (serotype 2), ATCC 700669 (serotype 23F), JJA (serotype 14), 70585 (serotype 5), Taiwan 19F (serotype 19F), Hungary 19A (serotype 19A), G54 (serotype 19F) and CGSP14 (serotype 14) (from the KEGG database) was also determined.

Gaps and other discrepancies of less than 1 kb were subjected to BLASTn and BLASTx searches of the NCBI database. 545 discrepancies were recorded within the coding sequence of strain 1, and 88 discrepancies were recorded within the coding sequence of strain 1861. The entire lists of these discrepancies are shown in Tables A.3 and A.4 (Appendix). Colour coding for the ORF descriptions in tables of BLAST search results are described in Table 5.5. In order to provide a focus for the present study, only assembly gaps greater than 1-kb in size and discrepancies within genes predicted to be

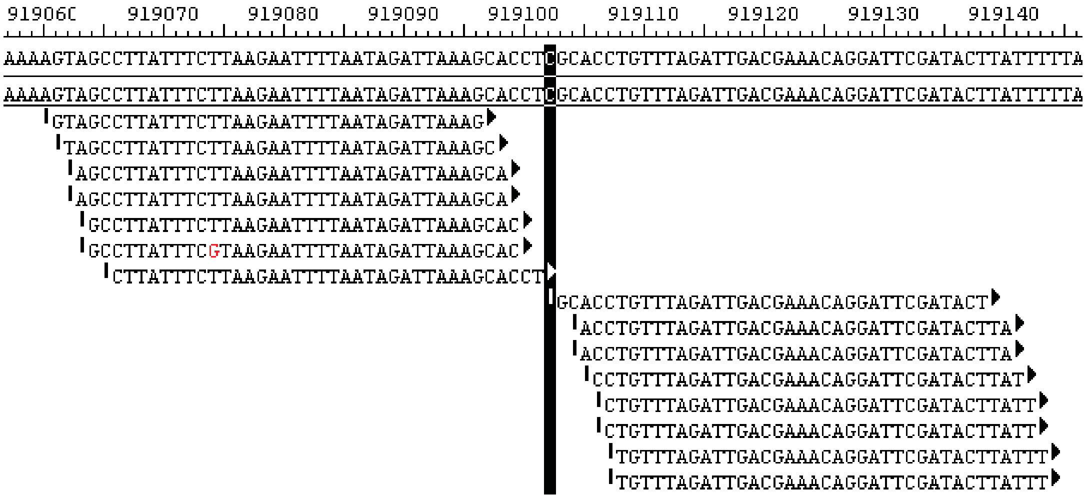
Figure 5.6 Categorisation of discrepancies

Discrepancies between the P1031 genome and the assembled sequences of strain 1 and 1861 were identified in the assembly view of SeqMan (Section 2.6.1). Files of at least 150-bp of sequence were generated for each discrepancy. Throughout this section discrepancies between the assembled consensus sequence were categorised as deletions or poor assemblies. Examples of this categorisation are shown opposite and overleaf. P1031 sequences within the assembly gap are highlighted in black.

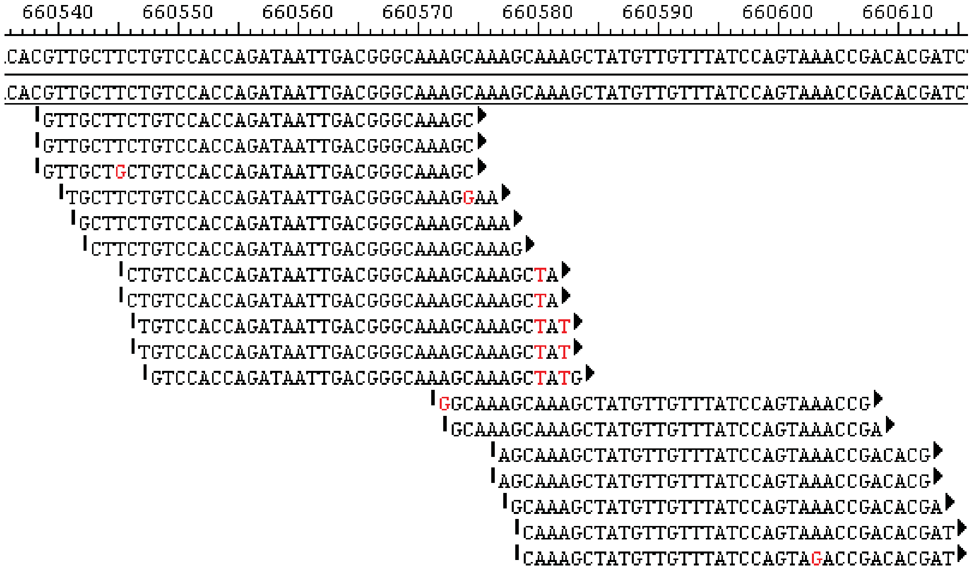
41-bp assembly gap


91290
91300
91310
91320
91330
91340
91350
91360
91370
91380
91390
91400
9
TCTGGGTGTGGTTGTCGCATCCAAGGGCTACCCGCTAGATTATTCAAAGGGTGTTGAGTTGCCAGTCAAAACCGATGGTGACATCATTACCTACTATGCAGGGGCTAAGTTTGCGGAAAATAGCA

1-bp assembly gap


919060
919070
919080
919090
919100
919110
919120
919130
919140
AAAAGTAGCCTTATTTCTTAAGAATTTTAATAGATTAAAGCACCTCGCACCTGTTTAGATTGACGAAACAGGATTCGATACTTATTTTTA

Poor assembly



```
 660540    660550    660560    660570    660580    660590    660600    660610
.CACGTTGCTTCTGTCCACCAGATAATTGACGGGCAAAGCAAAGCAAAGCTATGTTGTTTATCCAGTAAACCGACACGATC
.CACGTTGCTTCTGTCCACCAGATAATTGACGGGCAAAGCAAAGCAAAGCTATGTTGTTTATCCAGTAAACCGACACGATC
    GTTGCTTCTGTCCACCAGATAATTGACGGGCAAAGC
    GTTGCTTCTGTCCACCAGATAATTGACGGGCAAAGC
    GTTGCTGCTGTCCACCAGATAATTGACGGGCAAAGC
      TGCTTCTGTCCACCAGATAATTGACGGGCAAAGGAA
       GCTTCTGTCCACCAGATAATTGACGGGCAAAGCAAA
        CTTCTGTCCACCAGATAATTGACGGGCAAAGCAAAG
           CTGTCCACCAGATAATTGACGGGCAAAGCAAAGCTA
           CTGTCCACCAGATAATTGACGGGCAAAGCAAAGCTA
            TGTCCACCAGATAATTGACGGGCAAAGCAAAGCTAT
            TGTCCACCAGATAATTGACGGGCAAAGCAAAGCTAT
             GTCCACCAGATAATTGACGGGCAAAGCAAAGCTATG
                                     GGCAAAGCAAAGCTATGTTGTTTATCCAGTAAACCG
                                      GCAAAGCAAAGCTATGTTGTTTATCCAGTAAACCGA
                                          AGCAAAGCTATGTTGTTTATCCAGTAAACCGACACG
                                          AGCAAAGCTATGTTGTTTATCCAGTAAACCGACACG
                                           GCAAAGCTATGTTGTTTATCCAGTAAACCGACACGA
                                            CAAAGCTATGTTGTTTATCCAGTAAACCGACACGAT
                                            CAAAGCTATGTTGTTTATCCAGTAGACCGACACGAT
```

involved in virulence will be considered in detail. PCR and localised sequencing will be used to confirm such discrepancies between the strain 1 consensus sequence and P1031 (Section 5.5).

Table 5.5 Colour coding for ORF descriptions of BLAST results

Colour key for gene description	
Virulence factors	🟧
Transporters	🟨
Metabolism-associated	🟩
Regulation of transcription	🟨
Phage-associated	🟥
Transposase/transposon-associated	🟦
Hypothetical proteins	⬜
Other types of genes	🟪

5.4.1 Discrepancies between the P1031 sequence and the assembled strain 1861 consensus sequence

As described in Section 5.3, the strain 1861 consensus sequence represented approximately 99.68% of the P1031 genome, whereas the strain 1 consensus sequence represented 92.22% of the P1031 genome. Therefore, it was decided to catalogue the discrepancies between the strain 1861 consensus sequence and the P1031 genome first in order to ascertain whether the discrepancies identified later for strain 1 also represented differences between the strain 1 and strain 1861 genomes. The complete list of discrepancies identified between the strain 1861 consensus sequence and the P1031 genome are catalogued in Table A.4. Table 5.6 shows the list of 8 discrepancies that are greater than 100-bp in size. For the most part, the list of genes predicted to be interrupted or lost were within genes that would be expected to be most prone to sequence variation, such as phage-, transposon- and transposase-associated genes. However, the gene encoding the virulence factor NanA (Section 1.3.1), was predicted to contain two deletions that are 415-bp and 525-bp in size in strain 1861 (Table 5.6).

Table 5.6 P1031 ORFs missing or interrupted in strain 1861 assembly gaps greater than 100-bp in size

	Annotation/Putative function	ORF ID from annotated *S. pneumoniae* strains										Gap length
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
1	1,4-beta-N-acetylmuramidase	0667	0579	0688	06030	0618	0726	0691	0762	0608	0623	180bp
2	Tn*5253* CAAX amino terminal protease family	1346	1180	1364	12350	1248	1386	-	1482	1287	0847	115bp
3	Hypothetical	-	-	1177	12830	-	-	-	1248	1274	1307	625bp
4	*UmuD/MucA* homolog, transcriptional regulator (Tn*5253*) *UmuC/mucB* homolog (Tn*5253*)	- -	- -	1190 1191	- -	- -	- -	- -	- -	- -	- -	2,078bp
5	Tn*5253* conserved hypothetical protein	1348	1182	1147	12370	1270	1388	-	1235	1289	-	126bp
6	Neuraminidase A	1326	1504	1709	16920	1586	1731	1630	1797	1600	1665	415bp
7	Neuraminidase A	1326	1504	1709	16920	1586	1731	1630	1797	1600	1665	525bp
8	Hypothetical	2093	-	2148	-	-	-	-	-	-	-	440bp

Colour key for gene descriptions are described in Table 5.5

Dark shaded ORF IDs indicate annotation as pseudogenes

However, the mosaic nature of *nanA* has been previously reported (King *et al*., 2005), and in the case of strain 1861 has not led to an inability to cause disease. The contribution of such sequence variation to the assembly gaps of *nanA* was supported by the presence of *nanA* sequence in contig 671 of the strain 1861 boneyard (Table A.2). Direct sequencing of the gene in strain 1861 would be required to precisely define the differences in *nanA* sequence between strain 1861 and P1031. In addition, a putative 1,4-β-N-acetylmuraminidase (*SP_0667*), which shares some homology with the autolysin, LytC, appeared to be disrupted in strain 1861 (Table 5.6). However, the homologous gene in P1031 was annotated as a pseudogene due to a frameshift. Therefore, it is not clear whether the gene is functional in either strain 1861 or P1031. Localised sequencing would be required to confirm whether a deletion exists at this site or whether the gene in strain 1861 exhibits sequence variability compared to P1031.

5.4.2 Discrepancies between the P1031 sequence and the assembled strain 1 consensus sequence

5.4.2.1 Strain 1 sequencing gaps greater than 1-kb in size

As described above, P1031 sequence corresponding to assembly gaps in the strain 1 consensus sequence of greater than 1-kb in size were used to generate translation overviews, which are presented as each region is investigated. Subsequently, BLASTp searches of ORFs predicted within the translation overviews were used to predict the function of individual genes. In addition, genes of unknown function were submitted to the HHpred search engine in order to attribute a function to such genes using a more sensitive search methodology (Section 2.6.2). Of the 16 assembly gaps that were greater than 1-kb in size, two were predicted to encode transposases and so were not examined further, leaving the remaining list of 14 regions shown in Table 5.7.

Table 5.7 P1031 ORFs missing or interrupted in strain 1 assembly gaps greater than 1-kb in size

	Annotation/Putative function	ORF ID from annotated *S. pneumoniae* strains										Gap Length
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC 700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary-19A (SPH)	G54 (SPG)	CGSp14 (SPCG)	
1	Integrase	-	-	0028	-	-	0028	-	0026	-	-	33,844bp
	Hypothetical	-	-	0029	-	-	0029	-	0028	-	-	
	Plasmid addiction system poisin protein	-	-	0032	00290	0034	-	-	-	-	-	
	Conserved hypothetical	-	-	0033	00300	0035	-	-	-	-	-	
	Phage transcriptional regulator, Cro/CI family	-	-	0034	00310	0036	0031	-	-	-	-	
	Phage protein	-	-	0038	-	-	-	-	-	-	-	
	gp15	-	-	0039	-	-	0035	-	0034	-	-	
	Phage protein	-	-	0040	-	-	0036	-	0035	-	-	
	Hypothetical	-	-	0041	-	-	0037	-	0036	-	-	
	gp19	-	-	0043	-	-	0039	-	0038	-	-	
	gp21 (DNA replication protein)	-	-	0045	-	-	0041	-	0040	-	-	
	gp24	-	-	0049	-	-	0045	-	0091	-	-	
	Hypothetical, phage related	-	-	0050	-	-	0046	-	0092	-	-	
	Hypothetical	-	-	0053	-	-	0050	-	-	-	-	
	gp29	-	-	0054	-	-	0051	-	-	-	-	
	Prophage lSa2 site-specific recombinase	-	-	0056	-	-	0054	-	0046	-	-	
	Phage-related hypothetical	-	-	0058	-	-	-	-	-	-	-	
	Phage terminase, small subunit	-	-	0059	-	-	-	-	-	-	-	
	Phage terminase, large subunit	-	-	0060	-	-	-	-	-	-	-	
	Phage portal protein	-	-	0062	-	-	-	-	-	-	-	
	Phage prohead protease	-	-	0063	-	-	-	-	-	-	-	
	Phage major capsid protein	-	-	0064	-	-	-	-	-	-	-	
	Prophage pi2 protein 39	-	-	0070	-	-	-	-	-	-	-	
	Prophage LambdaSa04, tail tape measure protein	-	-	0072	-	-	-	-	0060	-	-	
	Hypothetical	-	-	0073	-	-	-	-	0061	-	-	
	Host specificity protein	-	-	0074	-	1850	0074	-	0062	-	-	
	PblB-like protein	-	-	0075	-	1850	0074	-	0062	-	-	
	Hypothetical	-	-	0076	-	1850	0074	-	0062	-	-	
	Phage holin 1	-	-	0081	-	-	0077	-	0065	-	-	
	Prophage lSa2, holin LLH superfamily	-	-	0082	-	-	0078	-	0066	-	-	

	Annotation/Putative function	ORF ID from annotated *S. pneumoniae* strains										Gap Length
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC 700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary-19A (SPH)	G54 (SPG)	CGSp14 (SPCG)	
2	Nitroreductase family protein	0574	0500	0590	05190	0532	-	-	0673	-	0538	2,656bp
	DEAD/DEAH Box helicase	0575	0500	0590	05190	0532	-	-	0673	-	0541	
3	ZmpB	0664	0577	0684	05990	1074	0723	0688	0759	0605	0620	166bp
	ZmpB	0664	0577	0684	05990	1074	0723	0688	0759	0605	0620	2,322bp
4	Sodium dependent transporter	0737	0642	0747	06610	0676	0783	0753	0838	0669	0687	Poor assembly
	Sodium dependent transporter	0737	0642	0747	06610	0676	0783	0753	0838	0669	0687	1,896bp
	MerR Family transcriptional regulator (regulator of pmrA)	0739	0643	0750	06630	0678	0785	0754	-	0670	0689	
	MuT/Nudix family protein	0740	0644	0751	06640	0679	0786	0755	-	0671	0690	
5	Neopullulanase	1046	0927	1049	09660	0983	1125	1101	1147	0972	1026	6,317bp
	Hypothetical	1047	0928	1051	09670	0984	1126	1104	1148	-	1027	
	PezA (antitoxin)	1050	0930	1053	09700	0987	1128	-	-	-	1029	
	PezT (toxin)	1051	0931	1054	09710	0988	1129	-	-	-	1030	
	Hypothetical	1052	0932	1055	09720	0989	1130	-	-	-	1031	
	Tn5252 ORF 10 protein	1054	0934	1056	09740	0991	-	-	-	-	1033	
	Tn5252 relaxase	1056	0936	1058	09780	0993	1135	-	-	0973	1035	
6	*Tn5252* , relaxase	-	0938	1059	09780	0995	1135	-	-	0975	-	9,505bp
	Rgg/GadR/MutR family transcriptional regulator	-	0939	1060	09790	0996	-	-	-	0976	-	
	Rgg/GadR/MutR family transcriptional regulator	-	0939	1060	09790	0996	-	-	-	0976	-	
	3-hydroxyisobutyrate dehydrogenase	-	-	1062	-	-	-	-	-	0977	-	
	Hypothetical	-	-	1063	-	-	-	-	-	0978	-	
	Prephenate dehydratase	-	-	1064	-	-	-	-	-	0979	-	
	Hypothetical	-	-	1065	-	-	-	-	-	0980	-	
	Hypothetical	-	-	1066	-	-	-	-	-	0981	-	
	UDP-glucose-4-epimerase	-	-	1067	-	-	-	-	-	0982	-	
	Biotin carboxylase	-	-	1068	-	-	-	-	-	0983	-	
	Transporter, Major facilitator superfamily	-	0950	1069	-	-	-	-	-	0984	-	
	Transporter, Major facilitator superfamily	-	0950	1069	-	-	-	-	-	0985	-	

	Annotation/Putative function	ORF ID from annotated *S. pneumoniae* strains										Gap Length
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC 700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary-19A (SPH)	G54 (SPG)	CGSp14 (SPCG)	
7	IgA1 protease	1154	1018	1140	10580	1073	1207	1181	1229	1053	1143	63,754bp
	ZmpD	-	-	1141	10590	1074	-	-	-	1054	1142	
	Replication initiator protein A, N-terminus (RepA)	-	-	1144	13160	-	-	-	1232	1292	1340	
	Type II DNA modification methyltransferase	1336	-	1145	-	-	1071	-	1233	1291	1339	
	Tn*5253* , conserved hypothetical	1349	1183	1146	13140	1251	1072	-	1234	1290	1338	
	Caax amino protease	1346	1180	1149	13110	1248	1075	-	1237	1287	1336	
	Tn*916* , Transposase	-	-	1150	-	-	-	1931	1402	-	1334	
	RNA polymerase σ-70 region 4 family protein	-	-	1152	13070	-	-	1929	1406	-	1331	
	TetM	-	-	1155	13050	-	-	1919	1409	1231	0171	
	Tn*916* , Hypothetical	-	-	1157	13040	-	-	1917	1411	1232	0169	
	NLP/P60 family	-	-	1158	13030	-	-	1916	1412	1233	0168	
	Conjugative transposon membrane protein	-	-	1159	13020	-	-	1915	1413	1234	0167	
	Conjugative transposon protein	-	-	1160	13010	-	-	1914	1414	1235	0166	
	Tn*5251* , Cro/CI family transcriptional regulator	-	-	1164	12970	-	-	1910	1422	1242	0162	
	Tn*916* , FtsK/SpoIIIE family	-	-	1165	12960	-	-	1909	1423	1243	0161	
	TraG/TraD family protein	-	-	1171	12890	-	1077	-	1242	1280	1313	
	Tn*5252* , Orf23	-	-	1173	12870	-	1079	-	1244	1278	1311	
	Type IV sec pathway, VirB4 component	-	-	1175	12850	-	-	-	1246	1276	1309	
	M23 peptidase	-	-	1176	12840	-	1083	-	1247	1275	1308	
	Hypothetical	-	-	1177	12830	-	-	-	1248	1274	1307	
	SNF2 family protein	-	-	1179	-	-	-	-	-	1272	1305	
	Parvulin-like peptidyl-prolyl isomerase	-	-	1183	12680	-	-	-	1257	1267	1301	
	DNA primase	-	-	1184	12670	-	-	-	1258	1266	1300	
	Helix-turn-helix domain protein	-	-	1188	12630	-	1128	-	1262	-	1296	
	Signal particle GTPase	-	-	1189	12620	-	1129	-	1263	-	1295	
	UmuD/MucA homolog, transcriptional regulator	-	-	1190	-	-	-	-	-	-	-	
	UmuC/mucB homolog	-	-	1191	-	-	-	-	-	-	-	
	Tn*5252* , relaxase	-	-	1196	12430	-	-	-	-	1250	1276	
	Integrase	-	-	1198	-	-	-	-	-	-	-	

	Annotation/Putative function	ORF ID from annotated *S. pneumoniae* strains										Gap Length
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC 700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary-19A (SPH)	G54 (SPG)	CGSp14 (SPCG)	
8	High-affinity Fe2+/Pb2+ permease	1300	1155	1340	-	-	-	-	1442	1194	1267	1,696bp
	Dyp-type peroxidase	-	1156	1342	-	-	-	-	-	-	1268	1,696bp
	Transposase	+	+	1343	+	+	+	+	+	+	+	1,696bp
	IS1380, transposase	+	-						+	+	+	1,702bp
	IS1380-Spn1, Transposase	+	-						+	+	+	2,225bp
9	N-acetylneuraminate lyase	1329	-	1350	12210	1245	1374	1615	1471	1222	1648	18,026bp
	Cytidine deaminase (EC 3.5.4.5)	-	1164	1351	-	-	1371	-	-	-	-	
	Phosphatidylglycerophatase A (EC 3.1.3.27)	-	1165	1352	-	-	1376	-	-	-	-	
	Glycoside hydrolase family protein (EC 3.2.1.26)	-	1166	1353	-	-	1377	-	0827	-	-	
	ABC transporter, ATP-binding protein (oligopeptide transport)	1888	1167	1354	-	-	1378	-	-	-	-	
	Oligopeptide transport, membrane spanning permease	1889	1168	1355	-	-	-	-	-	-	-	
	Oligopeptide transport, membrane spanning permease	1890	1169	1356	-	-	-	-	-	-	-	
	ABC transporter, substrate binding		1170	1357	-	-	-	-	-	-	-	
	Kelch-like protein	-	1171	1358	-	-	-	-	-	-	-	
	N-acetylmannosamine-6-phosphate 2-epimerase (EC 5.1.3.9)	-	1172	1359	-	-	-	-	-	-	-	
	Tn*5253* bacteriocin	-	1174	1361	12910	-	-	-	-	1282	1315	
	Hypothetical	-	-	1362	12920	-	-	-	1239	-	1316	
	Tn*5253* , hypothetical	-	-	1363	12930	-	-	-	1238	1285	1317	
	CAAX amino protease	1346	1180	1364	12350	1248	1386	-	1237	1287	1336	
	Arsenate reductase	1348	1182	1366	12370	1250	1388	-	1235	1289	-	
	Hypothetical	1349	1183	1367	12380	1251	1389	-	1234	1290	1338	
	Methyltransferase	-	-	1368	-	-	-	-	1233	1291	1339	
	Replication initiation protein A, N-terminus	-	-	1371	12390	-	1390	-	1232	1292	1340	
10	Hypothetical	1612	-	1633	-	-	1652	1553	-	-	1594	1,545bp

	Annotation/Putative function	ORF ID from annotated *S. pneumoniae* strains										Gap Length
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC 700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary-19A (SPH)	G54 (SPG)	CGSp14 (SPCG)	
11	Transketolase (cytidylate kinase)	1615	-	1637	-	-	1657	1555	1729	-	-	7,287bp
	D-allulose-6-phosphate 3-epimerase	1616	-	1638	-	-	1658	1556	1730	-	-	
	PTS system IIC component (ribulose)	1617	-	1639	-	-	1659	1557	1731	-	-	
	PTS system, IIB component	1618	-	1640	-	-	1660	1558	1732	-	-	
	PTS system, IIA component	1619	-	1641	-	-	1661	1559	1733	-	-	
	PTS system, nitrogen regulatory component IIA	1620	-	1642	-	-	1662	1560	1734	-	-	
	Transcription antiterminator, BglG family	1621	-	1643	-	-	1663	1561	1735	-	-	
12	ABC-2 type transporter	-	-	1779	-	-	-	-	-	-	-	2,876bp
	Nod factor export ATP-binding protein I	-	-	1780	-	-	-	-	-	-	-	
	Transcriptional regulator, ArsR family	-	-	1781	-	-	-	-	-	-	-	
13	Ribosomal protein methyltransferase	1781	1572	1783	17940	spj1680	1827	1704	1898	1669	1759	multiple over 1,514bp
	Ribosomal protein L11 methyltransferase (prmA)	1782	1573	1784	17950	spj1681	1828	1705	1899	-	1760	
	MutT/nudix family (7,8-dihydro-8-oxoguanine-tripjoatase)	1783	1574	1785	-	-	1829	-	1900	1672	-	
14	Hypothetical	2093	-	2148	-	-	-	-	-	-	-	1,456bp

Colour key for gene descriptions are described in Table 5.5

Coloured ORF IDs indicate annotation as pseudogenes (Dark) or are positioned at an alternative location within the genome (green)

Spotted rows were not considered any further

Genes predicted to be encoded within these assembly gaps were deemed to be absent from strain 1 and if not also absent from strain 1861 (Section 5.4.1), were deemed to be a genetic difference between strains 1 and 1861.

Region 1 is a 33.8-kb assembly gap (Figure 5.7), predicted to contain ORFs homologous to *SPP_0028 – SPP_0082* of P1031 and was not detected by CGH (Section 5.2) due to the region's absence from both the TIGR4 and R6 genomes. The region was predicted to encode bacteriophage proteins, with a similar genetic structure to that of group 1 temperate pneumophage (Romero *et al*., 2009a). Of particular interest was the homologue of the platelet-binding protein B (PblB) of *Streptococcus mitis* (*SPP_0075*), which has been shown to mediate binding to human platelets and is required for virulence in an animal model of infective endocarditis (Bensing *et al*., 2001; Mitchell *et al*., 2007). Whilst the genes of region 1 are not present in the TIGR4 or R6 genomes, components of the region are present in the genomes of Hungary 19A, JJA and 70585 (Table 5.7).

Region 2 is approximately 2.7-kb and contains two ORFs (Figure 5.8). The two ORFs are annotated as a single pseudogene (*SPP_0590*) in the P1031 genome. BLASTx and BLASTp searches suggested that *SPP_0590* is a fragmented DEAD/DEAH box helicase family protein, homologous to *SPJ_0532* of JJA.

Region 3 includes two consecutive assembly gaps that were 166-bp and 2.3-kb in size. The whole region was homologous to sequence encoding the zinc metalloproteinase B (ZmpB) of P1031 (*SPP_0684*). Therefore, it is possible that a truncated ZmpB exists in strain 1, which may contribute to reduced virulence. Alternatively, allelic differences in *zmpB* (Hsieh *et al*., 2008) might contribute to assembly gaps in variable regions of the gene. The contribution of sequence variation to the assembly gaps of *zmpB* was supported by the presence of *zmpB* sequence in contigs

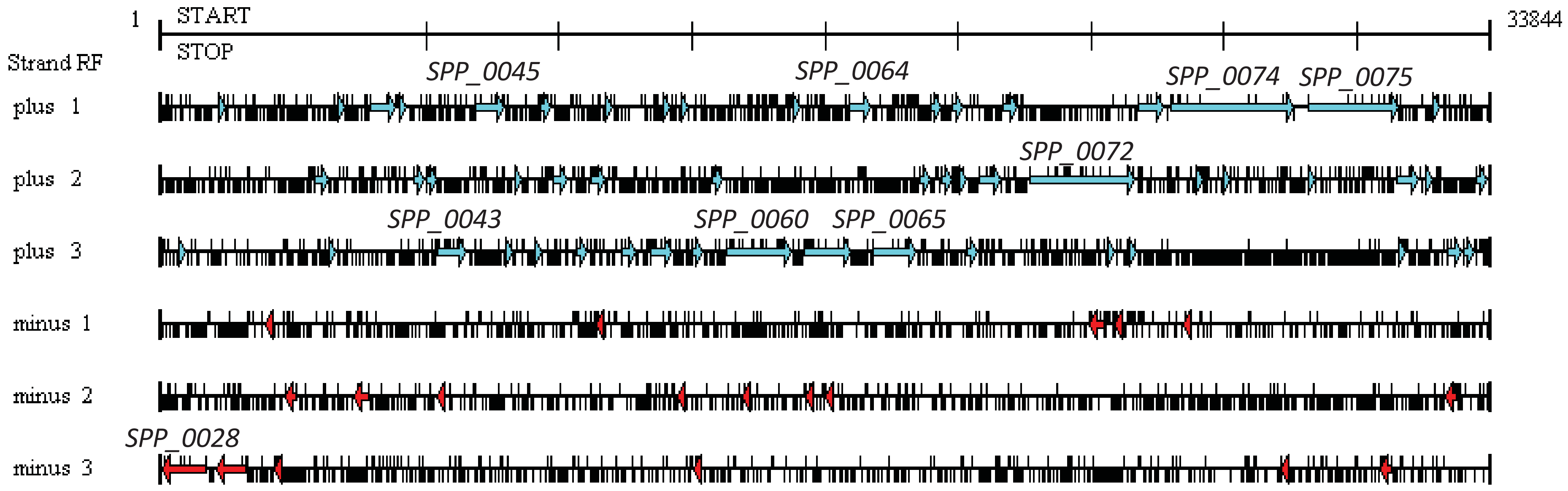


Figure 5.7 Translation overview of the assembly gap sequence of region 1
Translation overview in all six reading frames was generated using DNAMAN (Section 2.6.1), from P1031 sequence of the assembly gap of region 1 (Table 5.7). ORFs containing the start codon 'ATG' and of at least 80 amino acids in length are highlighted in blue (forward strand) and red (reverse strand). Vertical black bars indicate start codons (above) and stop codons (below).

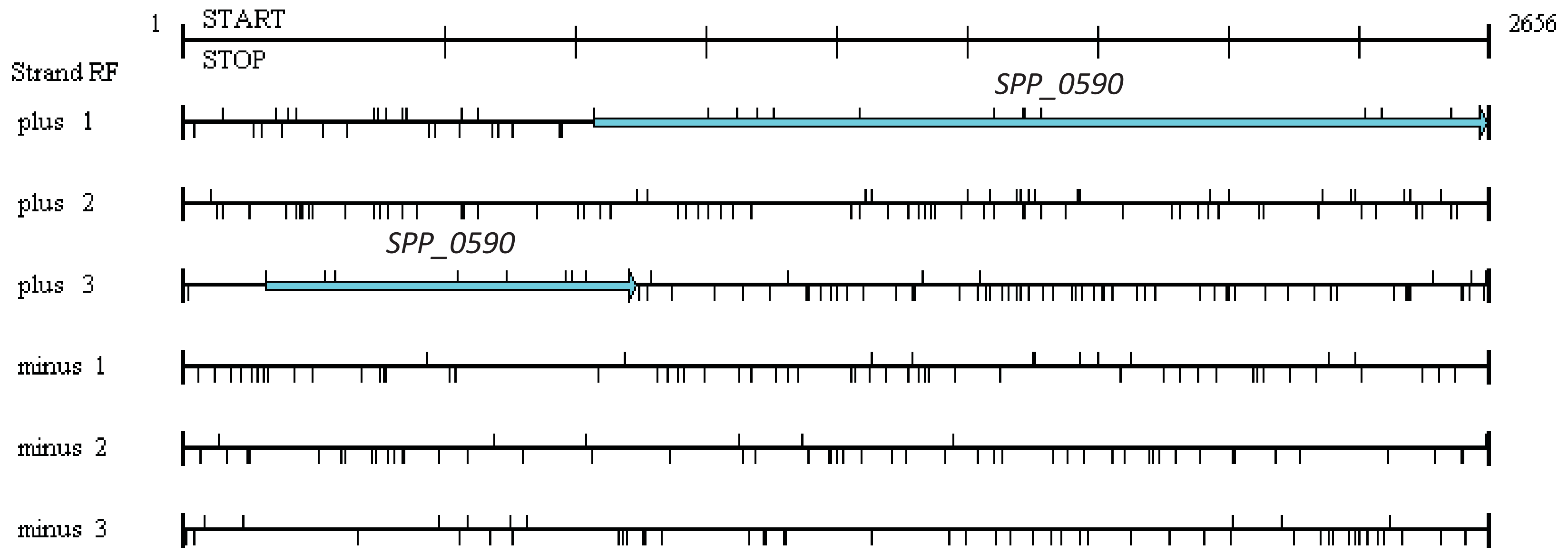


Figure 5.8 Translation overview of the assembly gap sequence of region 2
Translation overview in all six reading frames was generated using DNAMAN (Section 2.6.1), from P1031 sequence of the assembly gap of region 2 (Table 5.7). ORFs containing the start codon 'ATG' and of at least 80 amino acids in length are highlighted in blue (forward strand). Vertical black bars indicate start codons (above) and stop codons (below).

2028 and 1215 of the strain 1 boneyard (Table A.1). Of interest has been the suggestion that allelic variation of *zmpB* has an impact on pneumococcal virulence (Hsieh *et al*., 2008). Therefore, the allelic variation responsible for the discrepancies within region 3 could indicate that strains 1 and 1861 possess different alleles of *zmpB* that might contribute to their differences in virulence.

Region 4 includes two discrepancies within close proximity, which included poor assembly followed by a 1.9-kb assembly gap. Three ORFs were predicted within region 4 (Figure 5.9), which included a MerR-family transcriptional regulator (*SPP_0750*) on the forward strand and part of a fragmented sodium-dependent transporter (*SPP_0750*) and a MutT/Nudix family protein (*SPP_0751*) on the reverse strand. The fragmented sodium-dependent transporter was homologous to the full-length gene in D39 (*SPD_0642*), but is annotated as a pseudogene in P1031 (*SPP_0747*) due to an early frameshift. A similar frameshift was observed in ATCC 700669 (*SPN23F_06610*). Therefore, whilst much of the sodium-dependent transporter appeared to be absent from strain 1, the gene might not be functional in P1031. However, uninterrupted versions of a putative MerR transcriptional regulator and a MutT/Nudix family protein appeared to be missing from strain 1.

Regions 5 and 6 suggested that strain 1 lacks most of the PPI-1 variable region present in P1031, which has been shown to be the case in Chapter 3, and was also shown by CGH in Section 5.2. Homologous sequence shared by strain 1 and P1031 exists within the Tn*5252*-associated region of PPI-1 (Section 3.4.1) and corresponds to the sequence between the assembly gaps of regions 5 and 6.

Region 7 is the largest assembly gap in the strain 1 consensus sequence at approximately 63.8-kb in size (Figure 5.10), and includes ORFs homologous to *SPP_1140* – *SPP_1198* of P1031. Of particular interest was the presence of the IgA1

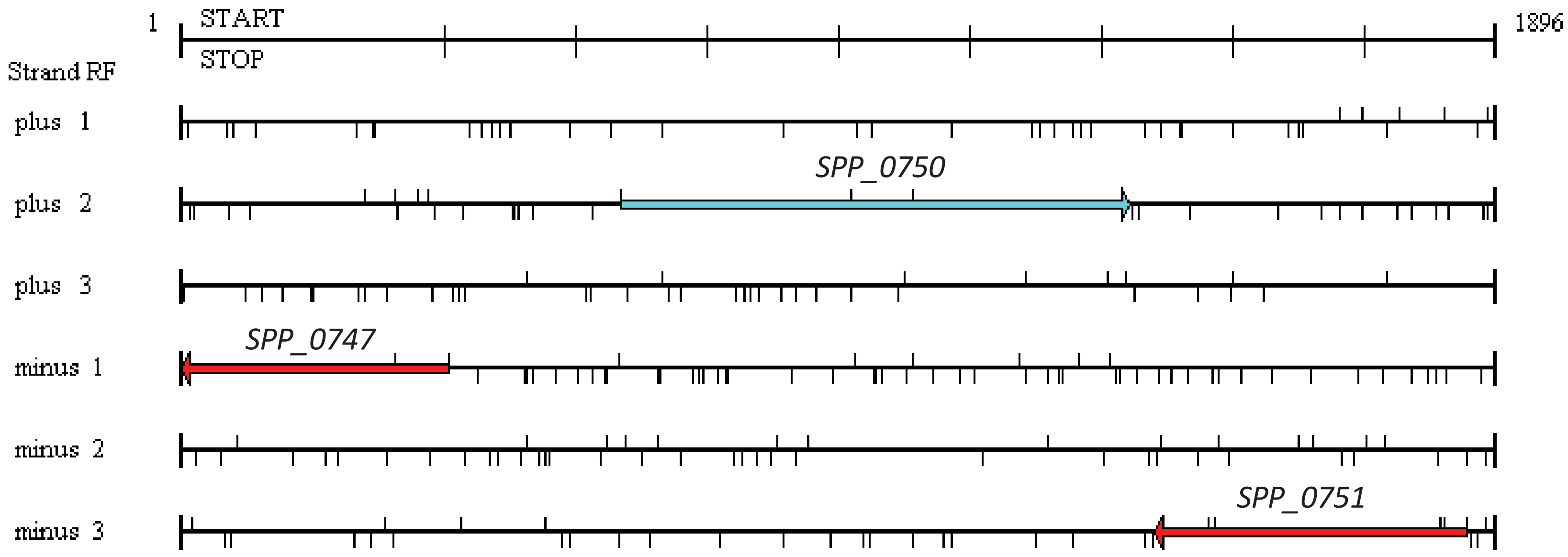


Figure 5.9 Translation overview of the assembly gap sequence of region 4
Translation overview in all six reading frames was generated using DNAMAN (Section 2.6.1), from P1031 sequence of the assembly gap of region 4 (Table 5.7). ORFs containing the start codon 'ATG' and of at least 80 amino acids in length are highlighted in blue (forward strand) and red (reverse strand). Vertical black bars indicate start codons (above) and stop codons (below).

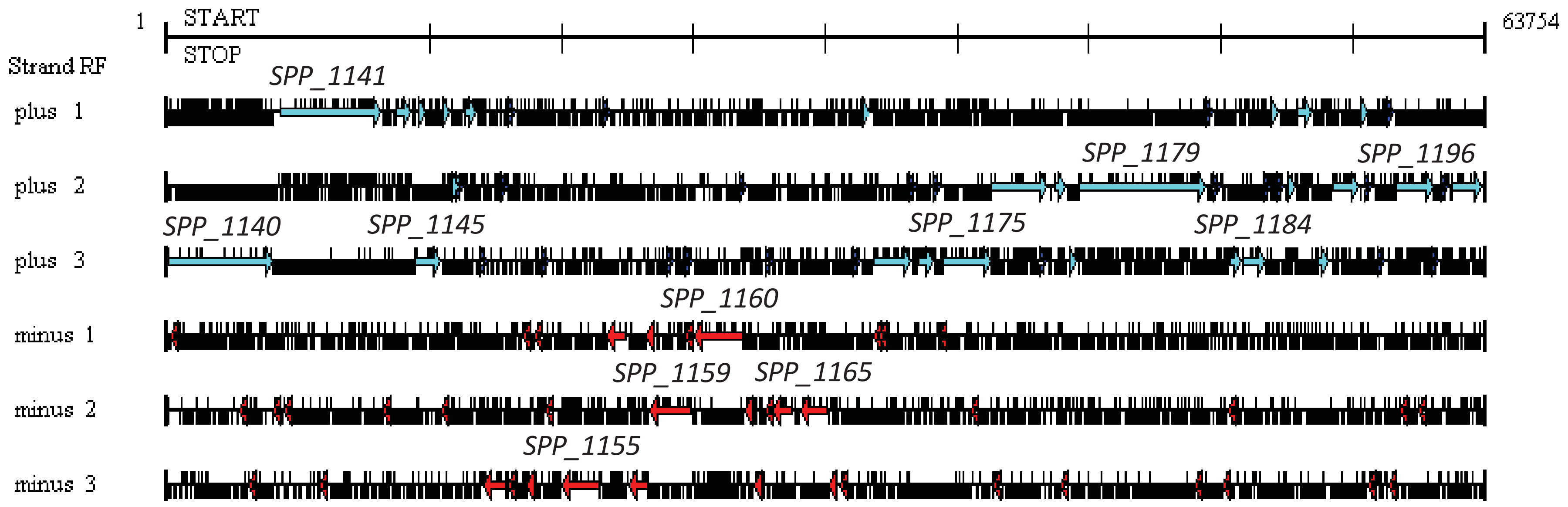


Figure 5.10 Translation overview of the assembly gap sequence of region 7
Translation overview in all six reading frames was generated using DNAMAN (Section 2.6.1), from P1031 sequence of the assembly gap of region 7 (Table 5.7). ORFs containing the start codon 'ATG' and of at least 80 amino acids in length are highlighted in blue (forward strand) and red (reverse strand). Vertical black bars indicate start codons (above) and stop codons (below).

protease gene (*iga*; *SPP_1140*), which plays important roles in survival of the pneumococcus *in* vivo (Section 1.3.1). *Iga* was followed by *zmpD* (*SPP_1141*), which has been suggested to be possessed by only 49% of strains (Chiavolini *et al*., 2003), and of the strains in the KEGG database only ATCC 700669, CGSP14, G54 and JJA possess the gene (Table 5.7). In P1031, *SPP_1141* has been annotated as *zmpB*, probably as a result of an error during automated annotation due to homology between *zmpB* and *zmpD*. The remaining 52 kb of region 7 encodes most of the conjugative transposon, Tn*5253*. Full-length Tn*5253* is 65.5-kb in size and is a composite element made up of the Tn*916*-like element, Tn*5251*, inserted into a Tn*5252* element (Ayoubi *et al*., 1991; Henderson-Begg *et al*., 2009). The composite nature of Tn*5253* explains the mixture of hits from Tn*916* and Tn*5252* for region 7 (Table 5.7). However, it seems that P1031 has lost part of the Tn*5252* portion, as the overall Tn*5253* region was 13-kb smaller than previously reported. Tn*5253* appears to exhibit limited distribution throughout the genomes of the KEGG database, as ATCC 700669 is the only other genome that was found to contain the majority of the region. G54, Hungary 19A and CGSP14 appeared to possess separate elements as their respective ORF IDs indicated that the Tn*5251* and Tn*5252* elements are located at different positions within the chromosome. Taiwan 19F appears to possess only Tn*5251*, whilst 70585 possesses parts of Tn*5252*. It is important to note that sequence homologous to *SPP_1190* and *SPP_1191* of Tn*5253* was identified as absent in strain 1861 (Section 5.4.1). In addition, a putative TA system (*SPP_1188* – *SPP_1189*) homologous to PezAT (Chapter 4) is also present within this region.

Region 8 is a 1.7-kb assembly gap in strain 1 (Figure 5.11) that contains ORFs homologous to *SPP_1340* – *SPP_1343* of P1031. These genes include a putative high-affinity Fe^{2+}/Pb^{2+} permease (*SPP_1340*), a dyp-type peroxidase (*SPD_1156*) and a

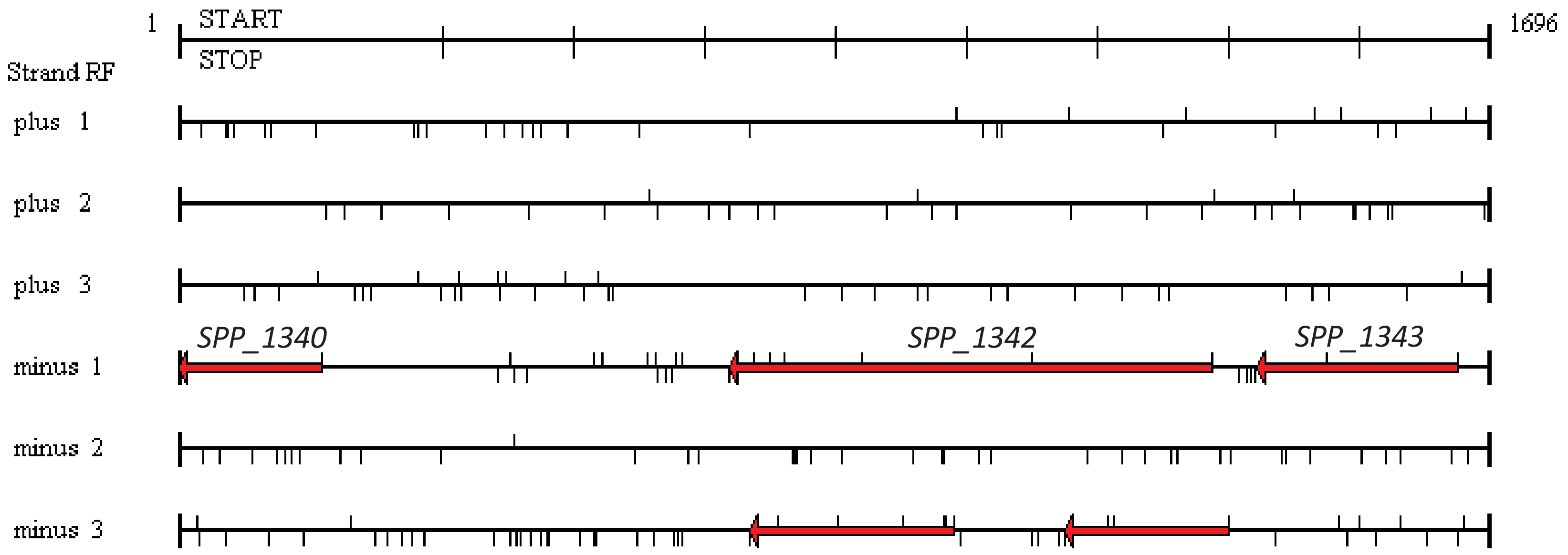


Figure 5.11 Translation overview of the assembly gap sequence of region 8
Translation overview in all six reading frames was generated using DNAMAN (Section 2.6.1), from P1031 sequence of the assembly gap of region 8 (Table 5.7). ORFs containing the start codon 'ATG' and of at least 80 amino acids in length are highlighted in red (reverse strand). Vertical black bars indicate start codons (above) and stop codons (below).

transposase (*SPP_1343*). The putative Fe^{2+}/Pb^{2+} permease was only partially present in the assembly gap indicating that at least part of the sequence encoding the gene was present in strain 1. In P1031 the dyp-type permease is annotated as a pseudogene (*SPP_1342*) due to fragmentation of the gene when aligned with *SPD_1156* of D39. This region exhibits mixed distribution with the Fe^{2+}/Pb^{2+} permease present in strains TIGR4 (*SP_1300*), D39 (*SPD_1155*), Hungary 19A (*SPH_1442*), G54 (*SPG_1194*) and CGSP14 (*SPCG_1267*) (Table 5.7). In contrast, the dyp-type permease was only present in strains D39 (*SPD_1156*) and CGSP14 (*SPCG_1268*) (Table 5.7).

Region 9 is an 18-kb assembly gap (Figure 5.12) that contains ORFs homologous to *SPP_1350 – SPP_1371* of P1031. These genes also share homology with ORFs *SPD_1164 – SPD_1174*, which correspond to part of AR 24 (Blomberg *et al*., 2009) and was also detected as a region of difference by CGH in Section 5.2.3. CGH suggested that the region was associated with heightened virulence, due to its presence in strains 1861 and 4496, but absence from the non-invasive and intermediately virulent strains. The region encodes putative enzymes involved in sialic acid degradation (*SPP_1350*, *SPP_1358* and *SPP_1359*), and a putative ABC transporter (*SPP_1354 – SPP_1357*). Region 9 displayed only limited distribution within the pneumococcal strains of the KEGG database, with only D39 possessing a majority of homologous ORFs.

Region 10 is a 1.5-kb assembly gap that encodes a putative serine/threonine protein kinase homologous to *SPP_1633* of P1031. Region 10 displayed mixed representation being present only in strains TIGR4, 70585, Taiwan 19F and CGSP14 (Table 5.7).

Region 11 was a 7.3-kb assembly gap (Figure 5.13) that was predicted to encode ORFs homologous to *SPP_1637 – SPP_1643* in P1031 and *SP_1615 – SP_1621* in

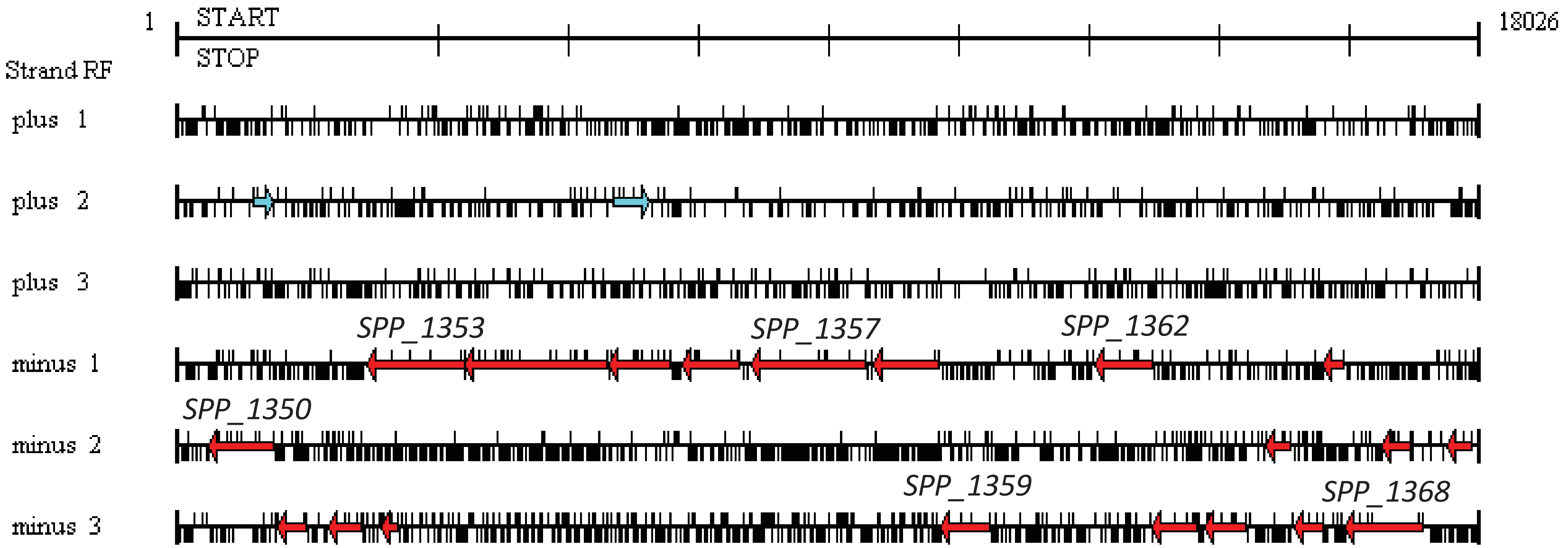


Figure 5.12 Translation overview of the assembly gap sequence of region 9
Translation overview in all six reading frames was generated using DNAMAN (Section 2.6.1), from P1031 sequence of the assembly gap of region 9 (Table 5.7). ORFs containing the start codon 'ATG' and of at least 80 amino acids in length are highlighted in blue (forward strand) and red (reverse strand). Vertical black bars indicate start codons (above) and stop codons (below).

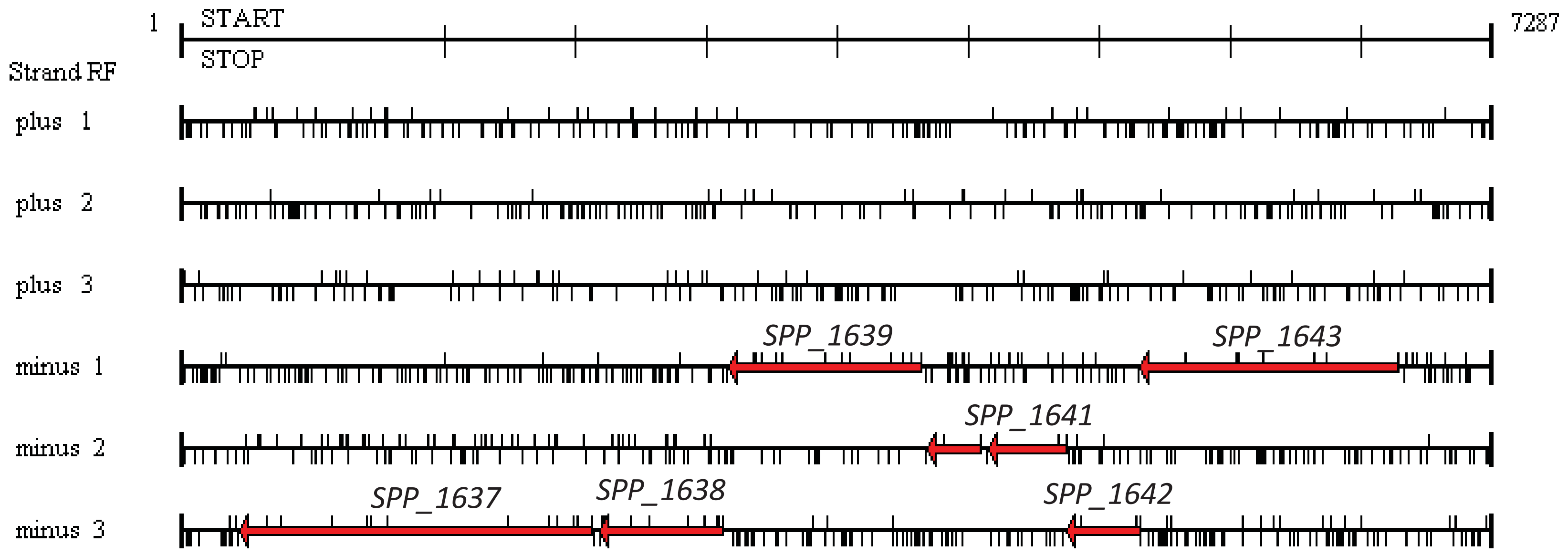


Figure 5.13 Translation overview of the assembly gap sequence of region 11
Translation overview in all six reading frames was generated using DNAMAN (Section 2.6.1), from P1031 sequence of the assembly gap of region 11 (Table 5.7). ORFs containing the start codon 'ATG' and of at least 80 amino acids in length are highlighted in red (reverse strand). Vertical black bars indicate start codons (above) and stop codons (below).

TIGR4, which corresponds to AR 34 (Blomberg *et al*., 2009). In Section 5.2.3, this region was present in strain 1861, but absent from the non-invasive strains, the intermediately virulent strains and strain 4496. In addition, the region is present in strains 70585, Taiwan 19F and Hungary 19A, suggesting mixed representation amongst pneumococcal strains. The region encodes a PTS thought to be involved in ribulose acquisition (Obert *et al*., 2006) and was previously discussed in Section 5.2.3.

Region 12 is a 2.9-kb region (Figure 5.14) that contains ORFs homologous to *SPP_1779 – SPP_1781* of P1031. However, these genes were not found in any of the other pneumococcal genomes of the KEGG database (Table 5.7). Within the region, the genes encoding the ABC-2 type transporter and Nod factor export ATP-binding protein may be co-transcribed, given their common orientation and overlap. The ArsR family transcriptional regulator (*SPP_1781*) may provide some regulatory control over the other two genes.

Region 13 was a 1.5-kb region (Figure 5.15) that contains ORFs homologous to *SPP_1784 – SPP_1785* of P1031. However, whilst these genes were also homologous to *SP_1781 – SP_1783* of TIGR4, the CGH of Section 5.2 did not identify an association between these genes and a virulence phenotype, as they were not present in strain 4496. The genes of region 13 appeared to be widely distributed amongst the KEGG *S. pneumoniae* genomes (Table 5.7).

5.4.2.2 Deletions in strain 1 within virulence factors

Table 5.8 summarises the assembly gaps identified within suspected virulence factors. A number of short assembly gaps and a poor assembly were identified within *pspA* (*SPP_0185*). Given that *pspA* is known to vary in sequence (Hollingshead *et al*., 2000), it was not surprising that *pspA* would contain such discrepancies.

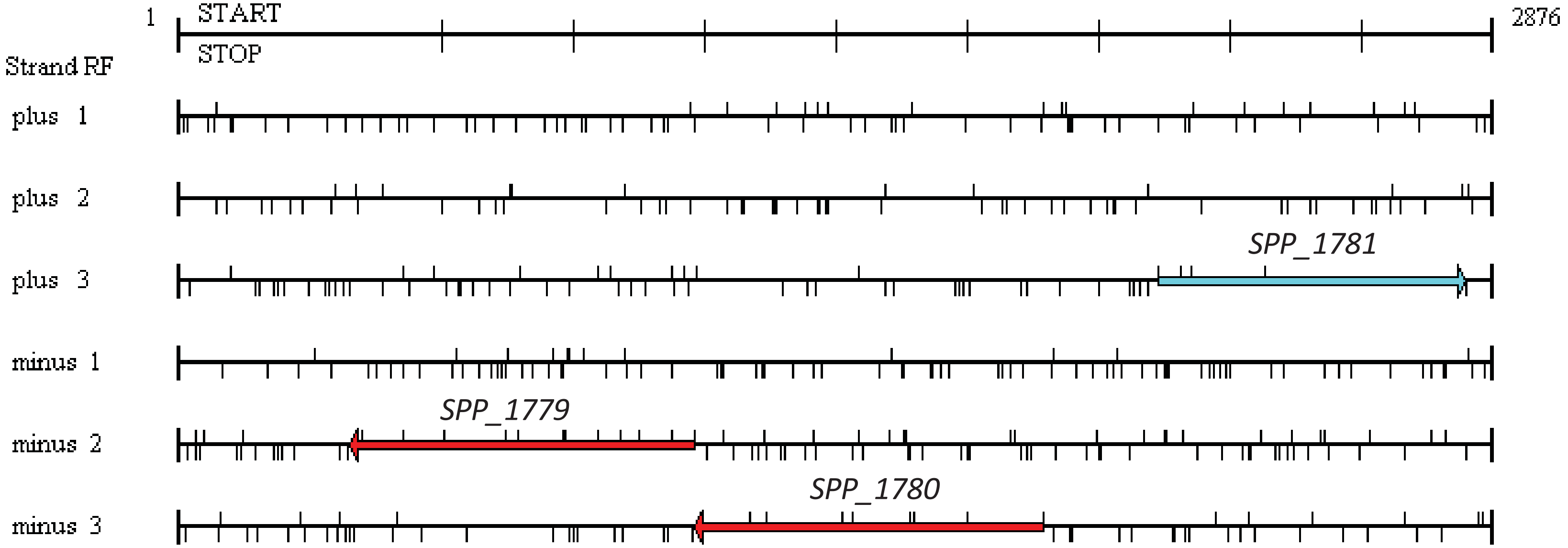


Figure 5.14 Translation overview of the assembly gap sequence of region 12
Translation overview in all six reading frames was generated using DNAMAN (Section 2.6.1), from P1031 sequence of the assembly gap of region 12 (Table 5.7). ORFs containing the start codon 'ATG' and of at least 80 amino acids in length are highlighted in blue (forward strand) and red (reverse strand). Vertical black bars indicate start codons (above) and stop codons (below).

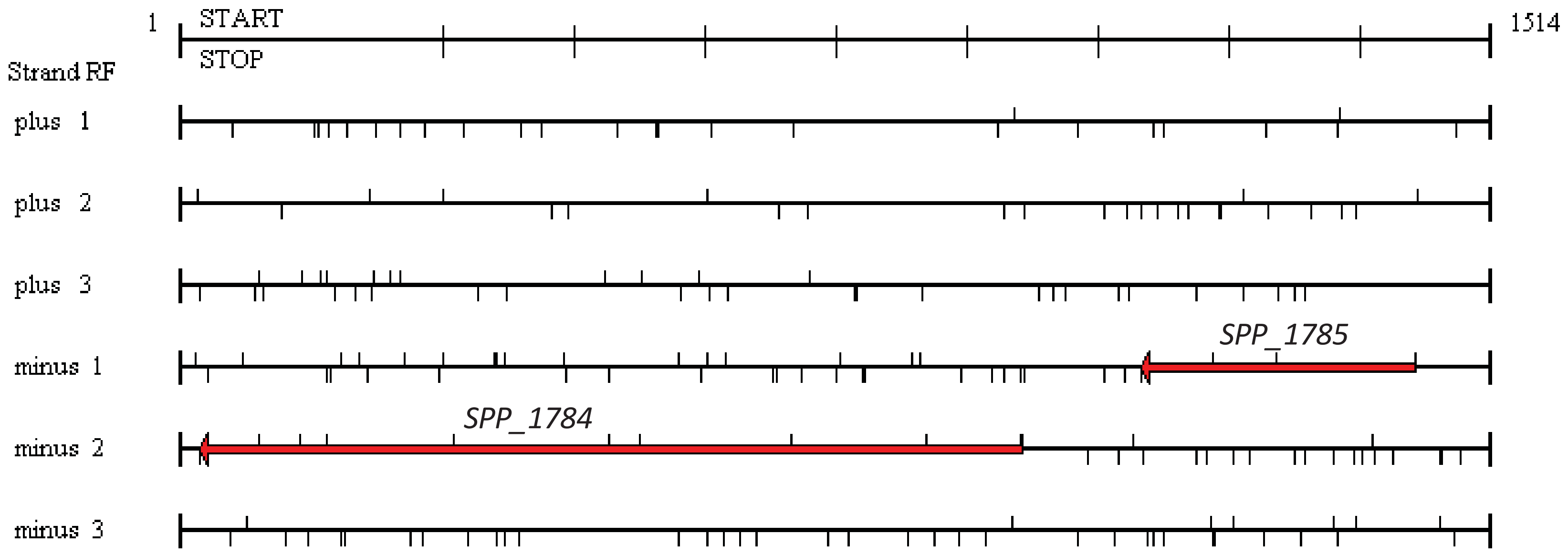


Figure 5.15 Translation overview of the assembly gap sequence of region 13
Translation overview in all six reading frames was generated using DNAMAN (Section 2.6.1), from P1031 sequence of the assembly gap of region 13 (Table 5.7). ORFs containing the start codon 'ATG' and of at least 80 amino acids in length are highlighted in red (reverse strand). Vertical black bars indicate start codons (above) and stop codons (below).

Table 5.8 P1031 virulence factors reported as interrupted in the strain 1 consensus sequence

Annotation/Putative function	ORF ID from annotated *S. pneumoniae* strains										Gap length
	TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
Pneumococcal surface protein A (PspA)	0117	0126	0185	01290	0148	0197	0163	0232	0121	0120	Multiple gaps
Hyaluronate lyase (HylA)	0314	0287	0350	02890	0310	0379	0364	0426	0285	0322	395-bp gap
Surface-associated subtilisin-like serine protease (PrtA)	0641	0558	0657	07590	0592	0702	0666	0733	0584	0599	Multiple gaps
Pneumococcal histidine triad protein B (PhtB)	1003	0889	1009	09290	0944	1043	1049	1104	0928	0977	Multiple gaps
Pneumococcal histidine triad protein D (PhtD)	1174	01037	1217	10770	1093	1226	1198	-	1073	1122	Multiple gaps

Coloured ORF IDs indicate annotation as pseudogenes

A 395-bp assembly gap was identified in the gene encoding hyaluronate lyase (*hylA*; *SPP_0350*) (Table 5.8), which has been shown to degrade hyaluronate, a major component of the host extracellular matrix (Berry *et al*., 1994; reviewed in Jedrzejas *et al*., 2004). Given the important role played by HylA in virulence, it would be important to confirm whether such a deletion exists.

The gene encoding a surface associated subtilisin-like serine protease (*prtA*; *SPP_0657*), contained multiple assembly gaps (Table 5.8). However, similar to *pspA*, *prtA* has been reported to possess regions of diversity within what is otherwise a conserved protein (Bethe *et al*., 2001). Such regions of diversity might be responsible for the assembly gaps that are present within the strain 1 consensus sequence at the position of *prtA* in the P1031 genome. However, sequence homologous to *prtA* was not found in contigs greater than 300-bp in size in the strain 1 boneyard (Table A.1).

The genes encoding the histidine triad proteins, PhtB (*SPP_1009*) and PhtD (*SPP_1217*) contained multiple assembly gaps (Table 5.8). Similar again to *pspA* and *prtA*, it is characteristic for the sequence of *phtB* and *phtD* to vary between strains (Adamou *et al*., 2001), which may contribute to the presence of assembly gaps. In addition, significant inconsistency exists in publicly available genomes regarding the annotation of *phtB* and *phtD*, probably due to the reported 87.2% sequence identity shared between the two genes (Adamou *et al*., 2001).

5.5 Clarification of discrepancies identified by genomic sequencing using PCR in strains 1, 2, 3415, 5482, 1861 & 4496

As the strain 1 consensus sequence was assembled from 36-bp reads, the assembly was vulnerable to repeated sequences of greater than 36-bp in length, such as at specific loci, or due to highly identical sequence elements, such as transposase sequence, present at multiple loci within the genome. A further complication for the interpretation of assembly gaps was sequence variation between the reference genome and the sequenced strain. In cases where the sequence identity between the sequence of the individual 36-bp reads and the reference fell below the 80% minimum required for assembly, such reads were placed in the boneyard (Tables A.3 and A.4).

Due to the limitations of assembling 36-bp sequence reads against a reference genome, it was decided to verify the discrepancies reported in Section 5.4.2, by PCR and direct sequencing. In addition, PCR and localised sequencing were used to assess whether the differences identified between the genomes of strains 1 and 1861 were associated with a virulence phenotype using non-invasive (1 and 2), intermediately virulent (3415 and 5482) and highly virulent (1861 and 4496) serotype 1 isolates.

The consensus sequences generated from strains 1 and 1861 were used to design primers that would bind to sequence shared by the two consensus sequences and flank the assembly gaps of Tables 5.7 and 5.8. These primers were designed to produce products of at least 100-bp in size within the strain predicted to lack the target sequence. In some instances reference was made to the genome of the lineage A strain, INV104B (serotype 1, ST227), which is available at http://www.sanger.ac.uk/Projects/S_pneumoniae.

5.5.1 Verification of strain 1 assembly gaps greater than 1-kb in size

As described in Section 5.4.2, the assembly gap of region 1 was approximately 33.8-kb in size. Attempts were made to amplify across the gap in the strains 1, 2, 3415, 5482, 1861 and 4496, using primers RH06Fa and RH06R (Table 2.3). It was expected that products of approximately 200-bp in size would be produced in strains lacking the P1031 sequence of region 1, such as strain 1. In addition, it was expected that the region would not be amplified in strains that possessed the region, due to the large size of the expected product (~34 kb). However, following numerous amplification attempts, the region could not be amplified in any of the six strains. Interestingly, subsequent comparisons between the primer-binding sites of RH06Fa and RH06R in the sequenced genomes of P1031 and INV104B revealed that in P1031, the RH06Fa-binding site was at position 24,071 and RH06R was at positions 58,193 and 1,803,843. In strain INV104B the RH06Fa-binding site was located at position 24,059, but in contrast to P1031, the RH06R-binding site was located at only position 1,826,379. Therefore, it was possible that in strain 1 the actual deletion of region 1 might be larger than that suggested by the consensus sequence. The fact that in P1031 RH06R was predicted to bind at two locations suggested that sequence at the 3’ end of region 1 in P1031 was homologous to sequence present at a second position on the chromosome. In contrast, the single primer-binding site in INV104B suggested that whilst the strain possessed the same sequence as above, it was only present at the second position. BLAST searches of the sequence at the RH06R-binding site of P1031 detected a phage-associated endolysin (*SPP_0083*). ClustalW alignment between the P1031 endolysin and strain 1 *lytA* showed that the two full-length genes shared 85.2% sequence identity, which had probably led to incorrect assembly of strain 1 *lytA* sequence against the phage endolysin sequence of the P1031 reference sequence. Therefore, RH06R$_{(3)}$ was designed to bind sequence at position 59,845 in P1031 and 24,342 in INV104B, which was downstream of the phage endolysin in P1031. Amplification using RH06Fa and RH06R$_{(3)}$, successfully generated products approximately 300-bp in size in strains 1, 2, 3415 and 5482 (Figure 5.16), which indicated that the non-invasive and intermediately virulent strains lacked the prophage encoded by P1031. As expected, products were not generated in strains 1861 or 4496, due to the excessive size of the expected product. In order to confirm the presence of the phage, the primers RHrtPblBF and RHrtPblBR were designed to detect *pblB*, which is a component of the phage (Table 5.7). As expected, detection of *pblB* was successful only in strains 1861 and 4496 and not the non-invasive or intermediately virulent strains. Therefore, it was found that the pneumophage, which encoded PblB, was present in only highly virulent strains. As discussed above, PblB has previously been reported to mediate binding by *S. mitis* to human platelets and is required for virulence in an animal model of endocarditis (Bensing *et al*., 2001; Mitchell *et al*., 2007). Therefore, it is possible that PblB might function as a virulence factor in the pneumococcus, and contribute to the heightened virulence of strains 1861 and 4496.

Primers RH152F and RH152Ra were used to verify the assembly gap of region 2 (Figure 5.17). However, the 3-kb product expected for strains carrying region 2 was only achieved from strain 1861 and not from the non-invasive strains, the intermediately virulent strains, or strain 4496. Instead a 650-bp product was produced from strains 1, 2, 3415, 5482 and 4496. Therefore, the genes of region 2 were not consistently associated with a virulence phenotype.

As described in Section 5.4.2, the assembly gap of region 3 was predicted to encode ZmpB. Therefore, primers RH208/09F and RH208/09R were designed to

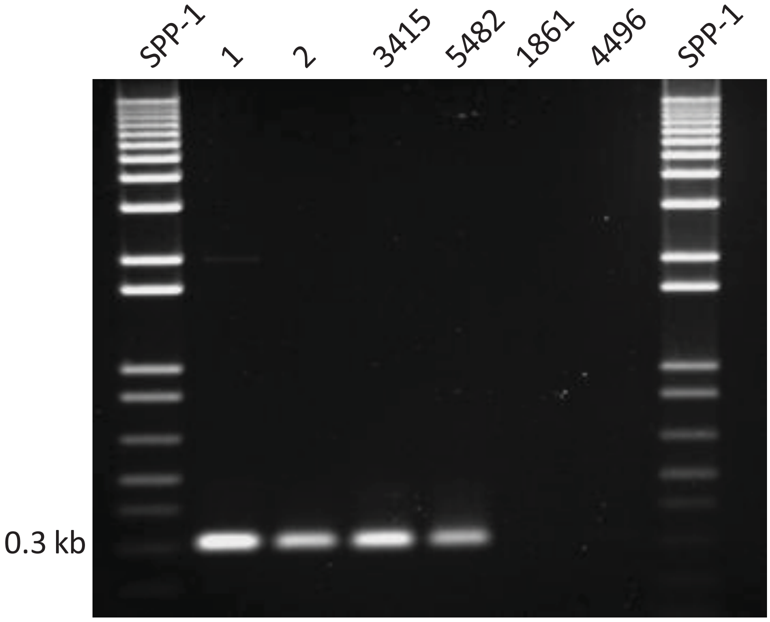


Figure 5.16 Amplification across the assembly gap of region 1
PCR amplification was performed using primers RH06Fa and RH06R$_{(3)}$ (Section 2.9.4), from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the 1kb plus marker, as described in Section 2.9.1.

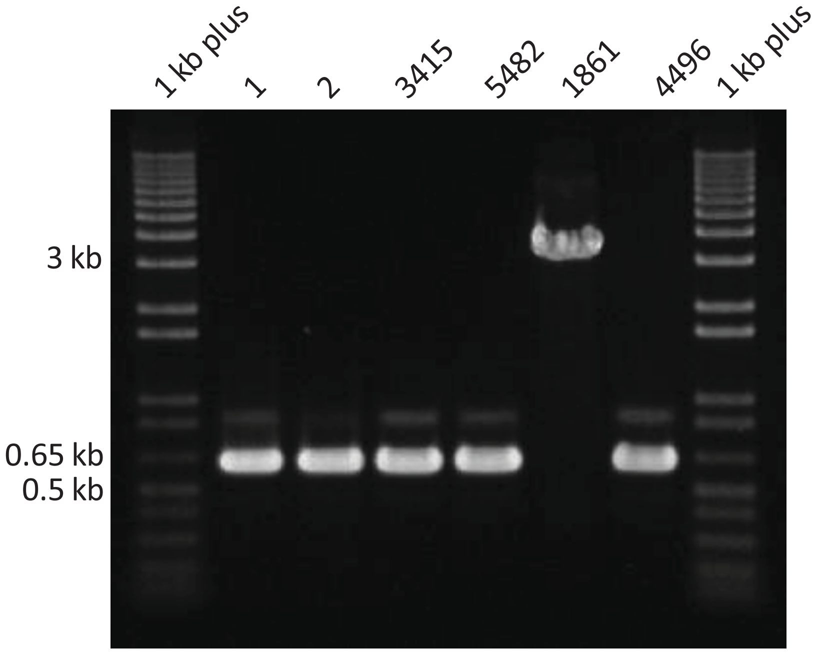


Figure 5.17 Amplification across the assembly gap of region 2
PCR amplification was performed using primers RH152F and RH152Ra, (Section 2.9.4) from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the 1kb plus marker, as described in Section 2.9.1.

amplify across the region in order to identify whether significant differences in size existed between *zmpB* in the non-invasive, intermediately virulent or highly virulent strains. Successful amplification was achieved from all strains, albeit weakly from strains 3415 and 4496, which produced a product of approximately 4.5-kb in size (Figure 5.18). Sequence variation in the primer-binding sites probably contributed to the differences in band intensity between strains. Therefore, it was unlikely that the size of *zmpB* differed significantly between strains, and instead the assembly gaps highlighted sequence variability between *zmpB* of strains 1 and 1861. This is supported by the presence of *zmpB* sequence in the strain 1 boneyard within contigs 2028, 1215 and 1687 (Table A.1). In future work, sequencing *zmpB* in all six strains would be required to confirm whether a particular allele is associated with a virulence phenotype.

Amplification across the assembly gap of region 4 sequence was attempted using RH224/5F and RH224/5R, which produced a 1.5-kb product from the non-invasive and intermediately virulent strains and a 2-kb product from strains 1861 and 4496 (Figure 5.19). The product produced from the non-invasive and intermediately virulent strains was larger than the expected 100-bp product that would have been produced had the assembly gap been due to only the absence of the region 4 P1031 sequence. Therefore, the 1.5-kb product of strains 1 and 5482 was sequenced using primers RHseq224/5F and RHseq224/5R. The resultant sequence was assembled using the sequence assembly tool in DNAMAN (Section 2.6.1), which produced the 1,531-bp consensus sequence used to generate the translation overview in Figure 5.20. Three ORFs were predicted to encode proteins of greater than 80 amino-acids in length, which included a truncated sodium-dependent transporter homologous to *SPD_0642* and protein A and B components of an IS*1381* transposase, which are homologous to multiple transposases within all *S. pneumoniae* genomes in the KEGG database.

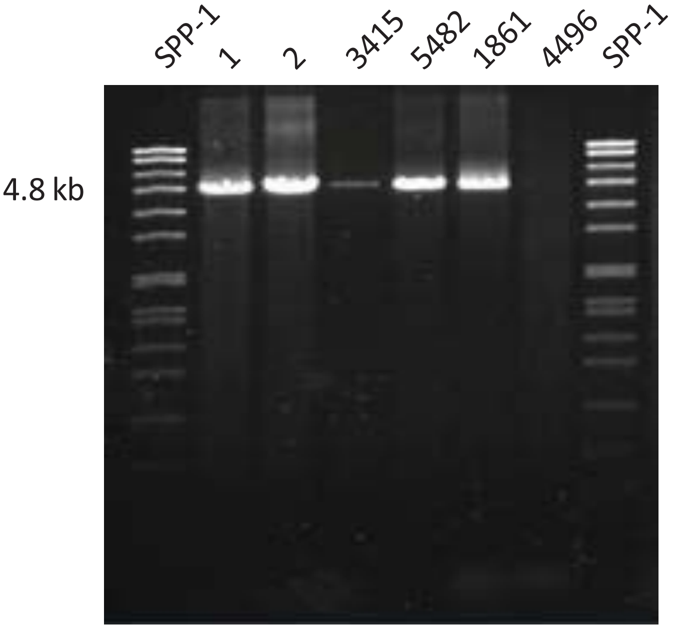


Figure 5.18 Amplification across the assembly gap of region 3
PCR amplification was performed using primers RH208/09F and RH208/09R (Section 2.9.4), from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the SPP1 marker, as described in Section 2.9.1.

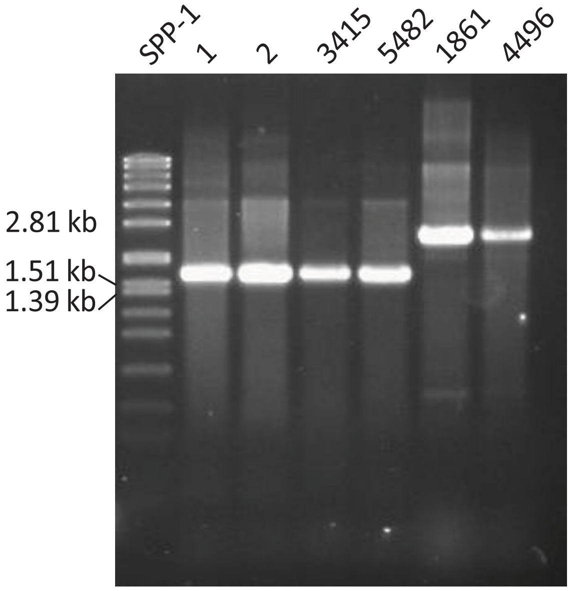


Figure 5.19 Amplification across the assembly gap of region 4
PCR amplification was performed using primers RH224/5F and RH224/5R (Section 2.9.4), from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the SPP1 marker, as described in Section 2.9.1.

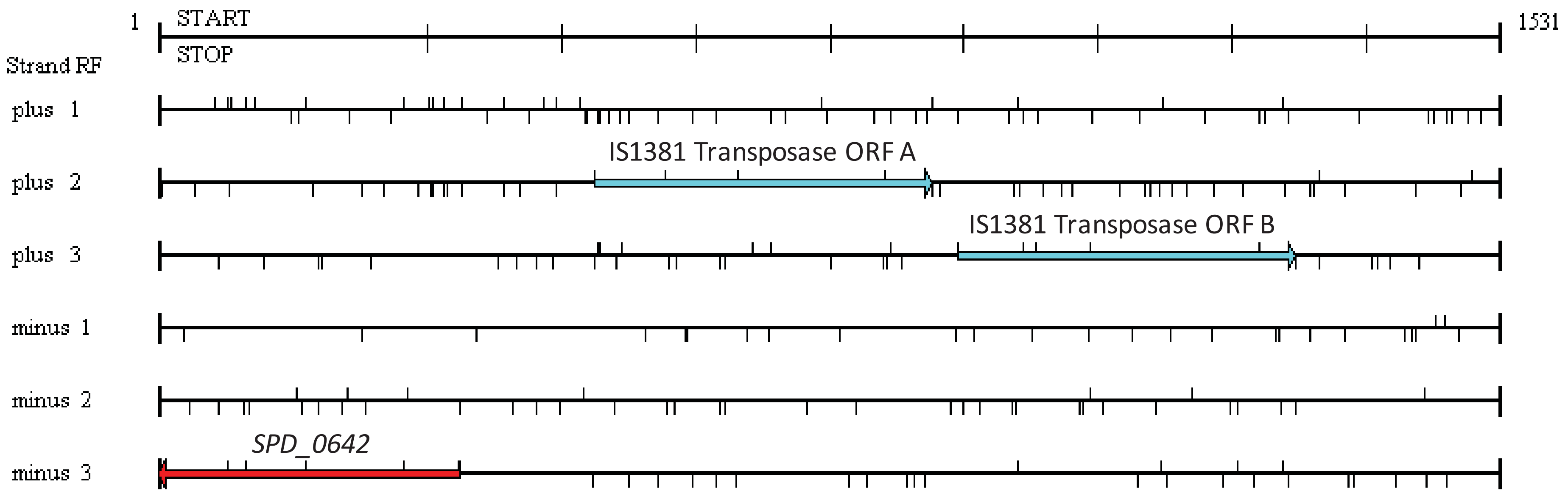


Figure 5.20 Translation overview of 5482 sequence in assembly gap 4
Translation overview in all six reading frames was generated using DNAMAN (Section 2.6.1), from sequencing of strain 5482 RHseq224/5F – RHseq224/5R product. ORFs containing the start codon ‘ATG’ and of at least 80 amino acids in length are highlighted in blue (forward strand) and red (reverse strand). Vertical black bars indicate start codons (above) and stop codons (below).

Therefore, region 4 of the non-invasive strains and intermediately virulent strains differed from the highly virulent strains, due to a 214 amino acid shorter sodium-dependent transporter (*SPD_0642*) and the replacement of *merR* (*SPP_0750*) and *mutT/nudix* (*SPP_0751*) with two transposase-related genes. Therefore, *SPP_0750* and *SPP_0751* were only present in the highly invasive serotype 1 isolates of this study and could have a role in virulence.

Amplification across the assembly gap of region 7 was attempted using primers RH344F and RH344R. However, successful amplification was not expected from strains 1861 and 4496, as a 64-kb product would need to be amplified if the region 7 sequence was indeed present. As expected, products were not produced from strains 1861 or 4496. However, 5-kb products were produced from the non-invasive and intermediately virulent strains (Figure 5.21), which were larger than the expected 150-bp product. Interestingly, when the 5-kb products from strains 1 and 3415 were sequenced using RHseq344F and RHseq344R, the product was confirmed to consist of *iga* sequence. Therefore, sequence variation between *iga* of the highly virulent strains and the lineage A strains probably led to the lack of assembly of strain 1 *iga* sequence against the P1031 reference. In addition, *iga* sequence of strain 1 was identified in the boneyard within contigs 1170, 1264, 1853, 1851, 1691 and 2125 (Table A.1). The distribution of *iga* sequence across so many contigs was probably as a result of some scattered assembly of strain 1 *iga* sequence against P1031, thus resulting in incomplete assembly of *iga* from the boneyard sequences. However, with the exception of *iga*, the P1031 genes of region 7 were absent from the non-invasive and intermediately virulent strains. The presence of *zmpD* in strains 1861 and 4496, and its absence from the lineage A strains was confirmed by PCR using RHrtzmpDF$_{(2)}$ and RHrtzmpDR$_{(2)}$. Subsequently, RHumuCF – RHumuCR, RHint5252F – RHint5252R and RHint916F –

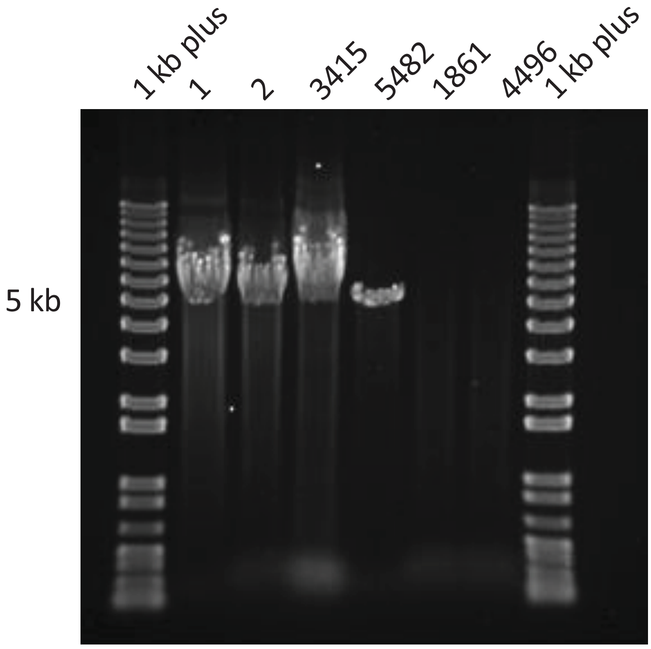


Figure 5.21 Amplification across the assembly gap of region 7
PCR amplification was performed using primers RH344F and RH344R (Section 2.9.4), from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the 1kb plus marker, as described in Section 2.9.1.

RHint916R were used to confirm the presence of key Tn*5253* components, in strains 1861 and 4496 and the absence of the region in the non-invasive and intermediately virulent strains. The significance of Tn*5253* is highlighted by the presence of genes conferring resistance to tet and cml (Ayoubi *et al*., 1991). In addition, *umuC/mucA*-(*SPP_1190*) and *umuC/mucB*-like (*SPP_1191*) homologs have been suggested to be involved in an SOS-like response (Munoz-Najar & Vijayakumar, 1999). However, since *SPP_1190* and *SPP_1191* appeared to be absent or interrupted in strain 1861 (Table 5.6), direct sequencing of this region would be required to confirm whether *SPP_1190* and *SPP_1191* could have a role in the virulence of strains 1861 and 4496.

Amplification across the assembly gap of region 8 was attempted using RH383F and RH383R, and produced the expected 2.8-kb product from strains 1861 and 4496 (Figure 5.22). Successful amplification was also achieved across the assembly gap in the non-invasive and intermediately virulent strains, producing the expected 400-bp product. Therefore, the P1031 ORFs *SPP_1340 – SPP_1343* were found to be associated with heightened virulence in the tested strains. However, since *SPP_1342* is annotated as a pseudogene due to a frameshift and exhibits 100% sequence identity to the strain 1861 consensus sequence, the gene may be non-functional in strain 1861, thus leaving only a transposase and a putative high-affinity Fe^{2+}/Pb^{2+} permease (*SPP_1340*) within region 8. The role of the putative high-affinity Fe^{2+}/Pb^{2+} permease in virulence might be worth investigating given the importance of metal ion transport *in vivo* (Section 1.3.1.3).

Amplification across the assembly gap of region 9 was attempted using RH386F and RH386R, but was only successful from strains 1861 and 4496 producing a product approximately 20-kb in size (Figure 5.23). Searches were undertaken for the relative

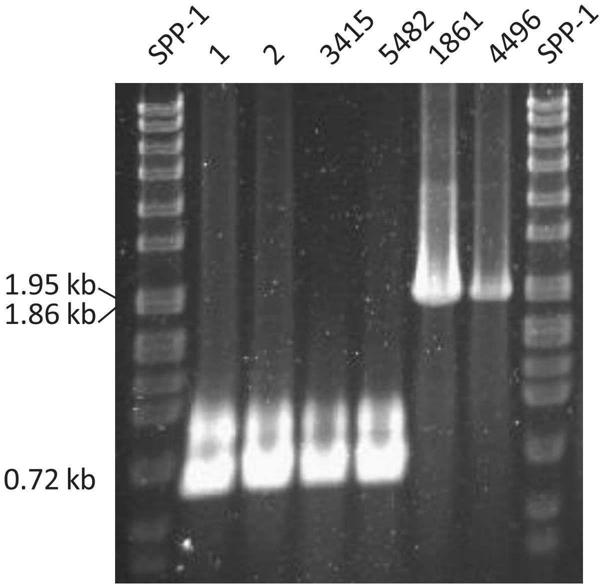


Figure 5.22 Amplification across the assembly gap of region 8
PCR amplification was performed using primers RH383F and RH383R (Section 2.9.4), from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the SPP1 marker, as described in Section 2.9.1.

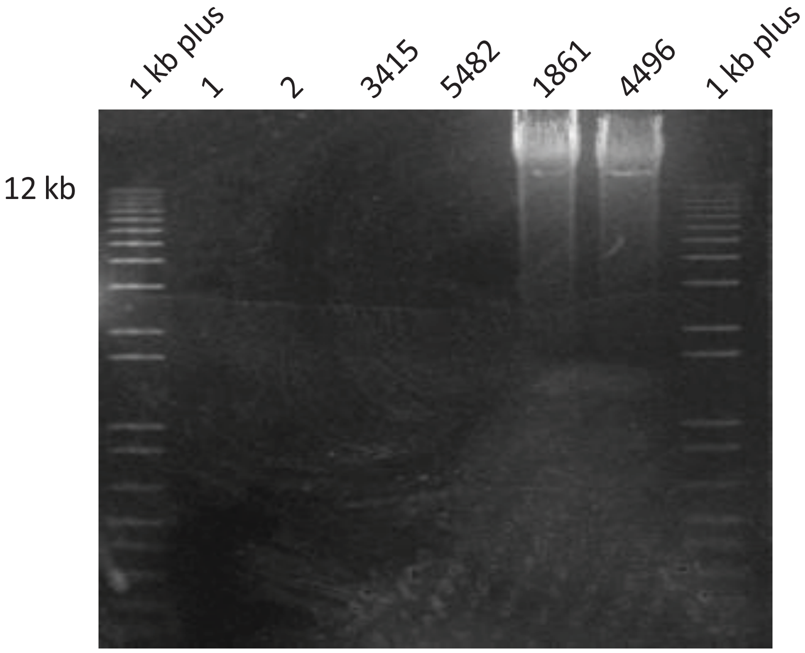


Figure 5.23 Amplification across the assembly gap of region 9
PCR amplification was performed using primers RH386F and RH386R (Section 2.9.4), from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the 1kb plus marker, as described in Section 2.9.1.

positions of the RHseq386F- and RHseq386R-binding sites in the INV104B and P1031 genomes (Table 5.8), which revealed that if the location of the primer-binding sites within the non-invasive and intermediately virulent strains were the same as that in INV104B, then amplification using RHseq386F and RHseq386R would not be possible. Therefore, the location of the RHseq383F primer-binding sites was determined in P1031 and INV104B (Table 5.8), and was found to be located 25.7-kb and 29.2-kb upstream of RHseq386R in P1031 and INV104B respectively. As the assembly gap of region 9 commenced approximately 7.6-kb downstream of RHseq383F, an alignment between the RHseq383F – RHseq386R sequence of P1031 and INV104B was performed using the ACT (Section 2.6.1) to determine the relative positions of homologous sequence between the two primer-binding sites in these strains (Figure 5.24). In order to identify ORFs that are shared by both P1031 and INV104B and those that are unique to either strain, a translation overview was generated for the RHseq383F – RHseq386R sequence of INV104B (Figure 5.25) and BLASTp searches were used to predict the function of the ORFs in this region of INV104B (Table 5.9). By comparing the alignment of Figure 5.24 and the BLASTp results of Tables 5.7 and 5.9, only ORFs *SPP_1344* and *SPP_1364* were shared between both strains. In order to compare the above alignment to the strains of this study, primers RH386F$_{(2)}$, RH386R$_{(2)}$, RH386F$_{(3)}$ and RH386R$_{(3)}$ were designed to bind to positions indicated in Figure 5.24 and produce the products in Table 5.10 for strains harbouring either the P1031 or INV104B regions. Amplification using RH386F$_{(2)}$ and RH386R$_{(2)}$ produced products 2-kb in size from strains 1 and 2, 1.5-kb in size from strain 3415 and 5482 and approximately 14-kb in size from strains 1861 and 4496 (Figure 5.26a), which were as expected for strains 3415, 5482, 1861 and 4496, but not strains 1 and 2. Therefore, strains 1 and 2 possess an additional 500-bp of sequence between the RH386F$_{(2)}$- and RH386R$_{(2)}$-binding sites.

Table 5.8 Relative positions of primers RHseq386F- and RHseq386R-binding sites in the genomes of P1031 and INV104B

	P1031 (strand)	**INV104B (strand)**
RHseq386F	1,260,346 (+)	613,214 (-)
RHseq386R	1,278,452 (-)	1,259,131 (-)
Approx. size 386F → 386R	18.1 kb	-
RHseq383Fa	1,252,707 (+)	1,229,970 (+)
Approx. size 383Fa → 386R	25.7 kb	29.2 kb

(+/-) Indicates '+' forward strand and '-' reverse strand

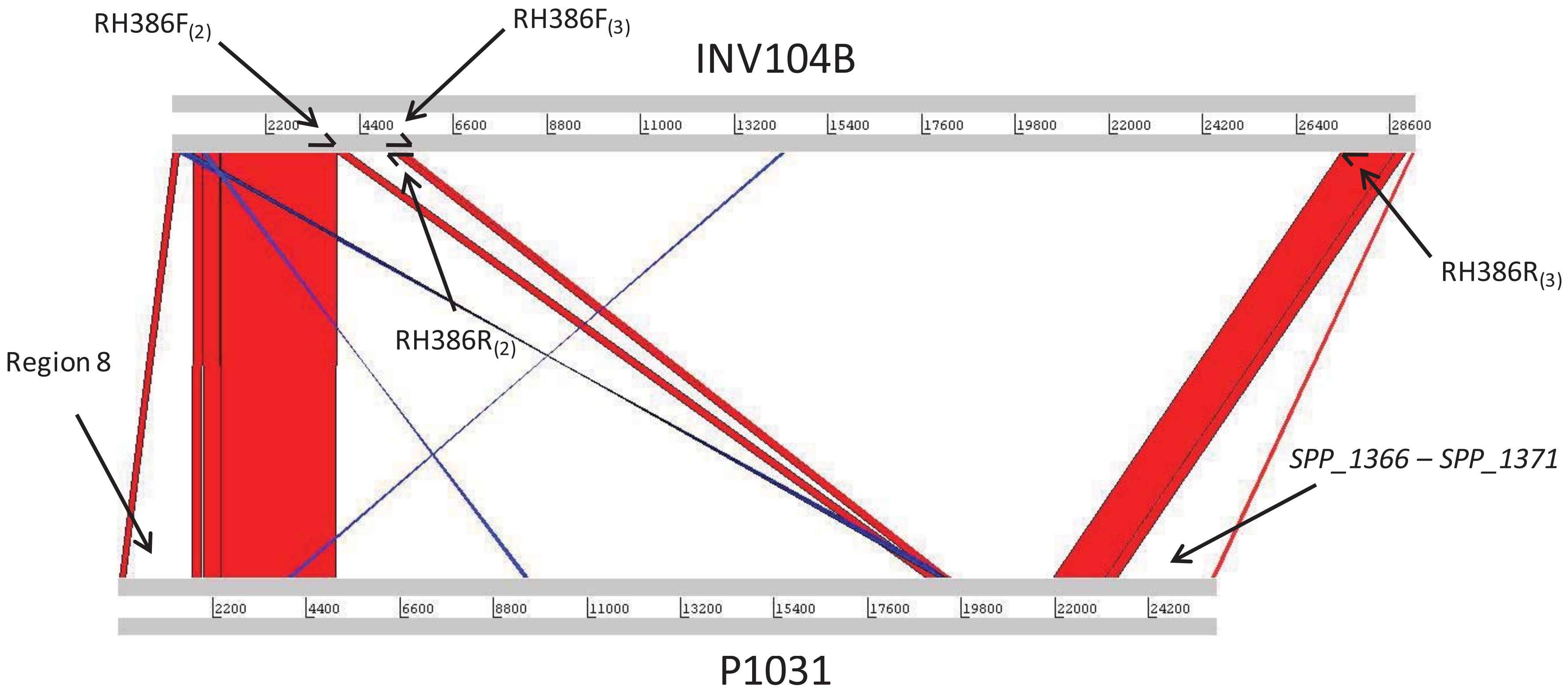


Figure 5.24 Alignment of sequence between the binding sites of primers 383F and 386R of INV104B and P1031
Generated as described in Section 2.6.1. Red indicates regions of at least 80% sequence identity over a region of at least 100 bp. Blue indicates regions of inverted sequence of at least 80% identity over a region of at least 65 bp. The relative position of the primers designed to amplify components of the assembly gap of region 9 are indicated above. The position of the assembly gap of region 8 and ORFs *SPP_1366 – SPP_1371* are also indicated.

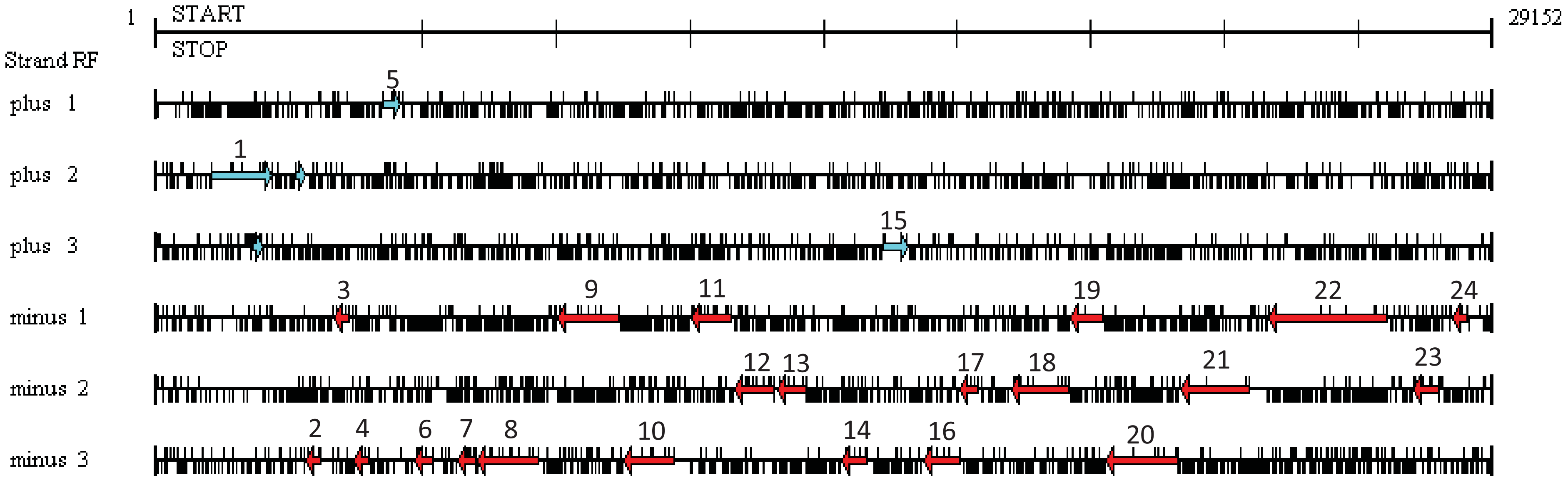


Figure 5.25 Translation overview of Region 9 in INV104B
Translation overview in all six reading frames was generated using DNAMAN (Section 2.6.1), from INV104B sequence between the primer-binding sites of RH383Fa and RH386R. ORFs containing the start codon 'ATG' and of at least 80 amino acids in length are highlighted in blue (forward strand) and red (reverse strand). BLASTp results of ORFs numbered 1 – 24 are shown in Table 5.9. Vertical black bars indicate start codons (above) and stop codons (below).

Table 5.9 BLASTp results of INV104B ORFs in region 9

	Annotation/Putative function	ORF ID from annotated *S. pneumoniae* strains									
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC 700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary-19A (SPH)	G54 (SPG)	CGSp14 (SPCG)
1	Glutamate dehydrogenase (EC 1.4.1.4)	1306	1158	1344	11970	1219	1369	0923	1446	1197	1272
2	ABC transporter, ATP-binding protein	-	1160	1346	09810	0998	1371	-	-	1199	-
3	Transposase, ORF 2	+	+	+	-	+	+	+	+	+	+
4	Putative transposase	+	+	+	+	+	+	+	+	+	+
5	IS630-SpnII, transposase	+	+	+	-	-	+	+	+	0665	+
6	Hypothetical	-	-	-	-	-	-	-	0833	-	-
7	Diadenosine tetraphosphate hydrolase, histidine-triad family protein	-	-	-	-	-	-	-	0830	-	-
8	Putative dihydrolipoamide dehydrogenase	-	-	-	-	-	-	-	0829	-	-
9	Bacterial extracellular solute-binding protein	-	-	-	-	-	-	-	0828	-	-
10	Glycoside hydrolase family protein (EC 3.2.1.26)	-	1166	1353	-	-	1377	-	0827	-	-
11	Binding-protein-dependent transport systems inner membrane component (LplC)	-	-	-	-	-	-	-	0826	-	-
12	Sugar ABC transporter substrate binding protein (LplB)	-	-	-	-	-	-	-	0825	-	-
13	Putative N-acetylmannosamine-6-phosphate epimerase	1685	1497	1702	12220	1579	1724	1623	0824	1593	1658
14	Phosphosugar-binding transcriptional regulator, RpiR family	-	-	-	12240	-	-	-	0823	-	-
15	Hypothetical	-	-	-	12250	-	-	-	0822	-	-
16	IS3-Spn1, transposase	+	+	+	+	+	+	+	+	+	+
17	IS861 transposase Orf1	-	-	-	12270	-	-	+	+	-	0026
18	Hypothetical protein	1340	1175	-	12290	-	1380	-	1477	-	1454
19	ABC transporter, ATP-binding protein	1341	1176	-	12300	-	1381	-	1478	-	1455
20	Drug efflux ABC transporter, ATP-binding/permease protein	1342	1177	-	12310	-	1382	-	1479	-	-
21	Prolyl oligopeptidase family protein	1343	1178	-	12320	-	1383	-	-	-	-
22	Putative lantibiotic synthetase	1367	1201	1387	13330	1267	1406	0906	1499	1308	1356
23	Tn*5253* CAAX amino terminal protease family	1346	1180	1364	12350	1248	1386	-	1482	1287	0847
24	Hypothetical	1348	1182	1147	12370	1250	1388	-	1235	1227	-

Colour key for gene descriptions are described in Table 5.5

Coloured ORF IDs indicate annotation as pseudogenes (Dark) or are positioned at an alternative location within the genome (green)

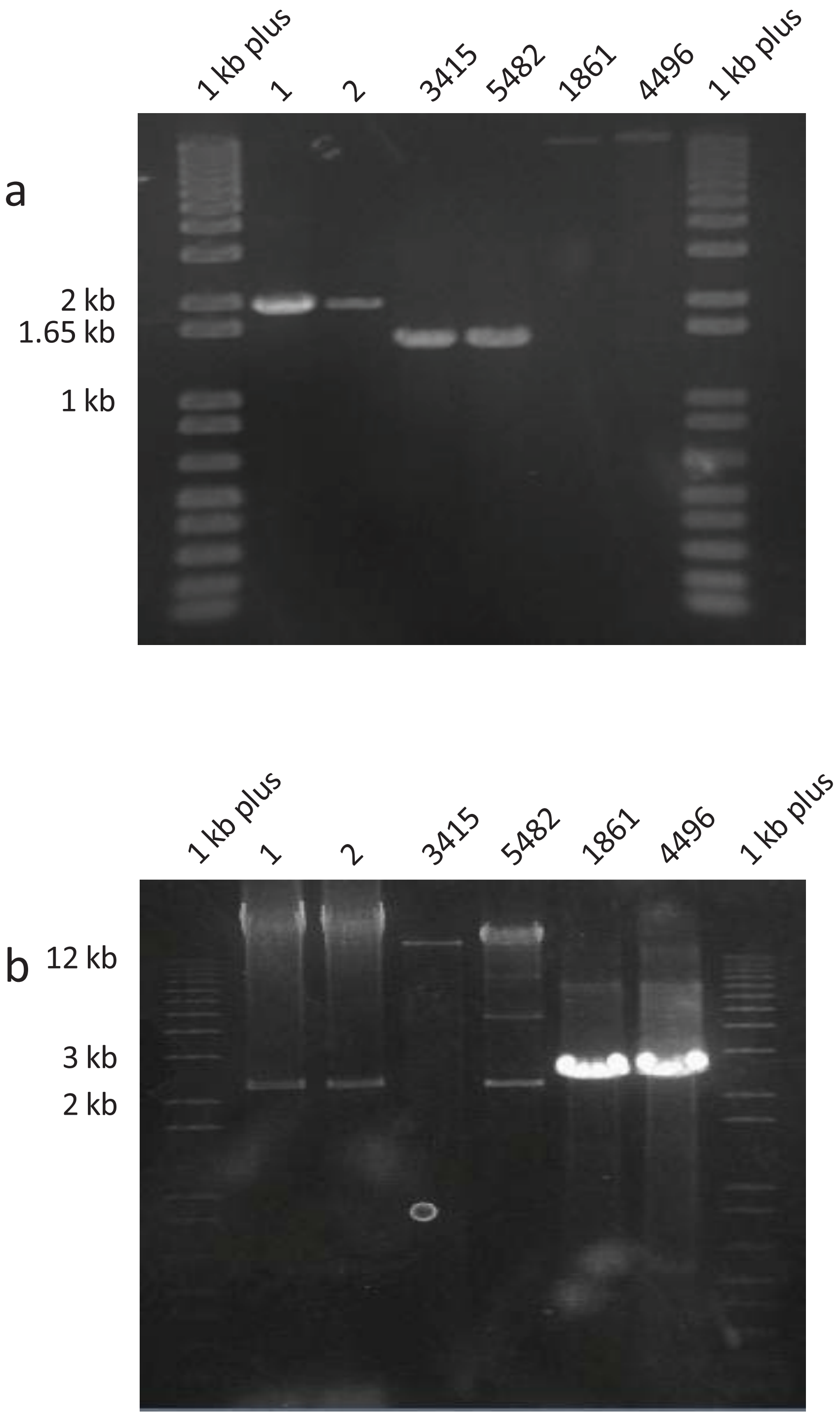


Figure 5.26 Amplification across the assembly gap of region 9
PCR amplification was performed using primers (a) RH386F$_{(2)}$ and RH386R$_{(2)}$ and primers RH386F$_{(3)}$ and RH386R$_{(3)}$ (Section 2.9.4), from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the 1kb plus marker, as described in Section 2.9.1.

Difficulty was encountered when attempting to amplify using primers RH386F$_{(3)}$ and RH386R$_{(3)}$, probably due to the repeated nature of the sequence to which these primers bind. However, the predominant bands produced from RH386F$_{(3)}$ – RH386R$_{(3)}$ amplification were >> 12 kb from strains 1, 2, 3415 and 5482 and 2.5-kb in size from strains 1861 and 4496 (Figure 5.26b). Therefore, strains 1, 2, 3415 and 5482 were likely to possess the sequence present between the RH386F$_{(3)}$- and RH386R$_{(3)}$-binding sites in INV104B, and strains 1861 and 4496 were likely to possess the P1031 region. Furthermore, strains 3415 and 5482 were suggested to harbour the INV104B version of region 9, whilst strains 1861 and 4496 were likely to harbour the P1031 version of region 9. In strains 1 and 2, the INV104B version of the region appeared to most likely be present, with the exception of an additional 500 bp in the first portion of the region (RH386F$_{(2)}$ – RH386R$_{(2)}$). In addition, the presence of sequence homologous to ORFs *SPH_0824 – SPH_0830* and ORFs *SPH_1777 – SPH_1779* in contigs 1318, 1280, 1531, 1652 and 1481 of the strain 1 boneyard, suggested that strain 1 possessed much of the INV104B version of region 9.

Table 5.10 Estimated sizes of PCR products used to compare the content of region 9 to either P1031 or INV104B

	P1031	INV104B
RH386F$_{(2)}$ – RH386R$_{(2)}$	14.3 kb	1.5 kb
RH386F$_{(3)}$ – RH386R$_{(3)}$	2.5 kb	22 kb

Key components of region 9 include an enzyme of the *N*-acetylneuraminate lyase subfamily (*SPP_1350*), a cytidine deaminase (*SPP_1351*; EC 3.5.4.5), a putative phosphatidylglycerophosphatase A (*SPP_1352*; EC 3.1.3.27), an invertase (*SPP_1353*; EC 3.2.1.26), an ABC transporter (*SPP_1354 – SPP_1357*), a kelch-like protein (*SPP_1358*), an N-acetylmannosamine 6-phosphate 2-epimerase (*SPP_1359*; EC

5.1.3.9), a CAAX amino protease (*SPP_1364*) and an arsenate reductase (*SPP_1366*), which are present in only the highly virulent strains. Enzymes of the N-acetylneuraminate lyase (NAL) sub-family have been shown to share a common catalytic step, but are involved in different biological pathways. In addition, the differences that distinguish members of the NAL sub-family are subtle and are difficult to confirm from sequence data alone (Barbosa *et al*., 2000). The putative ABC transporter encoded by *SPP_1354 – SPP_1357* has been suggested to be involved in oligopeptide transport. However, the actual substrate-specificity of the transporter is unknown. The kelch-like protein encoded by *SPP_1358* is homologous to YjhT of *E. coli*, which functions to relaese α-N-acetylneuraminic acid from complex sialoconjugates, and has been shown to provide an *in vivo* survival advantage in some bacteria (Severi *et al*., 2008). Interestingly, the putative *YjhT*-like gene is followed by a putative N-acetylmannosamine-6-phosphate-2-epimerase, which is involved in sialic acid catabolism (Plumbridge *et al*., 1999).

In INV104B the region can be divided into two parts flanked by transposases, with the first part encoding genes homologous to *SPH_0833 – SPH_0822* of Hungary 19A and the second part encoding genes homologous to *SPN23F_12290 – SPN23F_13330* of ATCC 700669 (Table 5.9). The second part of the region corresponds to part of AR 28 (Blomberg *et al*., 2009). However, the first part, which corresponds to a region of Hungary 19A, has not been characterised. The first region encodes a bacterial extracellular solute-binding protein and an ABC transporter homologous to LplB and LplC, involved in the import of lactose in *B. subtilis* (Quentin *et al*., 1999). The second part of region 9 in INV104B encodes a putative ABC transporter (*SPN23F_12300* and *SPN23F_12310*) and putative bacteriocin-modification enzymes homologous to *SPN23F_12320* and *SPN23F_12330* of ATCC 700669. Region

9 in both strains P1031 and INV104B, which correspond to the highly virulent and lineage A strains, respectively, encodes proteins required for transport and metabolism of sugars. Perhaps differences between the sorts of genes present in each strain may reflect differences in the range of sugars that can be utilised as an energy source by each group of strains.

Initially, amplification was attempted separately across the assembly gaps of regions 10 and 11 using primers RH430F – RHseq430R and RH434Fa – RH434Ra respectively. However, amplification was successful only from strain 1861. Closer analysis of the three P1031 ORFs present between gap 10 and 11 (*SPP_1634 – SPP_1636*) revealed the presence of transposase-like sequence, which suggested that strain 1 sequence could have been incorrectly assembled at this location. Therefore, amplification spanning both regions 10 and 11 was attempted using RHseq430F and RHseq434Ra (Figure 5.27). A 400-bp product was produced from the non-invasive strains, intermediately virulent strains and strain 4496, and a 10-kb product was produced from strain 1861. These PCR results confirmed the CGH data (Section 5.2.3), which detected homologues of *SP_1612 – SP_1621* only in strain 1861 and not in the non-invasive strains, the intermediately virulent strains or strain 4496. Therefore, the genes encoded by *SPP_1634 – SPP_1636*, which correspond to AR 31 (Blomberg *et al*., 2009), are not required for the virulence of strain 4496 and were not associated with a particular virulence phenotype.

Amplification across the assembly gap of region 12 was attempted using RH467F and RH467R (Figure 5.28). Products of approximately 300-bp in size were generated from the non-invasive and intermediately virulent strains and a 3-kb product was produced from strains 1861 and 4496, which suggested that the ABC-2 type

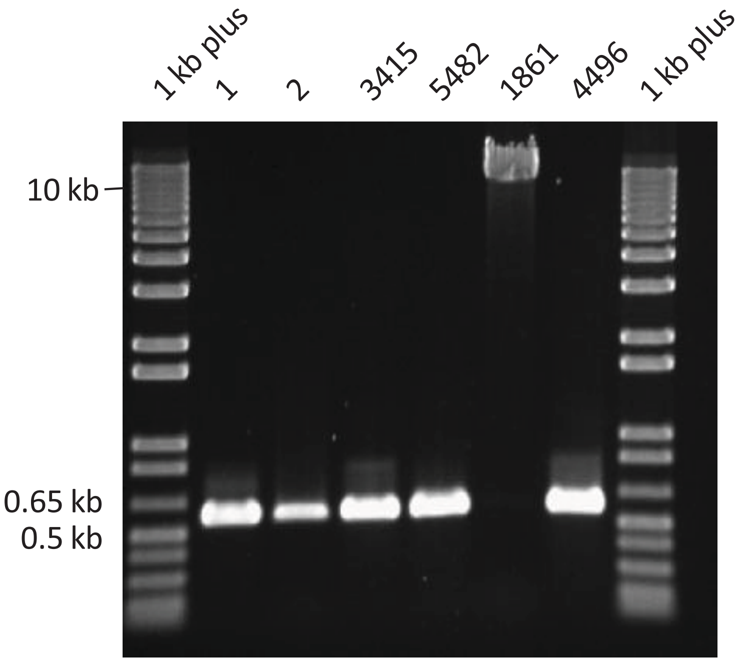


Figure 5.27 Amplification across the assembly gap of regions 10 and 11
PCR amplification was performed using primers RH430F and RH434Ra (Section 2.9.4), from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the 1kb plus marker, as described in Section 2.9.1.

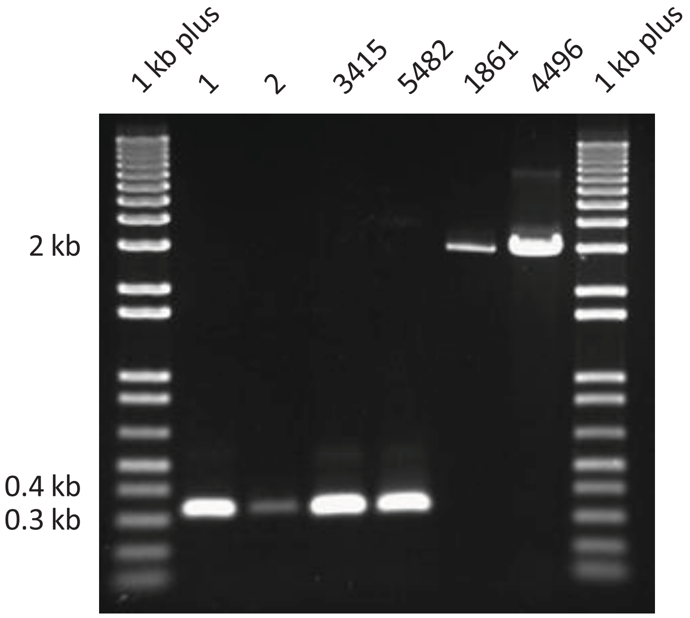


Figure 5.28 Amplification across the assembly gap of regions 12
PCR amplification was performed using primers RH467F and RH467R (Section 2.9.4), from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the 1kb plus marker, as described in Section 2.9.1.

transport protein (*SPP_1779*), the nod-factor export ATP-binding protein (*SPP_1780*) and the ArsR family transcriptional regulator (*SPP_1781*) of region 12 were associated with heightened virulence. However, it is not clear what role, if any, the products of these genes might play.

Amplification across the assembly gap of region 13 using RH469F and RH469R, produced 5-kb products from the two non-invasive strains and 1.5-kb products from the invasive strains 3415, 5482, 1861 and 4496 (Figure 5.29). Whilst the 1.5-kb product was expected from strains possessing the P1031 sequence of region 13, direct sequencing of the region in strains 1 and 2 was required to identify the content of the region in these strains. The 5-kb product of the non-invasive strains was sequenced using primers RHseq469F and RHseq469R (Section 2.9.8), which revealed the presence of a ribosomal protein L11 methyltransferase (*prmA*), which was homologous to *SPT_1705* of Taiwan 19F and *SPP_1783* of P1031. In addition, a putative AAA+ ATPase, homologous to *SPJ_1684* and a putative 7, 8-dihydro-8-oxoguanine-triphosphatase (nudix), homologous to *SPJ_1685* and *SPP_1785* of P1031 were also identified in strain 1. The discontinuous nature of assembly across region 13 of strain 1 sequences and the detection of some P1031 genes by direct sequencing suggested that the region contains a mix of genes vulnerable to variation. It is probably unlikely that the genes of region 13 play a role in virulence.

Amplification across the assembly gap of region 14 was attempted using primers RH513F and RH513R (Figure 5.30). Successful amplification was achieved from all six strains and confirmed the presence of the hypothetical protein encoded by *SPP_2148* in strain 1861 by producing the expected 1.5-kb product. In contrast, 100-bp products were produced from the non-invasive strains, the intermediately virulent strains and strain 4496, which confirmed the absence of *SPP_2148* in these strains. Therefore, *SPP_2148*

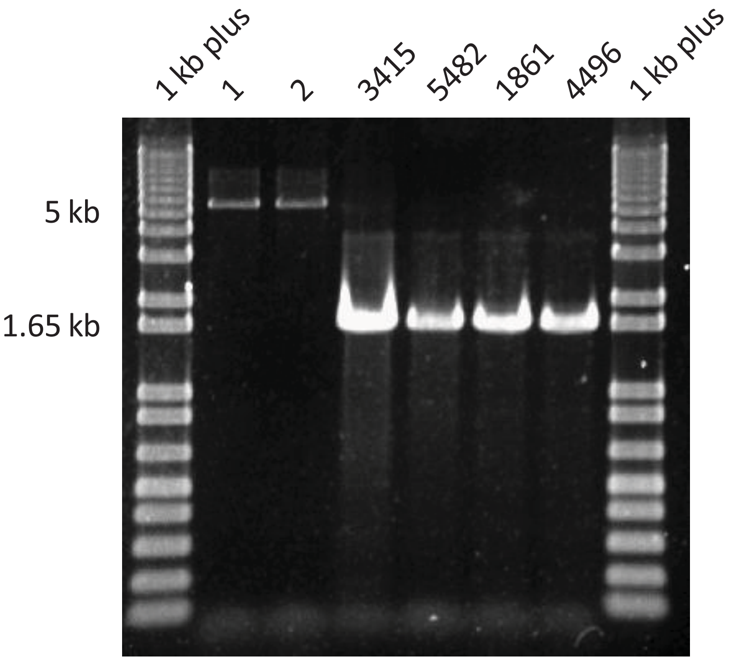


Figure 5.29 Amplification across the assembly gap of regions 13
PCR amplification was performed using primers RH469F and RH469R (Section 2.9.4), from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the 1kb plus marker, as described in Section 2.9.1.

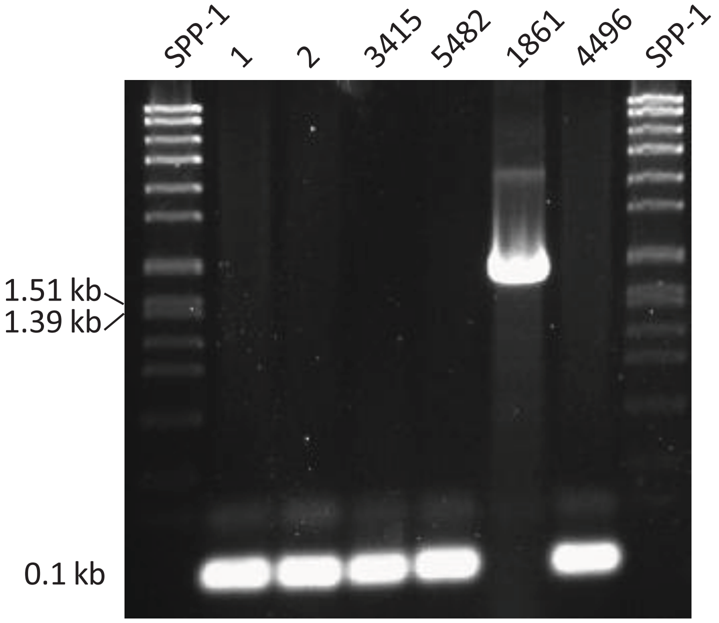


Figure 5.30 Amplification across the assembly gap of region 14
PCR amplification was performed using primers RH513F and RH513R (Section 2.9.4), from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the SPP1 marker, as described in Section 2.9.1.

was not found to be associated with a particular virulence phenotype and is not required for the virulence of strain 4496.

5.5.2 Verification of strain 1 sequencing gaps within virulence factors

In Section 5.4.2.2 a number of discrepancies were identified in the strain 1 consensus within sequence of the virulence factors *pspA*, *prtA*, *hylA*, *phtB* and *phtD*. RHPspAF and RHPspAR were used to verify the assembly gaps within *pspA* (Figure 5.31). 1.3-kb products were produced from strains 1, 2, 3415, 5482 and 1861. However, the failure to amplify from strain 4496 was probably due to some sequence variation in the primer-binding sites within *pspA*. Previous work discussed in Section 1.6.2, successfully detected PspA by Western blot, from strain 4496, confirming that 4496 does produce the protein. Therefore, the discrepancies that were identified in Section 5.4.2.2 were most likely due to sequence variability in the strain 1 gene compared to P1031, rather than *pspA* of strain 1 being shorter than that of P1031. Given the variable nature of *pspA* (Hollingshead *et al*., 2000), it is not surprising that *pspA* sequence variation exists between strains. However, in future work it would be interesting to identify by direct sequencing whether particular alleles of *pspA* are associated with the invasive potential of the serotype 1 isolates.

Similarly, amplification across the assembly gaps present in the strain 1 consensus sequence of *prtA* using RHPrtAF and RHPrtAR produced products that were the same size (4.5 kb) in strains where amplification was successful (Figure 5.32). A product was not produced at all from strain 3415 and a very weak band was produced from strain 4496. Poor amplification from strains 3415 and 4496 was probably due to variation in the sequence of the primer-binding sites within *prtA*. However, in any case an association between the size of the product produced for *prtA* and virulence was not

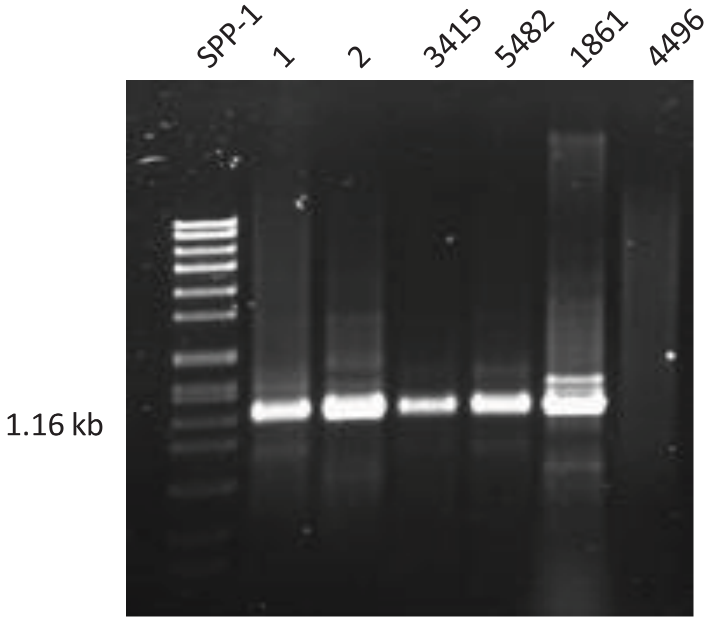


Figure 5.31 Verification of discrepancies in the strain 1 consensus sequence of *pspA*
PCR amplification was performed using primers RHPspAF and RHseqPspAR (Section 2.9.4), from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the SPP1 marker, as described in Section 2.9.1.

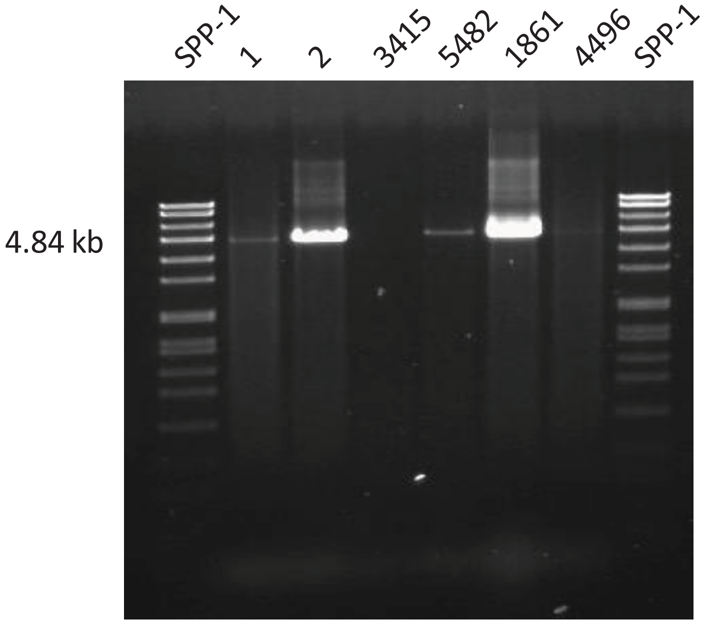


Figure 5.32 Verification of discrepancies in the strain 1 consensus sequence of *prtA*
PCR amplification was performed using primers RHPrtAF and RHPrtAR (Section 2.9.4), from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the SPP1 marker, as described in Section 2.9.1.

identified, which suggested that the discrepancies identified within the strain 1 consensus sequence of *prtA* were most likely due to sequence variation between *prtA* in strain 1 and *prtA* in P1031, rather than differences in the size of the gene. Similar to *pspA*, in future work it would be interesting to identify by direct sequencing whether any particular allele of *prtA* was associated with the invasive potential of the serotype 1 isolates.

Amplification was attempted across the assembly gap of *hylA* using primers RH316F and RH316R, which produced products 150-bp in size from strain 1 and 2 and products 400-bp in size from both the intermediately virulent and highly virulent strains (Figure 5.33). Therefore, a shorter version of *hylA* was associated with the non-invasive strains. In order to precisely characterise the deletion, sequencing was carried out on the *hylA* gene of strain 1, using RHhylAF, RHhylAR and RHseqhylAF$_{(2)}$ (Section 2.9.8). This confirmed that a 408-bp deletion was present in *hylA* of strain 1. A ClustalW alignment between *hylA* of strains 1 and 1861 showed that the deletion in strain 1 occurred in a region including both the catalytic domain and the cleft access-gate domain (Figure 5.34). However, the His399, Tyr408 and Asn349 residues required for catalytic activity (Jedrzejas, 2000) remained present in strain 1. In addition, a premature stop codon was introduced into strain 1 *hylA* by a nucleotide substitution. However, much of the sequence lost at the 3' end of *hylA* was not within a functional domain (Figure 5.34). Therefore, since it is not clear whether the 408-bp deletion in *hylA* would affect the activity of the translated protein, future work should compare the hyaluronidase activity of the non-invasive strains with the intermediately virulent and highly virulent strains to ascertain whether the activity of HylA correlates with virulence in these strains.

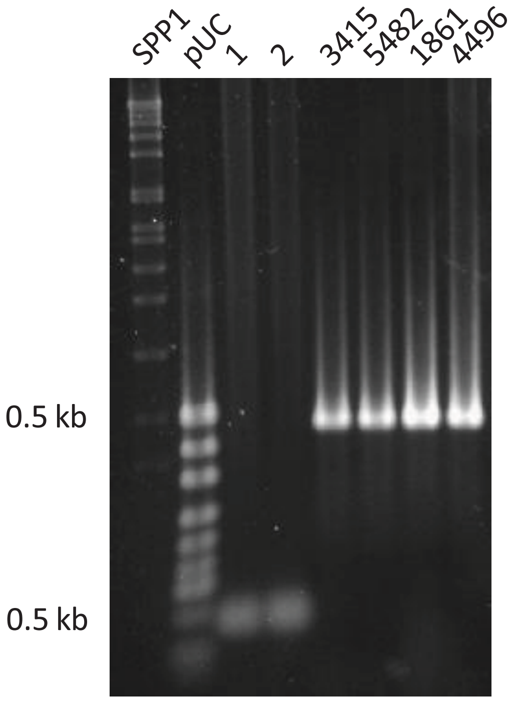


Figure 5.33 Verification of discrepancies in the strain 1 consensus sequence of *hlyA*
PCR amplification was performed using primers RH316F and RH316R (Section 2.9.4) from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the SPP1 and pUC markers, as described in Section 2.9.1.

Figure 5.34 ClustalW alignment of strain 1 and strain 1861 *hylA* amino acid sequence

The amino acid sequence was generated from nucleotide sequence using DNAMAN (Section 2.6.1). The signal sequence and domains of HylA are indicated by colour shading and labelled below the relevant region, as described by Rigden and Jedrzejas, (2003). The His399, Tyr408 and Asn349 catalytic residues are highlighted in orange, as described in Jedrzejas (2000).

```
             1                                                          60
Strain 1     MQTKTKKLIVSLSSLVLSGFLLNHYMTVGAEETTTNTIQQSQKEVQYQQRDTKNLVENGD
Strain 1861  MQTKTKKLIVSLSSLVLSGFLLNHYMTVGAEETTTNTIQQSQKEVQYQQRDTKNLVENGD
             ************************************************************
```

Signal Peptide

```
             61                                                        120
Strain 1     FGQTEDGSSPWTGSKAQGWSAWVDQKNS-ADASTRVIEAKDGAITISSPEKLRAALHRMV
Strain 1861  FGQTEDGSSSWTGSKAQGWSAWVDQKNSSADASTRVIEAKDGAITISSHEKLRAALHRMV
             *********.****************** ******************* ***********
```

Carbohydrate-binding domain

```
             121                                                       180
Strain 1     PIEAKKKYKLRFKIKTDNKVGIAKVRIIEESGKDKRLWNSATTSGTKDWQTIEADYSPTL
Strain 1861  PIEVKKKYKLRFKIKTDNKVGIAKVRIIEESGKDKRLWNSATTSGTKDWQTIEADYSPTL
             ***.********************************************************
```

```
             181                                                       240
Strain 1     DVDKIKLELFYETGTGTVSFKDIELVEVADQLSEDSQTDKQLEEKIDLPIGKKHVFSLAD
Strain 1861  DVDKIKLELFYETGTGTVSFKDIELVEVAAQLSEDSQTDKQLEEKIDLPIGKKHVFSLAD
             ***************************** ******************************
```

```
             241                                                       300
Strain 1     YTYKVENPDVASVKNGILEPLKEGTTNVIVSKDGKEVKKIPLKILASAKDAYTDRLDDWN
Strain 1861  YTYKVENPDVASVKNGILEPLKGGTTNVIVSKDGKEVKKIPLKILASVKDTYTDRLDDWN
             ********************** ************************.**:*********
```

Spacer domain

```
             301                                                       360
Strain 1     GIIAGNQYYDSKNEQMAKLNQELEGKVADSLSSISSQADRIYLWEKFSNYKTSANLTATY
Strain 1861  GIIAGNQYYDSKNEQMAKLNQELEGKVADSLSSISSQADRTYLWEKFSNYKTSANLTATY
             **************************************** *******************
```

```
            361                                                                                                             420
Strain 1    R K L E E M A K Q V T N P S S R Y Y K D E T V V R T V R D S M E W M H K H V Y N S E K S I V G N W W D Y E I G T P R A I
Strain 1861 R K L E E M A K Q V T N P S S R Y Y Q D E T V V R T V R D S M E W M H K H V Y N S E K S I V G N W W D Y E I G T P R A I
            * * * * * * * * * * * * * * * * * * : * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

            421                                                                                                             480
Strain 1    N N T L S L M K E Y F S D E E I K K Y T D V I E K F V P D P E H F R K T T D N P F K A L G G N L V D M G R V K V I A G L
Strain 1861 N N T L S L M K E Y F S D E E I K K Y T D V I E K F V P D P E H F R K T T D N P F K A L G G N L V D M G R V K V I A G L
            * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

                                                      Catalytic domain
            481                                                                                                             540
Strain 1    L R K D D Q E I S S T I R S I E Q V F K L V D Q G E G F Y Q D G S Y I D H T N V A Y T G A Y G N V L I D G L S Q L L P V
Strain 1861 L R K D D Q E I S S T I R S I E Q V F K L V D Q G E G F Y Q D G S Y I D H T N V A Y T G A Y G N V L I D G L S Q L L P V
            * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

            541                                                                                                             600
Strain 1    I Q K T K N P I D K D K M Q T M Y H W I D K S F A P L L V N G E L M D M S R G R S I S R A N S - - - - - - - - - - - - -
Strain 1861 I Q K T K N P I D K D K M Q T M Y H W I D K S F A P L L V N G E L M D M S R G R S I S R A N S E G H V A A V E V L R G I
            * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

            601                                                                                                             660
Strain 1    - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Strain 1861 H R I A D M S E G E T K Q R L Q S L V K T I V Q S D S Y Y D V F K N L K T Y K D I S L M Q S L L S D A G V A S V P R P S

            661                                                                                                             720
Strain 1    - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Strain 1861 Y L S A F N K M D K T A M Y N A E K G F G F G L S L F S S R T L N Y E H M N K E N K R G W Y T S D G M F Y L Y N G D L S
```

```
             721                                                        780
Strain 1     ---DGYWPTVNPYKMPGTTETDAKRADSDTGKVLPSAFVGTSKLDDANATATMDFTNWNQ
Strain 1861  HYSDGYWPTVNPYKMPGTTETDAKRADSDTGKVLPSAFVGTSKLDDANATATMDFTNWNQ
                ********************************************************* 

             781                                                        840
Strain 1     TLTAHKSWFMLKDKIAFLGSNIQNTSTDTAATTIDQRKLESSNPYKVYVNDKEASLTEQE
Strain 1861  TLTAHKSWFMLKDKIAFLGSNIQNTSTDTAATTIDQRKLESSNPYKVYVNDKEASLTEQE
             ************************************************************
```

Cleft access gate domain

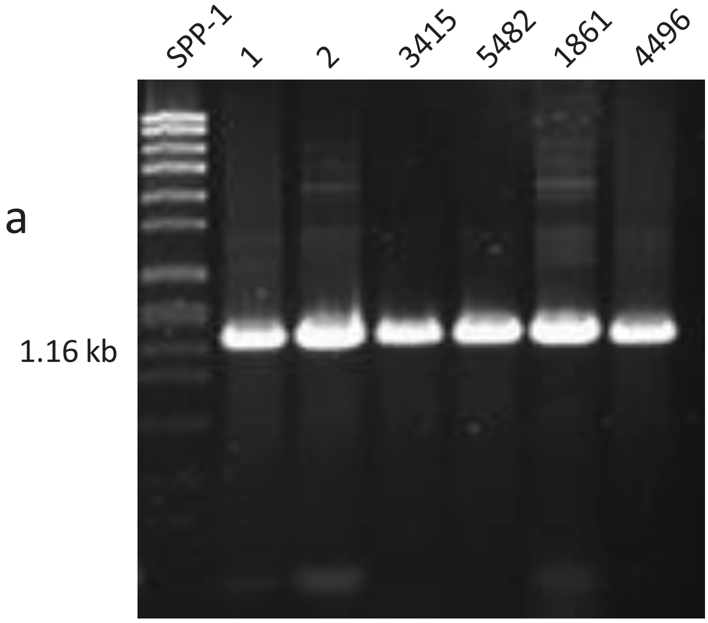
```
             841                                                        900
Strain 1     KDYPETQSVFLESSDSKKNIGYFFFKKSSISMSKALQKGAWKGINEGQSDKEVENEFLTI
Strain 1861  KDYPETQSVFLESSDSKKNIGYFFFKKSSISMSKALQKGAWKDINEGQSDKEVENEFLTI
             ******************************************.*****************

             901                                                        960
Strain 1     SQAHK-------------------------------------------------------
Strain 1861  SQAHKQNGDSYGYMLIPNVGRATFNQMIKKLESSLIENNETLQSVYDAKQGVWGIVKYDD
             *****

             961                                                       1020
Strain 1     ------------------------------------------------------------
Strain 1861  SVSTISNQFQVLKRGVYTIRKEGDEYKIAYYNPETQESAPDQEVFKKLEQAAQPQVQNSK

             1021                                         1067
Strain 1     -----------------------------------------------
Strain 1861  EKEKSEEEKNHSDQKNLPQTGEGQSILASLGFLLLGAFYLFRRGKNN
```

Amplification across the discrepancies identified within *phtB* and *phtD* were attempted in all six strains using RHPhtD(a)F – RHPhtD(a)R and RHPhtD(b)F – RHPhtD(b)R for *phtB* and *phtD*, respectively (Figure 5.35). For both *phtB* and *phtD*, products of the same size were produced in all strains, suggesting that assembly gaps had probably occurred within the gene due to sequence variability between the genes in strain 1 compared to P1031.

5.6 Comparison of *in vivo* expression of key 1861 genes between different niches of the mouse

As described in Section 5.5, a number of regions of the strain 1861 genome were also present in the genome of strain 4496, but absent from the genomes of the non-invasive and intermediately virulent strains. Genes of particular interest included *pblB* (*SPP_0075*) and the phage-associated endolysin (*SPP_0083*) of region 1, the putative sodium-dependent transporter (*SPP_0747*) of region 4, *zmpD* (*SPP_1141*) of region 7, the high-affinity iron/lead permease (*SPP_1340*) of region 8 and the glycoside hydrolase family protein (*SPP_1353*) of region 9. *SPP_1353* of region 9 was chosen to reflect the expression of the major component of the region, which includes the ABC transporter. However, due to potential non-specific binding to other ABC transporter genes, *SPP_1353* was chosen as it was likely to be co-transcribed with the transporter due to their close proximity and the absence of any obvious stem-loop structures between *SPP_1353* and *SPP_1354*.

As described in Chapter 4, *in vivo* expression analysis can provide important information about the expression requirements of target genes during different stages of pneumococcal pathogenesis. Therefore, *in vivo* expression comparisons using real-time RT-PCR (Section 2.12.6) were performed on the above genes in order to detect niche-



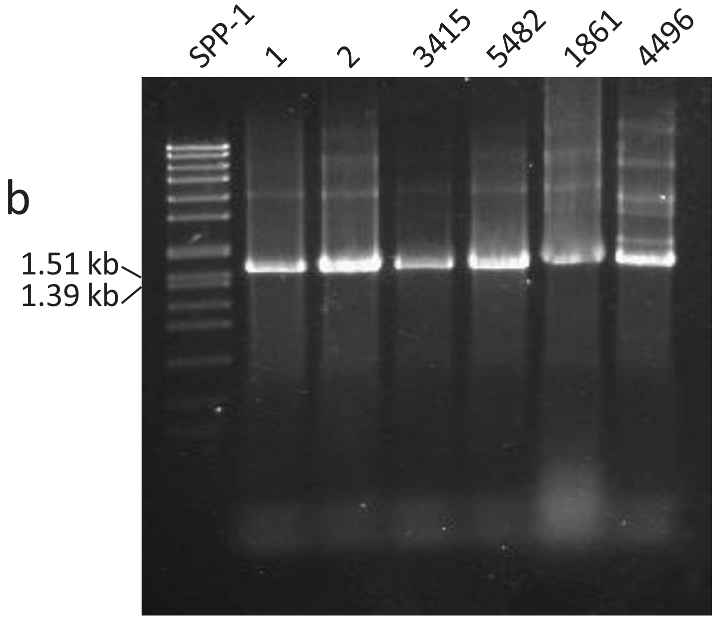


Figure 5.35 Verification of discrepancies in the strain 1 consensus sequence of *phtB* and *phtD*

PCR amplification was performed using primers (a) RHPhtD(a)F and RHPhtD(a)R and primers RHPhtD(b)F and RHPhtD(b)R (Section 2.9.4), from chromosomal DNA prepared from strains 1, 2, 3415, 5482, 1861 and 4496 (Section 2.9.2). PCR products were visualised on a 0.8% agarose gel using the 1kb plus marker, as described in Section 2.9.1.

dependent expression changes in strains 1861 and 4496. The same RNA used for the expression analysis of the PPI-1 genes (Section 4.3.2), was used as the template for the real-time RT-PCR analysis of this section. The primers used are shown in Table 5.11 and were designed as described in Section 2.3.2. Whilst *zmpD* was indeed a gene of interest, numerous attempts to design primers that could amplify a clean product failed, probably due to the multiple repeats within the gene. Therefore, *zmpD* expression was not analysed in this study.

Table 5.11 Primers used for real-time RT-PCR of potential virulence genes identified by genomic sequencing

Gene	Primers*
PblB (*SPP_0075*)	RHrtPblBF/RHrtPblBR
Phage-associated endolysin (*SPP_0083*)	RHrt0083F/RHrt0083R
Na^+-dependent transporter (*SPP_0747*)	RHrt0747F/ Hrt0747R
Fe^{2+}/Pb^{2+} permease (*SPP_1340*)	RHrt1340F/RHrt1340R
Glycoside hydrolase family protein (*SPP_1353*)	RHrt1353F/RHrt1353R

*Primer sequences are included in Table 2.3.

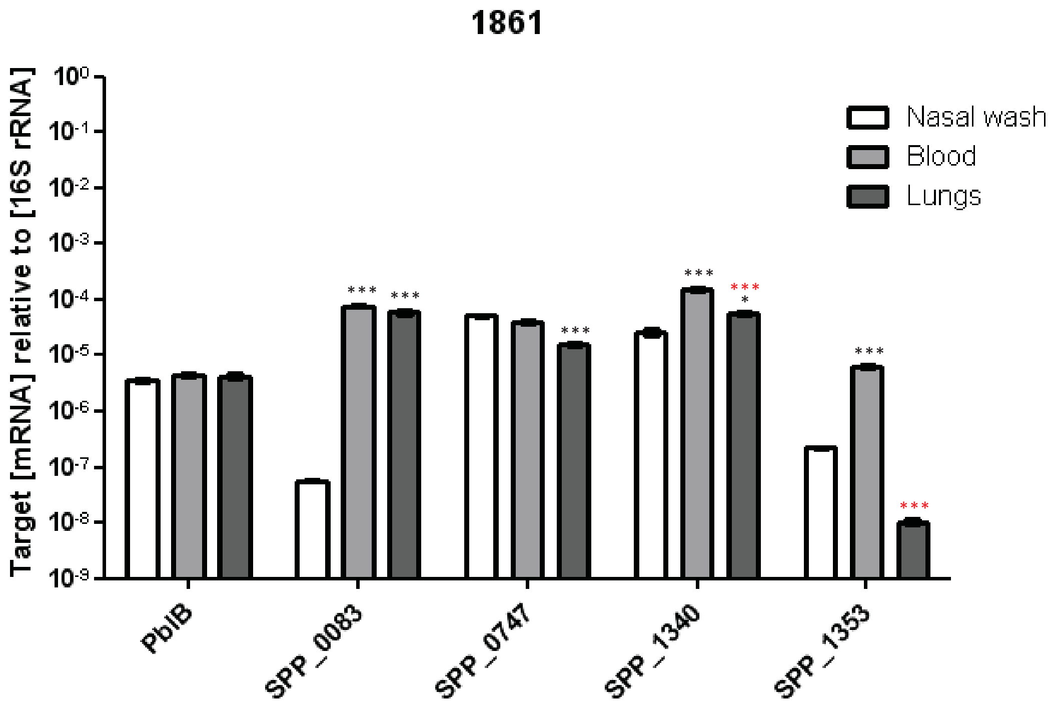
Five reactions were performed for each gene, with three reactions containing template and two no-template controls. In addition, melt-curve analysis was performed to ensure that only a single product was amplified in each reaction and to ensure that the melt temperature of each amplified product was the same as that amplified from RNA from an *in vitro* broth culture that had been checked by agarose gel electrophoresis (Section 2.9.1). For each reaction, a target of 150 – 300 ng of template RNA was desired. However, the exact quantity of template varied between samples due to differences in the quantity of contaminating eukaryotic RNA, and due to differences in the efficiency of linear RNA amplification. Differences in the amount of RNA between samples were normalised against 16S rRNA that was measured in each sample, using $RH16SF_{(3)}$ and $RH16SR_{(3)}$ (Table 2.3). In some instances genes were not detected at cycles significantly different from that of the no-template controls and were deemed to be below the limit of detection. Statistical significance between the amounts of target mRNA relative to 16S rRNA was calculated using the two-tailed unpaired *t*-test, where *P*< 0.05 was considered statistically significant.

Figure 5.36 presents the mean expression of each target gene relative to 16S rRNA in each niche for each strain. The fold change in expression of individual genes between niches is presented in Tables 5.12, 5.13 and 5.14. The expression of *pblB* was not significantly different between the blood and the surface of the nasopharynx (Table 5.12). However, expression of the phage-associated endolysin was significantly greater in the blood than at the nasopharyngeal surface (*P*<0.001), which suggested that there might be a greater requirement for this gene in the blood than at the nasopharyngeal surface. Greater expression of the endolysin suggests that phage-mediated lysis may occur more frequently in the blood than at the surface of the nasopharynx. Expression of the sodium-dependent transporter was not significantly different between the blood and the surface of the nasopharynx. However, expression of the high-affinity iron/lead permease was significantly greater in the blood than at the nasopharyngeal surface of both strain 1861- and 4496-infected mice (*P*<0.001). Similarly, the expression of the glycoside hydrolase family protein was greater in the blood than at the nasopharyngeal surface (*P*<0.001). Therefore, it is possible that expression of the ABC transporter upstream of this enzyme was also greater in the blood than at the nasopharyngeal surface.

PblB expression was not significantly different between the nasopharyngeal surface and the lungs of strain-1861 infected mice. However, there was a significant (2.43-fold) reduction in the lungs relative to the surface of the nasopharynx in 4496-infected mice (*P*<0.01) (Table 5.13). Significantly greater expression of the phage-

Figure 5.36 Relative expression of select genes absent from the lineage A strains and present in strains 1861 and 4496 in the nasopharynx, lungs and blood compared to 16S rRNA

The amount of target mRNA relative to 16S rRNA in the nasal wash, blood and lung samples of 1861-infected and 4496-infected mice was determined by real time RT-PCR (Section 2.12.6). Error bars indicate the standard deviation of triplicate reactions for each gene per niche. Statistical significance between the relative expression of individual genes in different niches was determined by unpaired *t*-test (*, *P*<0.05; **, *P*<0.01; ***, *P*<0.001). Black asterisks indicate comparison with nasal wash and red with blood.


1861
Target [mRNA] relative to [16S rRNA]
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
Nasal wash
Blood
Lungs
***
***
***
***
***
*
***
***
PblB
SPP_0083
SPP_0747
SPP_1340
SPP_1353

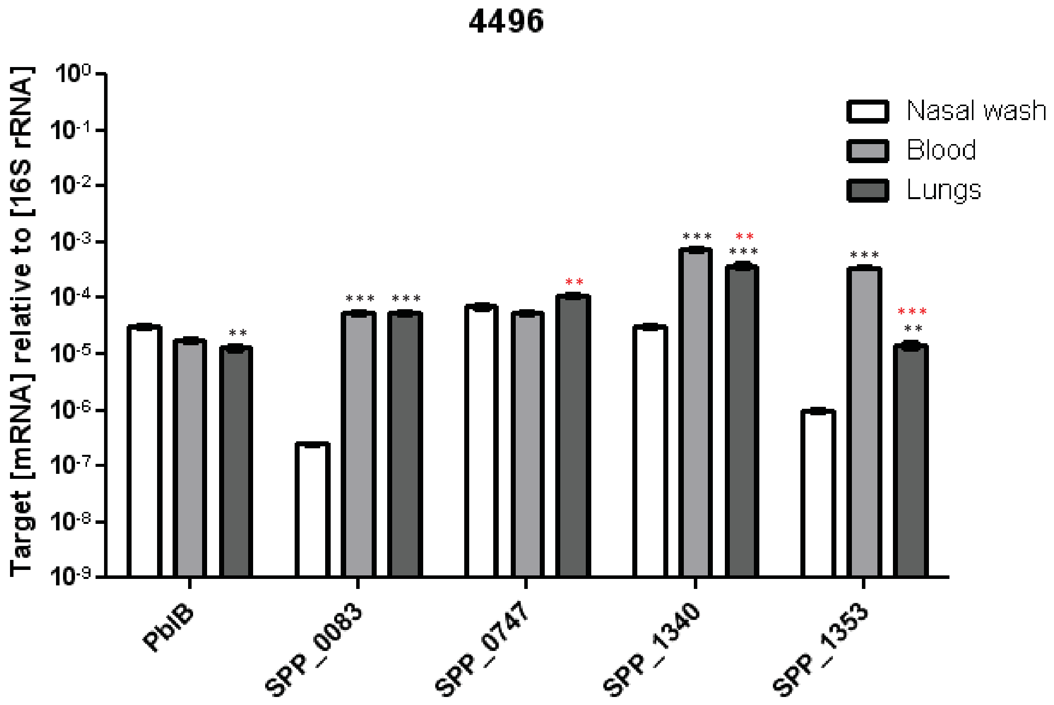

4496
Target [mRNA] relative to [16S rRNA]
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
Nasal wash
Blood
Lungs
**
***
***
**
***
**
***
***
***
**
PblB
SPP_0083
SPP_0747
SPP_1340
SPP_1353

Table 5.12 Relative expression of PPI-1 genes in the blood versus the nasal wash

ORF	1861		4496	
	Fold change	**Direction**	**Fold change**	**Direction**
PblB	1.22^{ns}	up	1.79^{ns}	down
SPP_0083	>1311.20*^{c}	up	>221.32*^{c}	up
SPP_0747	1.28^{ns}	down	1.28^{ns}	down
SPP_1340	5.92^{c}	up	24.36^{c}	up
SPP_1353	>29.18*^{c}	up	>369.65*^{c}	up

*Indicates target was not detected from nasal wash-derived RNA
Results of statistical analysis: ns. not significant (includes values <2); *a*, *P*<0.05; *b*, *P*<0.01; *c*, *P*<0.001
The relative fold difference was calculated using the ΔCt method, as described in Section 2.12.7.

Table 5.13 Relative expression of PPI-1 genes in the lungs versus the nasal wash

ORF	1861		4496	
	Fold change	**Direction**	**Fold change**	**Direction**
PblB	1.17^{ns}	up	2.43^{b}	down
SPP_0083	>1028.74*^{c}	up	>220.30*^{c}	up
SPP_0747	3.15^{c}	down	1.54^{ns}	up
SPP_1340	2.17^{ns}	up	12.10^{ns}	up
SPP_1353	-$^{\#}$	-$^{\#}$	>14.66*^{b}	up

*Indicates target was not detected from nasal wash-derived RNA
#Indicates transcript was below the limit of detection in both samples
Results of statistical analysis: ns. not significant (includes values <2); *a*, *P*<0.05; *b*, *P*<0.01; *c*, *P*<0.001
The relative fold difference was calculated using the ΔCt method, as described in Section 2.12.7.

Table 5.14 Relative expression of PPI-1 genes in the blood versus the lungs

ORF	1861		4496	
	Fold change	**Direction**	**Fold change**	**Direction**
PblB	1.04^{ns}	up	1.36^{ns}	up
SPP_0083	1.27^{ns}	up	1.00^{ns}	-
SPP_0747	2.47^{b}	up	1.97^{ns}	down
SPP_1340	2.73^{a}	up	2.01^{c}	up
SPP_1353	>624.55*^{c}	up	25.22^{c}	up

*Indicates target was not detected from lung-derived RNA
Results of statistical analysis: ns. not significant (includes values <2); *a*, *P*<0.05; *b*, *P*<0.01; *c*, *P*<0.001
The relative fold difference was calculated using the ΔCt method, as described in Section 2.12.7.

associated endolysin was detected in the lungs than at the nasopharyngeal surface (*P*<0.001), which suggests that more frequent phage-mediated lysis may also occur in the lungs than at the nasopharyngeal surface. Expression of the sodium-dependent transporter (*SPP_0747*) was significantly greater at the nasopharyngeal surface than in the lungs of strain 1861-infected mice (*P*<0.001). However, expression of the same gene was not significantly different between the two niches in strain 4496-infected mice. Therefore, similar to *pblB*, it was not possible to confirm that the sodium-dependent transporter exhibits niche-specific expression between the nasopharyngeal surface and the lungs. The high-affinity iron/lead permease exhibited significantly greater expression in the lungs than at the nasopharyngeal surface of both strain 1861- (*P*<0.05) and 4496- (*P*<0.001) infected mice, which suggests that the product of this gene might be more important in the former niche. The expression of the glycoside hydrolase family protein gene was below the limit of detection in both the nasal wash- and lung-derived RNA samples of strain 1861-infected mice. However, expression of this gene was significantly greater in the lungs than at the nasopharyngeal surface of strain 4496-infected mice.

Expression of *pblB* and the phage-associated endolysin was not significantly different between the blood and lungs of either strain 1861- or 4496-infected mice (Table 5.14). The expression of the sodium-dependent transporter was significantly greater in the blood than the lungs in strain 1861-infected mice (*P*<0.05), but not in strain 4496-infected mice. In contrast, significantly greater expression of the high-affinity iron/lead permease was detected in the blood than the lungs in both strain 1861- (*P*<0.05) and 4496- (*P*<0.001) infected mice. The expression of the glycoside hydrolase was significantly greater in the blood than the lungs of both strain 1861- and 4496-

infected mice (*P*<0.001), which suggests that the expression of the ABC transporter might also exhibit greater expression in the former niche.

In summary, *pblB* did not exhibit niche-dependent changes in expression, whereas the endolysin encoded downstream of *pblB* exhibited significantly greater expression in both the blood and lungs than at the nasopharyngeal surface. The significance of the endolysin expression was highlighted in Mitchell *et al*., (2007), where it was shown that endolysin-deficient mutants of *S. mitis* exhibited significantly reduced ability to bind human platelets, which was correlated with a significant reduction in the surface expression of PblB. Interestingly, this phenotype was reversed by the addition of exogenous PblB to the culture media (Mitchell *et al*., 2007). Therefore, in the case of strains 1861 and 4496, the greater expression of the endolysin in the blood and lungs could contribute to increased PblB surface expression and thus increase adherence to human platelets and other host surfaces. Generally the sodium-dependent transporter (*SPP_0747*) did not exhibit niche-dependent changes in expression, which suggests that the requirement for this transporter does not appear to vary during disease progression. In contrast, niche-dependent changes in expression of the high-affinity iron/lead transporter were detected, with the greatest expression detected in the blood, followed by the lungs and finally the nasopharyngeal surface. The expression of the glycoside hydrolase and possibly the ABC transporter of region 9 were greatest in the blood when compared to the lungs and nasopharyngeal surface. Whilst gene expression analysis of a greater number of genes within this region is required, it appears that at least a component of this region exhibits greater expression in the blood than in either the lungs or on the nasopharyngeal mucosa. Therefore, it might be possible that strains 1861 and 4496 exhibit greater sialic acid degradative properties in the blood that could contribute to invasive potential.

5.7 Discussion

In previous studies, comparative genomics using technology such as CGH had been performed in order to identify ARs of the pneumococcal genome that are associated with increased invasive potential (Blomberg *et al*., 2009; Obert *et al*., 2006; Silva *et al*., 2006; Bruckner *et al*., 2004). Whilst one study suggested that ARs such as the *PsrP-secAY2A2* region are associated with strains of high invasive potential (Obert *et al*., 2006), another was unable to identify an AR that was responsible for the high-attack rate of serotypes such as 1 and 7F (Blomberg *et al*., 2009). However, even though a certain amount of redundancy is likely to exist between different ARs in their respective contribution to invasive potential, comparisons between serotypes can be complicated by serotype differences in the protective properties of the capsule (Section 1.3.1.5). Therefore, the most valuable comparisons are made between strains of the same serotype in order to identify serotype-independent differences in the genome that are associated with specific virulence phenotypes. Subsequently, regions found to be responsible for serotype-independent differences in virulence can be investigated in multiple serotypes to ascertain whether their contribution to virulence is consistent.

Previous work has shown in both mice and humans that serotype 1 strains can vary considerably in virulence potential (Section 1.6.1; Smith-Vaughan H *et al*., 2009; Antonio *et al*., 2008; Nunes *et al*., 2008). Therefore, this chapter aimed to identify ARs in addition to the PPI-1 variable region that could contribute to the differences in invasive potential between non-invasive, intermediately virulent and highly virulent serotype 1 isolates.

5.7.1 Identification of virulence phenotype-associated genes by CGH

Initially, CGH was carried out between non-invasive (1 and 2), intermediately virulent (3415 and 5482) and highly virulent (1861 and 4496) serotype 1 clinical isolates. As described in Section 5.2.2, the *PsrP-secAY2A2* region (*SP_1758 – SP_1771*; AR 34) was found to be present in both intermediately virulent strains and strain 4496, whilst being absent from the non-invasive strains. Therefore, it is possible that the absence of *PsrP-secAY2A2* from the non-invasive strains might contribute to their inability to invade and survive in the blood. However, the highly virulent strain 1861 also lacks *PsrP-secAY2A2*, which suggests that the region is not required for the virulence of this strain. CGH identified three regions that were present only in the highly virulent strains and absent in the non-invasive and intermediately virulent strains. These regions included an ABC transporter (*SPR_1191 – SPR_1195*; AR 24) the PPI-1 variable region (*SP_1047 – SP_1053*; AR 22 [Chapters 3 & 4]) and a PTS with a number of enzymes thought to be involved in the uptake of ribulose (*SP_1615 – SP_1621*; AR 31) (Obert *et al*., 2006). However, closer analysis revealed that AR 31 was only present in strain 1861, and not in strain 4496. Therefore, the region is not required for virulence in strain 4496.

In summary, only the PPI-1 variable region and the ABC transporter of AR 24 were found to be consistently associated with heightened virulence using CGH.

5.7.2 Identification of genes associated with hypervirulent serotype 1 isolates by genomic sequencing

In order to comprehensively compare the genomes of the serotype 1 clinical isolates, next generation genome sequencing technology was utilised to sequence a representative non-invasive strain (strain 1) and a representative highly virulent strain

(strain 1861). Assembly of the sequenced genomes was performed against the serotype 1 (ST303) genome of P1031. As very few discrepancies were identified between the strain 1861 consensus sequence and the P1031 genome, the differences that were identified between the genomes of strain 1 and P1031 were generally representative of the differences between the genomes of strains 1 and 1861. PCR verification of strain 1 consensus sequence assembly gaps that were greater than 1-kb in size confirmed that a number of regions were associated with the heightened virulence. A group 1 pneumophage (Romero *et al*., 2009a) was present only in the highly virulent strains and encoded PblB and an endolysin. In addition, a region encoding a putative sodium-dependent transporter, a MerR family transcriptional regulator and a MutT/Nudix family protein were also present only in the highly virulent strains. ZmpD and a large conjugative transposon (Tn*5253*) were also found to be associated with heightened virulence. Given that other zinc metalloproteinases play roles in virulence (Section 1.5.3.2), examination of the contribution made by ZmpD to the invasive potential of strains 1861 and 4496 is warranted. A high-affinity iron/lead permease, a dyp-type peroxidase, an ABC-2 type transporter gene, a nod factor export ATP-binding protein and a transcriptional regulator of the ArsR family (*SPP_1781*) were also identified as associated with heightened virulence. However, very little is known about the function of many of these genes making it difficult to predict their role, if any, in the virulence of strains 1861 and 4496. Finally, a complex region of putative catabolic enzymes and an ABC transporter were associated with heightened virulence. BLAST and HHpred search results suggested that the region was involved in the metabolism of sialic acid. The ability of some strains, such as the highly virulent strains, to take advantage of specific sugar sources could provide a survival advantage and contribute to increased invasive potential.

Genomic sequencing also identified allelic differences between *iga*, *pspA*, *prtA*, *phtB* and *phtD* of strains 1 and 1861. However, direct sequencing of these genes in each strain would be required to show whether specific alleles are associated with invasive potential. In addition, a deletion was identified in *hylA* of only the non-invasive strains and not in either the intermediately virulent or highly virulent strains. Whilst it was difficult to predict whether the deletion in *hylA* would impact on the activity of the translated protein, reduced HylA activity could reduce the passage of the pneumococcus through host tissues (Section 1.3.1.1), thus reducing invasive potential.

5.7.3 Differential *in vivo* expression of key genes associated with heightened invasiveness

Key genes of regions that were associated with heightened virulence were selected for expression comparisons between the nasopharyngeal surface, the blood and the lungs. The expression of *pblB* remained unchanged between niches, which suggested that continuous production of PblB occurs at all stages of pathogenesis. However, the phage-associated endolysin, encoded within the same prophage as *pblB*, exhibited significantly greater expression in both the lungs and blood of infected mice. Given the role of the product of this gene in lysis of the host cell (reviewed in Young *et al*. [2000]), it might be possible that increased lysis by strains 1861 and 4496 in the blood and lungs could augment the inflammatory properties of LytA activity (Section 1.3.1.2). Interestingly, previous work in *S. mitis* showed that lysis of the bacterium was required for the surface expression of PblB (Mitchell *et al*., 2007). Therefore, greater expression of endoylsin in the blood and lungs compared to the nasopharyngeal surface suggests that upon reaching these niches the prophage encoded within strains 1861 and 4496 enters the lytic cycle in at least a subpopulation of cells, thus enabling the release of cytoplasmic PblB. Attachment of extracellular PblB to the cell wall of unlysed pneumococci may enhance the ability of the bacterium to adhere to host cell surfaces, such as platelets (Bensing *et al*., 2001; Mitchell *et al*., 2007). Furthermore, the ability of the pneumococcus to bind plasmin has been linked to the bacterium's ability to travel across epithelial surfaces (Attali *et al*., 2008). Therefore, these attributes suggest that PblB may have a role in the heightened virulence of strains 1861 and 4496. Since the expression of the sodium-dependent transporter was not found to change between niches the transporter may be required equally in all niches or its activity could be regulated by other mechanisms, such as substrate concentrations or post-translational regulation. In addition, regulation of the transporter's expression may no longer occur if the truncation predicted for *SPP_0750* in strain 1861 had prevented the translation of a functional protein. In contrast, the high-affinity iron/lead permease exhibited greater expression in the blood than the lungs, and was greater in both than at the nasopharyngeal surface. Therefore, if greater expression of this gene enables greater uptake or excretion of important metal ions, then the permease may provide a survival advantage for strains 1861 and 4496 in the blood and lungs. Finally, *in vivo* expression analysis showed that expression of the glycoside hydrolase of region 9 appeared to be greater in the blood than at the nasopharyngeal surface. Therefore, it is possible that the expression of the ABC transporter encoded immediately upstream of this enzyme is also greater in the blood than at the surface of the nasopharynx. Given that many of the enzymes that flank the ABC transporter are predicted to be involved in sialic acid degradation, it is possible that this region promotes the utilisation of sialic acid as an energy source in the blood, which could provide a survival advantage in this niche.

In summary, a number of ARs were identified in the genomes of hypervirulent serotype 1 isolates that were not present in the genomes of either non-invasive or intermediately virulent serotype 1 isolates. Interestingly, many of the ARs that were highlighted in this study had not been previously identified as ARs, due to their absence from the genomes of TIGR4 and R6. The ARs that were identified have potential roles ranging from the transport and metabolism of unknown host sugars to adherence and migration across epithelial surfaces. Future studies involving the mutagenesis of these genes will confirm the contribution of these ARs to the invasive potential of strains 1861 and 4496.

Chapter 6 – Competence in Serotype 1 Clinical Isolates

6.1 Introduction

As mentioned in Chapter 4, repeated attempts to genetically manipulate the serotype 1 isolates used in this study were unsuccessful. Inability to transform the highly virulent serotype 1 strains complicated attempts to compare the virulence of PPI-1 variants constructed in the serotype 1 background, and necessitated indirect studies in the D39 background (Chapter 4). Such difficulty has previously been encountered with some clinical isolates and the reasons for this phenomenon are not understood (Pozzi *et al*., 1996). In particular, it is not known whether the lack of transformability is primarily due to strain-strain variation in the optimum *in vitro* conditions, or due to defects in the competence system that render these strains non-transformable both *in vitro* and *in vivo*.

The method routinely used by our laboratory for genetic transformation of *S. pneumoniae* is outlined in Section 2.8.2 and is based on the method developed by Martin *et al*. (1995). This method involves culturing the relevant strain to mid-exponential phase in cCAT media (Section 2.8.1) before diluting ten-fold in fresh media with increased concentration of calcium and BSA (CTM) and growing to early exponential phase (A_{600} 0.2). An increase in calcium concentration facilitates the competence induction *in vitro* during exponential growth (Trombe *et al*., 1992). Subsequently, the culture is resuspended in fresh CTM with an elevated pH (pH 7.8), which reduces the culture density at which the competent state develops and increases the length of time for which the competent state exists (Chen & Morrison, 1987).

Genetic competence in the pneumococcus is regulated by a complex quorum sensing and gene expression network (Section 1.3.4). It has been shown that when the accumulation of CSP in the local extracellular environment reaches a critical concentration, the two-component signal transduction system encoded by genes *comD* and *comE* is activated, leading to expression of the early competence genes (Pestova *et al*., 1996; Havarstein *et al*., 1996; Alloing *et al*., 1998). A number of alleles of *comC* have been identified in pneumococci, of which *comC1* (CSP-1) is the most common, followed by CSP-2. By contrast, CSP alleles 3 – 6, are rarely found (Pozzi *et al*., 1996; Whatmore *et al*., 1999). In addition, four *comD* alleles have been characterised, of which 1, 3 and 4 have been shown to be preferentially activated by CSP-1, whilst allele 2 is preferentially activated by CSP-2 (Iannelli *et al*., 2005). Amino acid substitutions which impact on CSP specificity have been shown to be confined to 70 amino acids in the N-terminal region of ComD (Iannelli *et al*., 2005). Early competence genes are expressed within 5 minutes of CSP addition to an *in vitro* culture and include *comX*, *comW*, *comA*, *comB*, *comC* and *comD* (Peterson *et al*., 2004). Early competence genes are primarily involved in the regulation of competence and the synthesis and export of CSP. ComX is an alternative sigma factor that induces the expression of >60 late competence genes, which are involved in DNA uptake and the processing and recombination of environmentally acquired DNA into the chromosome (Peterson *et al*., 2004; Campbell *et al*., 1998; Pestova & Morrison, 1998). However, maximal expression of the late competence genes and efficient transformation requires ComW, which is thought to stabilise and promote accumulation of cytoplasmic ComX (Piotrowski *et al*., 2009; Sung & Morrison, 2005). ComW and ComX are the only known link between the quorum sensing system of ComC, ComD and ComE, and the expression of the late competence genes, which are required for successful transformation (Lee & Morrison, 1999; Luo & Morrison, 2004). Of the 15 early competence genes and the >60 late competence genes, approximately 23 are essential for transformation (Peterson *et al.*, 2004).

This chapter summarises the attempts to optimise the conditions for transformation of strains 1861 and 4496, and sought to identify possible causes for the lack of transformability of strains 1861 and 4496.

6.2 Sequencing *comC* and *comD* in 4496

In order to determine the CSP pherotype that is compatible with the *comD* alleles in strains 1861 and 4496, *comC* and *comD* were sequenced in these strains using primers 3, 4 and 5, as described in Whatmore *et al*. (1999). Final sequences were assembled using the sequence assembly tool in DNAMAN (Section 2.6.1). The predicted *comC* amino acid sequence of both strains 1861 and 4496 were aligned with the amino acid sequences of the six known alleles in *S. pneumoniae* (Figure 6.1) and showed that *comC* of strains 1861 and 4496 shares 100% sequence identity with *comC1*. Therefore, strains 1861 and 4496 encode CSP-1. Subsequently, the first 70 amino acids deduced from *comD* of strains 1861 and 4496 were aligned with the sequences of ComD1, ComD2, ComD3 and ComD4 (Figure 6.1). The alignment showed that ComD of strain 1861 shared 100% amino acid sequence identity with ComD1 and that ComD of strain 4496 shared 100% amino acid sequence identity with ComD4. Therefore, ComD1 of strain 1861 is most sensitive to CSP-1. However, whilst ComD4 has been shown to be less sensitive to CSP-1 than ComD1, the concentration of CSP-1 used in this study exceeded that required for maximal activation of ComD4 (Ianelli *et al*., 2005). Therefore, the sequence of *comD* is unlikely to have contributed to the inability to transform strains 1861 and 4496.

ComC

```
             1                                              48
comC1        MKNTVKLEQVALKEKDLQKIKGGEMRLSKF----FRDFI----LQRKK
Strain 1861  MKNTVKLEQVALKEKDLQKIKGGEMRLSKF----FRDFI----LQRKK
Strain 4496  MKNTVKLEQVALKEKDLQKIKGGEMRLSKF----FRDFI----LQRKK
comC2        MKNTVKLEQVALKEKDLQKIKGGEMRISRI----ILDFL----FLRKK
comC3        MKNTVKLEQVALKEKDLQKIKGGEMRKMNEKSFNIFNFFNFFNFFRRR
comC4        MKNTVKLEQVALKEKDLQKIKGGEMRKMNEKSFNIFNFFNF---FRRR
comC5        MKNTVKLEQVALKEKDLQKIKGGESRLPKI----LLDFL----FLRKK
comC6        MKNTVKLEQVALKEKDLQKIKGGEMRLPKI----LRDFI----FPRKK
             ************************ *  .     : :*:      *::
```

ComD

```
             1                                                                    70
comD1        MDLFGFGTVIVHFLIISHSYHFICKGQINRKELFVFGAYTLLTEIVFDFPLYILYLDGLGIERFLFPLGL
Strain 1861  MDLFGFGTVIVHFLIISHSYHFICKGQINRKELFVFGAYTLLTEIVFDFPLYILYLDGLGIERFLFPLGL
comD2        MDLLGFGTVIVHFLIISHSYRLICKGRINRKELYVFGAYTLLTEIVLEFSFYLLYLDKIGIERFLFPLGL
comD3        MDLLGFGTVIVHLLIISHNYHLICKGQINRKELFVFGAYTLLTEIVFDFPLYILYLDGLGIAIFLFPLGL
comD4        MDLFGFGTVIVHFLIISHSYHLICKGQINRKELFVFGAYTLLTEIVFDFPLYILYLDGLGIATFLFPLGL
Strain 4496  MDLFGFGTVIVHFLIISHSYHLICKGQINRKELFVFGAYTLLTEIVFDFPLYILYLDGLGIATFLFPLGL
             ***:********:*****.*::****:******:************::*.:*:**** :**  *******
```

Figure 6.1 The *comC* and *comD* deduced amino acid sequences of strains 1861 and 4496 aligned with previously published alleles

The amino acid sequences of *comC* and the first 70 amino acids of *comD* of strains 1861 and 4496 were aligned with the previously published alleles of *comC* (Whatmore *et al*., 1999) and *comD* (Iannelli *et al*., 2005), respectively. The amino acid sequences of *comC* and *comD* of strains 1861 and 4496 were determined following amplification and sequencing of the target genes, as described in Section 6.2. The amino acid sequences of *comC* sharing 100% sequence identity to strains 1861 and 4496 are highlighted in yellow, the amino acid sequences of *comD* sharing 100% sequence identity to strain 1861 are highlighted in blue and amino acid sequences sharing 100% sequence identity to strain 4496 are highlighted in green.

6.3 Optimisation of conditions for transformation of strain 4496 *in vitro*

As discussed in Section 6.1, the timing of competence is affected by the pH of CTM in the presence of CSP-1. However, despite trialling a range of incubation times with CSP-1, successful recombinants were not recovered from either strain 1861 or strain 4496. In addition, in case capsule obstructs DNA uptake in these strains, transparent-phase pneumococci were used throughout this study (Section 1.3.2). Therefore, aspects of the transformation methodology other than CSP incubation time and opacity phase were modified in order to identify the *in vitro* conditions required for competence induction in strain 4496.

6.3.1 Comparison between D39 and 4496 growth in cCAT

Calcium is an important environmental signal for the induction of competence (Trombe *et al*., 1992). However, whilst exponentially growing pneumococci tend to become competent in response to calcium, pneumococci within the stationary phase can instead undergo autolysis (Trombe *et al*., 1992). In the transformation methodology described in Section 2.8.2, an increase in calcium concentration is applied when the mid exponential-phase culture is diluted in CTM. Therefore, the phase of growth at which the culture is diluted in CTM might be important for the development of competence. Therefore, it was decided to compare the growth of strain 4496 and D39 in cCAT in order to ensure that strain 4496 did not enter the stationary phase at a lower culture density than D39, inadvertently triggering autolysis in response to the calcium in CTM. The experiment was carried out using D39 and 4496 frozen starter cultures that were prepared as described in Section 2.7. A_{600} readings at 30 min intervals over 6 h were recorded following the inoculation of fresh cCAT with an aliquot of the relevant frozen starter culture stock. Duplicate cultures were measured per strain and two independent experiments were performed (Figure 6.2). In both experiments, strain 4496 reached the stationary phase at a higher culture density than D39. Since strain 4496 consistently entered the stationary phase at A_{600} 1.3, it seems unlikely that at A_{600} 0.5 (Section 2.8.2) strain 4496 would undergo autolysis in a growth phase-dependent response to calcium.

6.3.2 The effect of culture density in CTM on transformation efficiency of D39 and strain 4496

Previous work has shown that competence develops spontaneously in a culture density-dependent fashion (Tomasz & Hotchkiss, 1964). However, the methodology for transformation of *S. pneumoniae* described in Section 2.8.2, assumes that a competent state will develop at a A_{600} 0.2. Since it is possible that strain 4496 could develop competence at a different culture density, the transformation efficiency of strain 4496 was compared between a range of CTM culture densities. In addition, the impact of CTM culture density on the transformation efficiency of D39 was also assessed. In the first experiment, the culture density was determined at 0, 40, 50, 60, 70, 80, 90, 100, 110 and 120 min following dilution into CTM (Figure 6.3a). These time points were chosen as previous experience had shown that A_{600} 0.2 would be reached by D39 following 60 to 80 min incubation in CTM. At each time point the remainder of the transformation protocol was performed to assess transformation efficiency at each culture density. Chromosomal DNA from the multi-drug resistant strain DP1617 was used as the donor DNA in these experiments. However, following the experiment, antibiotic resistant recombinants were not recovered from strain 4496, which suggested that the culture densities that were used did not affect the ability of the strain to undergo transformation. Unexpectedly, D39 recombinants were recovered from all time points and with no significant differences in transformation efficiency. However, despite the fact that the CTM culture density did not appear to affect D39 transformation efficiency, it was decided to repeat the experiment with earlier time points due to the early time

Figure 6.2 Growth of strain D39 and strain 4496 in cCAT

Duplicate cultures of D39 and strain 4496 were grown from frozen starter culture stocks (Section 2.7) in 20 ml of pre-warmed cCAT. Mean A_{600} readings from duplicate cultures were taken for each strain at 30 min intervals starting at 0 until 6 h. Two independent experiments were undertaken and are shown as (a) and (b). Error bars represent the standard error of the mean (SEM) of duplicate readings at each time point.

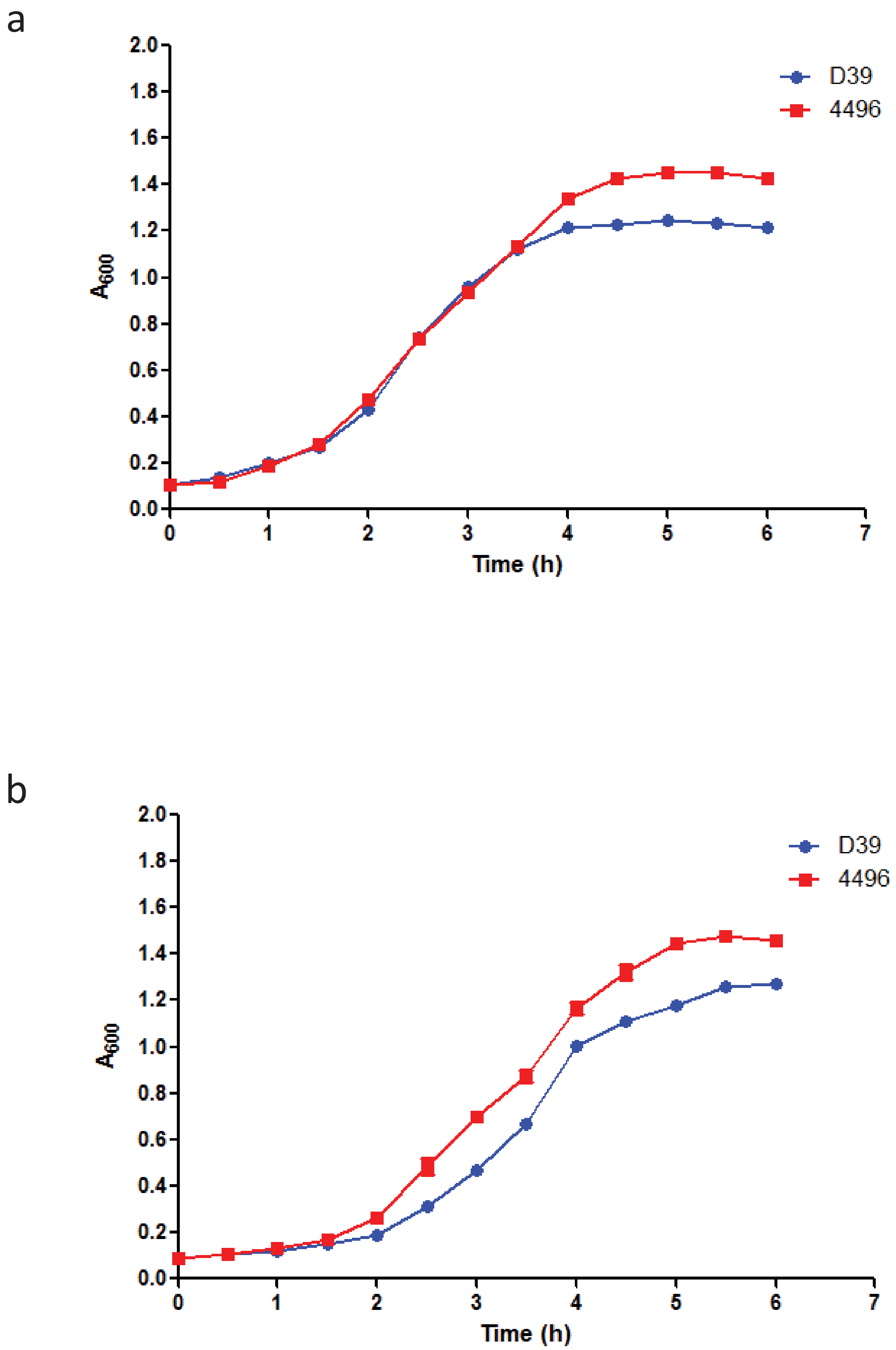
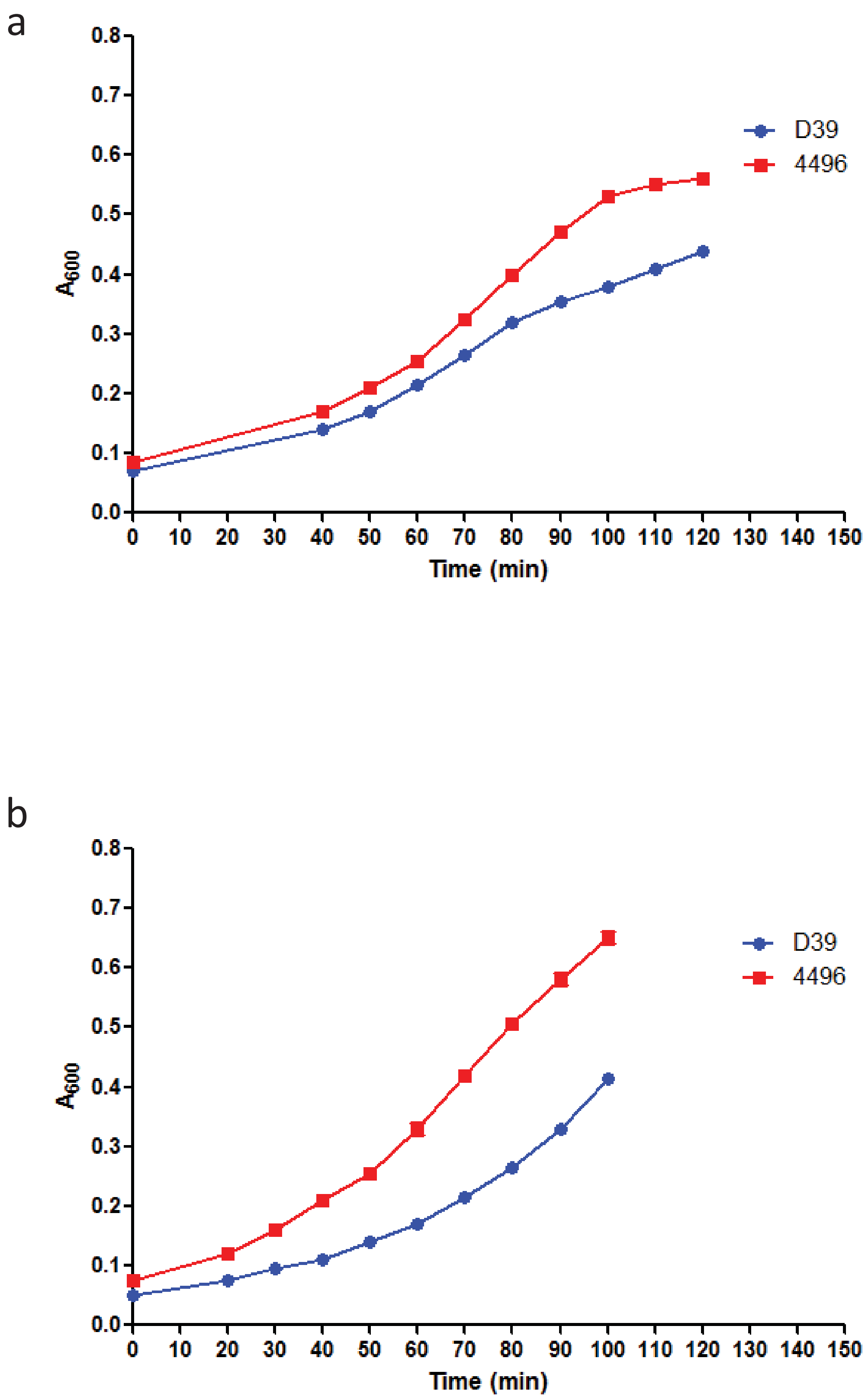

a
A600
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
Time (h)
D39
4496
b
A600
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
Time (h)
D39
4496

Figure 6.3 Growth of strains D39 and 4496 in CTM

Duplicate cultures of D39 and strain 4496 were grown from frozen starter culture stocks in cCAT (Section 2.7) to mid-exponential phase (A_{600} 0.5) before diluting tenfold in CTM. Mean A_{600} readings were taken at (a) 0, 40, 50, 60, 70, 80, 90, 100, and 110 min incubation in CTM (experiment 1) and at (b) 0, 20, 30, 40, 50, 60, 70, 80, 90, 100 min incubation in CTM (experiment 2). The transformation protocol was subsequently performed on each time point aliquot. Transformation mixes were incubated for 10 min with CSP-1 before adding chromosomal DNA and following the remainder of the protocol (Section 2.8.2). Error bars represent the SEM of duplicate readings at each time point.


a
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
A_{600}
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
Time (min)
D39
4496
b
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
A_{600}
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
Time (min)
D39
4496

point at which A_{600} 0.2 was reached by strain 4496 (Figure 6.3a). Therefore, the experiment was repeated with time points at 0, 20, 30, 40, 50, 60, 70, 80, 90 and 100 min in CTM (Figure 6.3b). However, similar to the first experiment, recombinants of strain 4496 were not recovered at any time point, and no difference in the transformation efficiency of D39 was detected between time points. It is also worth noting that A_{600} readings of D39 growth in CTM were consistently lower than 4496 in both experiments. However, it is not clear whether this difference affects transformability.

In summary, the culture density of CTM did not have a detectable impact on the transformation efficiency of either D39 or strain 4496. Therefore, the culture density used in the transformation attempts of strain 4496 was unlikely to be responsible for the lack of transformability.

6.3.3 The effect of different DNA donors on the transformation efficiency of strain 4496

The acquisition, processing and recombination of extracellular DNA into the chromosome is a complex process that in strain 4496 could potentially be defective at a number of different stages (reviewed by Claverys *et al*., 2009). Initial attempts at transformation of strain 4496 were primarily carried out using PCR product as the source of transforming DNA. These PCR products consisted of an antibiotic resistance cassette flanked by homologous DNA to allow recombination into the host chromosome. Therefore, in order to test whether the inability to transform strain 4496 was due to a defect in recombination and not DNA uptake, transformations were attempted on strain 4496 and D39 competent cells (Section 2.8.1), using plasmid DNA (pVA891). The plasmid pVA891 has been shown to be maintained in pneumococcal strains such as D39. Following multiple attempts, successful transformants were recovered only from D39 and not 4496. Therefore, it appears that strain 4496 was either unable to take up the plasmid DNA from the media, or was unable to maintain replication of this plasmid.

6.3.4 The effect of alternative media on the transformation efficiency of strain 4496

Following failed transformation attempts using CAT as the base medium, CAT was replaced in a series of transformation attempts with either THY or TSB. However, neither media enabled the transformation of strains 1861 or 4496. In addition, previous work has shown that the expression of competence genes is increased in pneumococci in the sessile state (Oggioni *et al*., 2006). Therefore, it was decided to attempt the transformation of strain 1861 and 4496 on solid media using the method described in Iannelli and Pozzi (2004). However, recombinants were not recovered from either strain 1861 or 4496 using such media. D39 was successfully transformed using all media tested in this section.

6.4 Alternative methods for genetic manipulation of serotype 1 isolates

Electrotransformation is a method commonly used to introduce recombinant DNA into a wide variety of bacteria. Whilst electrotransformation has been shown to be difficult in *S. pneumoniae* (Lefrancois & Sicard, 1997; Lefrancois *et al*., 1998), it was attempted in this study on strains 1861 and 4496 in order to overcome potential defects in the competence system of these strains. D39 was used as a control strain. Electrotransformation was performed with plasmid DNA (pVA891) and chromosomal DNA (DP1617) using a method loosely based on that described in Lefrancois and

Sicard (1997) (Section 2.8.3). However, electrotransformation was unsuccessful in both strains 1861 and 4496, and the D39 control.

Subsequently, it was decided to attempt transformation of strains 1, 4496 and D39 in the presence of Lipofectamine^{TM} 2000 reagent (Invitrogen). Lipofectamine^{TM} is routinely used for the transfection of eukaryotic cell lines and it was thought that the reagent might facilitate the uptake of DNA by strains D39, 1 and 4496. Transformation of strains 4496, 1 and D39 was performed in the presence of CSP-1, as described in Section 2.8.2. However, the protocol was modified by incubating the transforming DNA with Lipofectamine^{TM} at DNA:Lipofectamine ratios ranging from 1:0.5 to 1:5, according to the manufacturer's instructions. In addition, transformations lacking Lipofectamine^{TM} were also performed as a control. However, neither strain 1 nor 4496 transformants were recovered. Furthermore, treatment with Lipofectamine^{TM} had no effect on the transformation efficiency of D39.

6.5 Expression of key competence genes in the presence and absence of CSP-1

As discussed in Section 6.1, competence genes can be categorised as either early competence genes involved in competence regulation or late competence genes involved in DNA uptake and recombination. Therefore, expression of the key competence regulatory genes *comD*, *comX* and *comW* was compared between D39 and strain 4496 in the presence and absence of CSP-1. In addition, late competence genes involved in DNA uptake (*cglA*) and recombination (*coiA*) were selected to compare the expression of late competence genes between strains 4496 and D39.

The transformation protocol was performed on each strain in the presence and absence of CSP-1. The experiment was performed on two independently prepared batches of competent cells per strain (Section 2.8.1). RNA was prepared for gene expression analysis from cells immediately prior to adding the donor DNA (Section 2.12.1). The remainder of the transformation mixes were used to continue the transformation protocol (Section 2.8.2) to confirm whether or not transformation was successful in each case.

The relative expression of *comD*, *comX*, *comW*, *cglA* and *coiA* between samples was determined by real time RT-PCR using the ΔCt method, as described in Section 2.12.7. Approximately 100 ng of RNA was used as template for each reaction. However, differences in the actual amount of RNA in each sample were normalised against the amount of 16S rRNA, using $\text{RH16SF}_{(3)}$ and $\text{RH16SR}_{(3)}$ (Table 2.3). Five reactions were performed for each gene, with three reactions containing template and two lacking template. Melt-curve analysis was performed to ensure that only a single product was amplified in each reaction. Statistical significance between the relative amounts of target mRNA in different samples of the same strain was calculated by comparing the mean amount of target mRNA relative to 16S rRNA in each niche using the two-tailed unpaired *t*-test, where $P< 0.05$ was considered statistically significant.

Primers used to amplify the target genes were designed from the strain 1861 consensus sequence generated in Section 5.3 and the D39 genomic sequence using the criteria described in Section 2.3.2. The primers used for each target gene are indicated in Table 6.1.

Table 6.1 Primers used for real-time RT-PCR of key competence genes

Gene	Primers*
comD	RHrtcomDF/RHrtcomDR
comX	RHrtcomXF/RHrtcomXR
comW	RHrtcomWF/RHrtcomWR
cglA	RHrtcglAF/RHrtcglAR
coiA	RHrtcoiAF/RHrtcoiAR

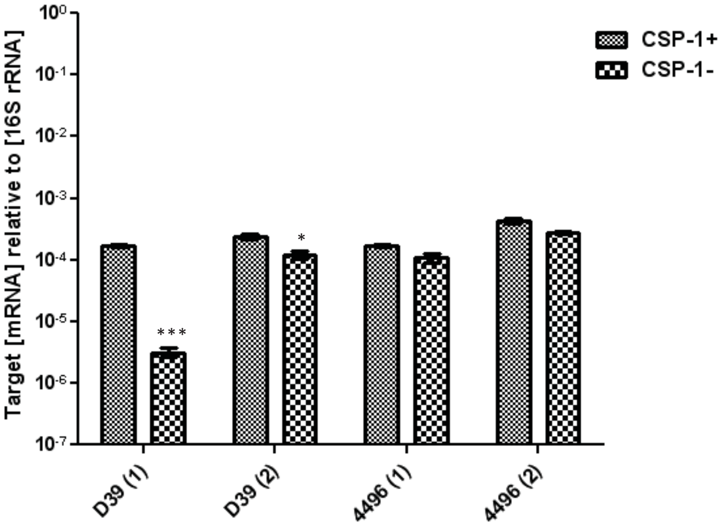
*Primer sequences are described in Table 2.3

The expression of *comD* was significantly greater in the presence of CSP-1 for both D39 experiments (*P*<0.001 and *P*<0.05, respectively). However, the fold increase in *comD* expression was much greater in experiment 1 due to the lower baseline expression in the sample lacking CSP-1 (Figure 6.4). In contrast, *comD* expression was not significantly different in the presence or absence of CSP-1 in either strain 4496 experiment. Nevertheless, the absolute level of expression of *comD* by strain 4496 (Figure 6.4) suggests that the CSP receptor is being transcribed at a level that, at least in D39, is sufficient for transformation. Therefore, *comD* expression is unlikely to account for the difference in transformability between the two strains. *ComX* expression was significantly greater in the presence of CSP-1 in both experiments of both strains (*P*<0.001), which indicated that CSP-1 was able to induce early competence gene expression in both D39 and 4496 (Tables 6.2 and 6.3). In contrast, CSP-1 was not associated with either a consistent increase or decrease in *comW* expression in either D39 or 4496. Interestingly, Figure 6.4 shows that the mean expression of *comW* in strain 4496 was significantly greater than in D39 (*P*<0.05), in a CSP-1-independent manner. The expression of *comW* in strain 4496 compared to D39 might indicate that *comW* over-expression could inhibit competence. However, successful transformation with expression of *comW* on a multicopy plasmid using a nisin-inducible promoter suggests that this is not case (Luo *et al*., 2004). Alternatively, *comW* over-expression could occur in response to the absence of some unknown negative feedback mechanism that might limit *comW* expression in D39. *CglA* expression was consistently greater in the presence of CSP-1 than without CSP-1 in both D39 and strain 4496. Furthermore, increased *cglA* expression in response to CSP-1 suggests that CSP-1 is able to induce the expression of some late competence genes in strain 4496. Therefore, *cglA* expression was unlikely to be linked to the lack of strain 4496 transformability. *CoiA* expression was consistently greater in the presence of CSP-1 than without CSP-1 in

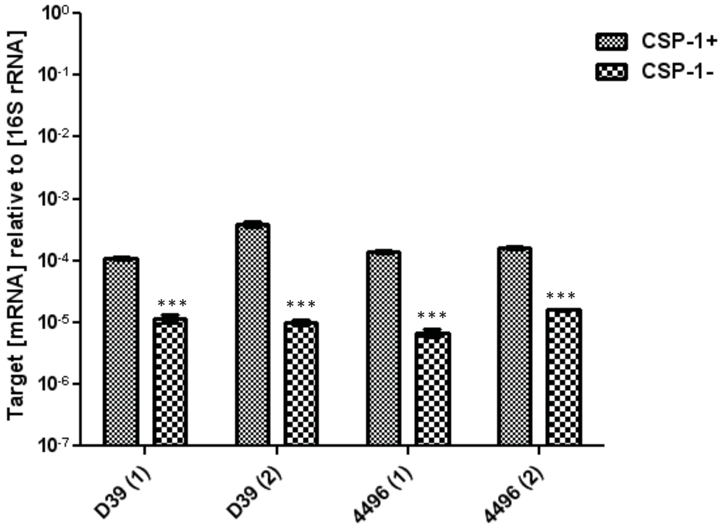
Figure 6.4 Expression of key competence genes relative to 16S rRNA in D39 and 4496 in response to CSP-1

The mean expression of the target competence genes relative to 16S rRNA in the presence and absence of CSP-1 was determined by real time RT-PCR from two duplicate transformations per strain. Competent cells were prepared and transformations were performed on two independently prepared batches of competent cells (Section 2.8). All competent cells were incubated in CTM pH 7.8 at 37°C with or without CSP-1, before adding chromosomal DNA. The remainder of the transformation protocol was followed, as described in Section 2.8.2. Immediately prior to adding DNA, aliquots of each transformation mix were taken for the extraction of RNA and DNase treatment as described in Section 2.12.1. Real-time RT-PCR was performed, as described in Section 2.12.6. The relative difference in expression was calculated using the◁ Ct method, as described in Section 2.12.7. Error bars indicate the standard deviation of triplicate reactions for each gene. Statistical significance was determined by unpaired two-tailed *t*-test (*, $P<0.05$; **, $P<0.01$; ***, $P<0.001$). Black asterisks indicate a significant difference in expression relative to 16S rRNA between the presence and absence of CSP-1. The red asterisk indicates differences in the the mean expression of *comW* relative to 16S rRNA between all four D39 samples compared to all four strain 4496 samples.

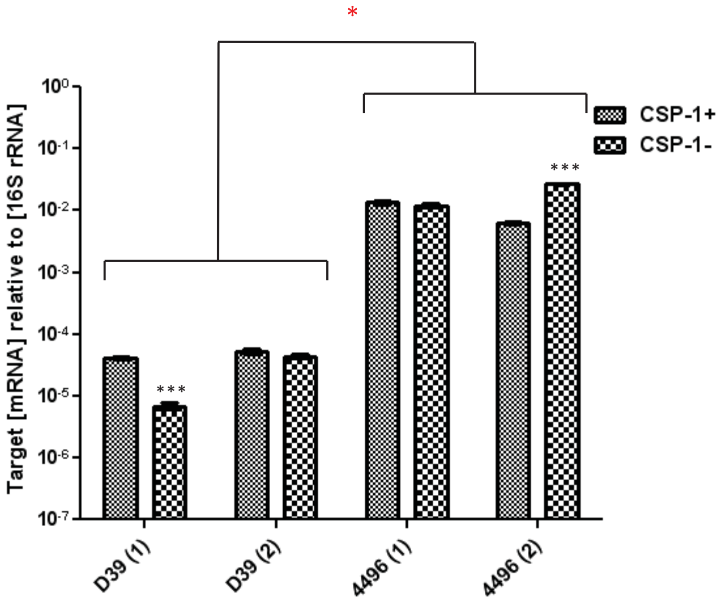
ComD



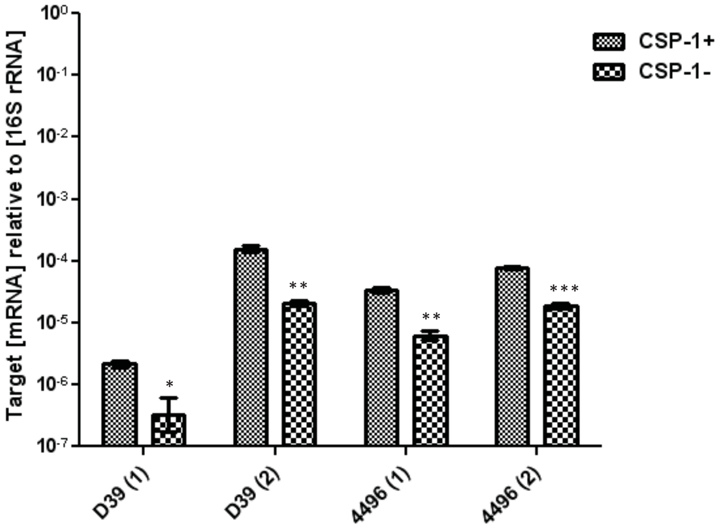
ComX



ComW

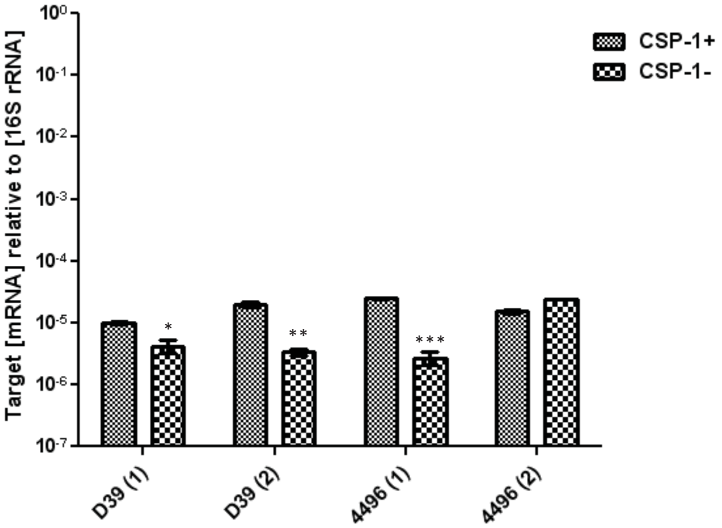


CglA



Target [mRNA] relative to [16S rRNA]
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
CSP-1+
CSP-1-
*
**
**
***
D39 (1)
D39 (2)
4496 (1)
4496 (2)

CoiA



Target [mRNA] relative to [16S rRNA]
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
CSP-1+
CSP-1-
*
**
***
D39 (1)
D39 (2)
4496 (1)
4496 (2)

Table 6.2 Relative expression of competence genes in D39 in the presence of CSP-1 versus the absence of CSP-1

ORF	Exp 1		Exp 2	
	Fold change	Direction	Fold change	Direction
comD	53.82^{c}	up	2.00^{a}	up
comX	9.30^{c}	up	38.76^{c}	up
comW	5.90^{c}	up	1.19^{ns}	up
cglA	>6.74*^{a}	up	7.53^{b}	up
coiA	2.38^{a}	up	5.84^{b}	up

*Indicates where the target was not detected in the absence of CSP-1

Results of statistical analysis: ns, not significant (includes values <2); *a*, *P*<0.05; *b, P*<0.01; *c, P*<0.001

Table 6.3 Relative expression of competence genes in 4496 in the presence of CSP-1 versus the absence of CSP-1

ORF	Exp 1		Exp 2	
	Fold change	Direction	Fold change	Direction
comD	1.59^{ns}	up	1.56^{ns}	up
comX	20.87^{c}	up	10.24^{c}	up
comW	1.13^{ns}	up	4.31^{c}	down
cglA	5.41^{b}	up	4.07^{c}	up
coiA	9.30^{c}	up	1.59^{ns}	down

Results of statistical analysis: ns, not significant (includes values <2); *a*, *P*<0.05; *b, P*<0.01; *c, P*<0.001

D39. Similarly, experiment 1 of strain 4496 showed that increased *coiA* expression occurred in the presence of CSP-1 ($P<0.001$). In the second 4496 experiment, *coiA* expression was not significantly altered by the addition of CSP-1. However, as the baseline expression level in this experiment was higher than the first 4496 experiment and both D39 experiments (Figure 6.4), the expression of this gene was probably not linked to the lack of transformability.

In summary, the expression of the competence genes *comD*, *comX*, *cglA* and *coiA* in strain 4496 was for the most part not particularly different from D39. Therefore, strain 4496 responds to CSP-1, by inducing the expression of the early competence gene, *comX*, which in turn induces the expression of the late competence genes, such as *cglA*. However, it was interesting that *comW* over-expression by strain 4496 occurred both in the presence and absence of CSP-1. Interestingly, over-expression in the competence regulatory system has been previously reported, but in ComW-deficient mutants that were also non-transformable (Sung & Morrison, 2005). In this case it was thought that the absence of ComW had inhibited a negative feedback mechanism that would have normally limited *comX* expression. Therefore, it is possible that *comW* expression might similarly be suppressed in a normally functioning competence system. In light of the expression data of this section, it is possible that some defect in a component of the late competence genes of strain 4496 might contribute to the strain's lack of transformability.

6.6 Essential competence genes present in strain 1861 and 4496 by CGH

As described in Section 6.1, competence induction is a complex process that requires the expression of 23 genes (Peterson *et al.*, 2004). However, given that *comW* was over-expressed in strain 4496 compared to D39 (Section 6.5), perhaps strain 4496 had some defect in a component of the competence system that if functional, would allow a negative feedback mechanism to limit *comW* expression and allow transformation. Therefore, the presence of essential competence genes was confirmed in strains 1861 and 4496 using the raw CGH data generated in Section 5.2. For genes that were not detected by CGH, searches of the 1861 consensus sequence (generated in Section 5.3) were performed to independently verify the CGH data. Since the genome sequence of only strain 1861 was available and that both strains 1861 and 4496 were non-transformable, it was assumed that the strain 1861 consensus sequence would be representative of both strains 1861 and 4496. 17 of the essential competence genes were detected in strains 1861 and 4496 by CGH (Table 6.4). Whilst *comA*, *comEA*, *comEC* and *ssbB* were not detected by CGH in either strain 1861 or 4496, analysis of the strain 1861 consensus sequence confirmed that these genes were indeed present in strain 1861. Therefore, it is probable that *comA*, *comEA*, *comEC* and *ssbB* are also present in strain 4496. In addition, alignments between *comA*, *comEA*, *comEC* and *ssbB* of TIGR4 and strain 1861 (data not shown) revealed that all genes contained small pockets of variable sequence that may have prevented hybridisation in CGH.

However, despite CGH confirming the presence of the majority of competence genes, the technology is unable to detect point mutations that could lead to premature termination of translation. In addition, sequence variation in non-coding sequence could interfere with the expression of some competence genes. Furthermore, there are many other apparently non-essential genes with unknown roles in competence that exhibit CSP-1 inducible expression (Peterson *et al*., [2004]).

Table 6.4 Detection of essential competence genes by CGH and genomic sequencing

Gene	TIGR4 ID	Putative role in competence*	1861	4496
comA	SP_0042	CSP export	+	-
comB	SP_0043	CSP export	+	+
comE	SP_2236	Response regulator for early competence gene expression	+	+
comX#	SP_0014 & SP_2006	Activator of late competence gene expression	+	+
comW	SP_0018	Stabilisation of ComX	+	+
coiA	SP_0978	DNA processing and recombination	+	+
cglA	SP_2053	DNA uptake	+	+
cglB	SP_2052	DNA uptake	+	+
cglC	SP_2051	DNA uptake	+	+
cglD	SP_2050	DNA uptake	+	+
cglG	SP_2047	DNA uptake	+	+
comFB	SP_2207	DNA uptake	+	+
ccla	SP_1808	DNA uptake	+	+
recA	SP_1940	DNA processing and recombination	+	+
ssbB	SP_1908	DNA processing and recombination	+	-
comEA	SP_0954	DNA uptake	+	-
comFA	SP_2208	DNA uptake	+	+
comEC	SP_0955	DNA uptake	+	-
dprA	SP_1266	DNA processing and recombination	+	+

ComC and *ComD* are not included above as their presence had already been confirmed (Section 6.2).
# Two identical copies exist in the chromosome, which are impossible to detect individually using either CGH or the strain 1861 consensus sequence.
*Roles are reviewed in Claverys *et al*. (2009) and Johnsborg and Havarstein (2009)
Red '+' Indicates those genes confirmed from the consensus sequence of strain 1861.

6.7 Discussion

A significant roadblock to definitively confirming the importance of genomic ARs to the invasive potential of strains 1861 and 4496 was the inability to genetically manipulate these strains. Therefore, this chapter summarised the attempts to transform strains 1861 and 4496. Whilst it has been previously reported that many clinical isolates of *S. pneumoniae* are difficult to transform *in vitro* (Pozzi *et al*., 1996), it has not been shown whether such difficulty is due to an inherent defect in the competence system of those strains or whether the inability to transform such strains is due to suboptimum *in vitro* conditions.

6.7.1 Compatibility of CSP pherotype and ComD

Since it is known that the concentration of CSP required for competence induction is dependent on the compatibility of the CSP pherotype and its receptor encoded by *comD* (Iannelli *et al*., 2005), it was decided to sequence *comC* and *comD* of strains 1861 and 4496 to ensure that the CSP of the correct pherotype was used in subsequent transformation attempts. Sequencing revealed that strains 1861 and 4496 both encoded CSP-1. In addition, strain 1861 encodes ComD1 and strain 4496 encodes ComD4. However, whilst tenfold greater concentration of CSP-1 is required for maximum activation of ComD4 compared to ComD1 (Iannelli *et al*., 2005), the concentration of CSP-1 used in this study exceeded this threshold. Therefore, the CSP receptors of strains 1861 and 4496 were unlikely to have limited the transformability of these strains.

6.7.2 Optimisation of conditions required for *in vitro* transformation of strain 4496

Given that it was not known whether the inability to successfully transform strain 4496 was due to the use of suboptimum *in vitro* conditions, transformation of strain 4496 was attempted at a range of culture densities and in a number of different media. However, whilst D39 was competent in all conditions, successful transformation of strain 4496 could not be achieved. In addition, a number of different DNA donors were used to transform strain 4496, including plasmid DNA, chromosomal DNA and PCR products. However, whilst all were able to successfully transform D39, transformation of strain 4496 could still not be achieved. The inability of strain 4496 to be transformed by plasmid DNA suggested that strain 4496 was unable to take up exogenous DNA.

Alternative methods of transformation were attempted, including electrotransformation and natural transformation in the presence of Lipofectamine^{TM} 2000 reagent. However, electrotransformation could not transform even D39 and Lipofectamine^{TM} did not have a detectable impact on the transformation efficiency of D39. In both cases successful transformation of strain 4496 was not achieved.

6.7.3 Comparison between the CSP-induced expression of key competence genes in D39 and strain 4496

Real-time RT-PCR was used to compare the expression of the competence genes *comD*, *comX*, *comW*, *cglA* and *coiA* between D39 and strain 4496, both in the presence and absence of CSP-1. The expression of *comD* was not thought to impact on the transformability of strain 4496, as its level of transcription in strain 4496 was at least equal, if not greater than that of D39, where transformation was successful. In addition, a significant increase in *comX* expression was observed in the presence of CSP-1 by both D39 and strain 4496. The increase in *comX* expression indicated that whilst strain 4496 was not successfully transformed, the strain was able to respond to CSP-1. Furthermore, the expression of late competence genes, such as *cglA* was increased in the presence of CSP-1, suggesting that the ability of ComX to activate late competence gene expression was probably not compromised. However, a significant difference was observed in *comW* expression between D39 and strain 4496, as the gene appeared to be over-expressed in strain 4496 in a CSP-independent manner. Therefore, it was thought that perhaps the over-expression of *comW* had inhibited competence. However, given the role of ComW in the stabilisation of ComX (Section 6.1), if ComW over-expression had inhibited competence then it would be expected that the induction of late gene expression would have been compromised, which did not appear to be the case. Instead it was thought that perhaps some defect in a downstream component of the competence system had prevented the successful transformation of strain 4496, and subsequently resulted in the failure of an unknown negative feedback mechanism that would normally suppress *comW* expression.

6.7.4 Search for missing genes known to be required for successful transformation

In Section 6.6, searches for the presence of essential competence genes were performed in case the absence of one or a number of such genes was responsible for the lack of transformability of strains 1861 and 4496. However, whilst CGH suggested that *comA*, *comEA*, *comEC* and *ssbB* were absent from both strains 1861 and 4496, verification using the genome sequence of strain 1861, showed that the sequence of these genes varied from that of TIGR4. Therefore, all essential competence genes identified by Peterson *et al*. (2004), were present in strain 1861, and probably strain 4496. Instead possible defects in the competence system that could not be detected by CGH might exist, such as single nucleotide frame-shift mutations in competence genes or sequence changes in non-coding regions that are important for the expression of some competence genes.

Chapter 7 – Final Discussion

The pneumococcus is a globally successful pathogen, responsible for significant mortality and morbidity (Section 1.1). However, such high mortality and morbidity occurs primarily as a consequence of the extraordinary rate of pneumococcal carriage, which has been recorded as high as 90% in some communities within both developed and developing countries (Section 1.2.1). Such successful global colonisation requires the ability to adapt to environments that can differ in terms of antibiotic usage, vaccination coverage, health of the host, host population density, climate and competition with other bacteria. However, presumably as a consequence of adaptation, differences in invasive potential exist between different strains of *S. pneumoniae*. The ability of the pneumococcus to adapt to the environment is thought to be facilitated by the plasticity of the pneumococcal genome. Such plasticity has promoted an enormous pool of non-core sequence within the *S. pneumoniae* pan-genome, which is at least as large in size as the core genome (Hiller *et al*., 2007). The pool of non-core sequence is divided into ARs, which are thought to not only promote environmental adaptation at the species level, but also influence the invasive potential of individual clones.

The primary objective of this study was to identify and characterise serotype-independent ARs that contribute to increased invasive potential. In preliminary work, the virulence of a selection of serotype 1 isolates that were either non-invasive or invasive in humans, was characterised in mice, which identified non-invasive, intermediately virulent and highly virulent isolates (Section 1.6.1). In addition, it was revealed that the highly virulent strains were able to rapidly invade and survive in the blood, whilst the non-invasive strains did not appear to progress any further than colonisation of the nasopharynx (Section 1.6.1). Preliminary genomic comparisons identified a chromosomally-encoded TA system (PezAT) within the variable region of PPI-1 that was only present in the highly virulent strains and not in either the non-invasive or intermediately virulent strains (Section 1.6.3). Interestingly, PezAT had previously been reported to be involved in the virulence of TIGR4 (Brown *et al*., 2004). Other studies have attempted to identify ARs associated with invasive potential (Section 1.5.3.2). However, such studies have usually been based on broad assumptions about invasive potential between serotypes and genotypes. In contrast, this study performed genomic comparisons on a group of serotype 1 isolates with known virulence in humans and mice, in order to identify ARs associated with invasive potential.

7.1.1 Genetic diversity of serotype 1 isolates

As mentioned above, the virulence of a selection of non-invasive and invasive serotype 1 human isolates were grouped as non-invasive, intermediately virulent and highly virulent following characterisation in mice. However, whilst the non-invasive and intermediately virulent strains were known to be in lineage A (Section 1.5.2.1), the genetic relatedness of these strains to the highly virulent strains was unknown. Therefore, the STs of the highly virulent strains were determined using MLST, which found strain 1861 to be of lineage B and strain 4496 to be of lineage C (Section 3.2.1). Interestingly, strain 1861 is very closely related to the hypervirulent clones responsible for epidemics of severe IPD with unusually high mortality rates (Antonio *et al*., 2008). Therefore, it is quite likely that strain 1861, and probably strain 4496, are also quite virulent in humans, which is consistent with the heightened virulence of these isolates in mice (Section 1.6.1).

7.1.2 Sequence analysis of the variable region of the Pneumococcal Pathogenicity Island 1

In preliminary genomic comparisons, PezAT was associated with the highly virulent strains (Section 1.6.3). The role of chromosomally encoded TA systems is controversial (reviewed in Van Melderen & Saavedra De Bast [2009]) and has been suggested to mediate processes such as resistance to environmental stresses and the genomic stability of non-core regions of the chromosome (Szekeres *et al*., 2007). Therefore, if PezAT had a role in maintaining the stability of non-core regions of the genome, it was possible that PezAT might function to prevent the loss of virulence-enhancing genes within the PPI-1 variable region of the highly virulent strains. Sequencing and subsequent analysis of the PPI-1 variable region in the non-invasive, intermediately virulent and highly virulent strains (Section 3.3), predicted that the non-invasive and intermediately virulent strains lacked *pezAT*, as expected, and possessed a truncated version of a putative neopullulanase encoded by *nplT*. A major feature of the PPI-1 variable region in these strains was the presence of a putative immunity system against the bacteriocin, mersacidin that is expressed *in vitro* by strain 1 (Section 4.2), and might provide a survival advantage in the presence of mersacidin-producing pneumococci and against other species present in the nasopharynx. However, the importance of mersacidin in the nasopharyngeal environment is unknown as it is unknown what species produce mesacidin and how common it is for the bacteriocin to be present in the nasopharynx.

In contrast, the highly virulent strains possess a full-length, but fragmented, *nplT* and encode PezAT (Section 3.3). However, the predominant feature of the PPI-1 variable region is a putative operon (Section 4.2) that encodes hypothetical proteins and putative enzymes responsible for catalysing the rate-limiting steps of BCAA catabolism, phenylalanine biosynthesis and UDP-sugar conversion (Section 3.3.2.2). In addition, a biotin carboxylase and a putative transporter of the major facilitator superfamily are present within the region of the highly virulent strains. In summary, one role for the PPI-1 variable region of the highly virulent strains might be to regulate multiple metabolic pathways following the activation of a single promoter (Section 4.2).

Alignments between the PPI-1 variable region of the strains of this study with the region in a number of strains with publicly available genomes highlighted the composite nature of the PPI-1 variable region (Section 3.5.3). The region could be divided into components including *pezAT,* Tn*5252*-associated sequence and an accessory region that could include a variety of genes with various functions. Therefore, it was decided to screen a large selection of *S. pneumoniae* strains belonging to various serotypes to assess the prevalence of different versions of the PPI-1 variable region (Section 3.4.2.2). Interestingly, the mersacidin immunity system present in the lineage A strains of this study was the most common component of the PPI-1 variable region in the tested strains, and was present in a range of serotypes and clones, including the pandemic carriage Spanish^{23F} ST81 clone (Croucher *et al*., 2009) and serotypes 18C, 3, 11, 19A, 9N and 24. In addition, the accessory region of strains 1861 and 4496 was the next most common component of the region and was present in serotypes including 19F and 11A. Therefore, whilst the PPI-1 variable region is quite diverse, the region appears to have an ordered structure. In summary, the PPI-1 variable region appears to be a hotspot for recombination within the genome that could promote adaptation and contribute to differences in the invasive potential of the serotype 1 isolates.

7.1.3 Functional characterisation of the Pneumococcal Pathogenicity Island 1

The expression of a number of genes that encode enzymes that catalyse the rate-limiting steps of a number of metabolic pathways were transcriptionally coupled in the PPI-1 variable region of highly virulent serotype 1 isolates (Section 4.2). Interestingly, this operon exhibited significantly greater expression in the blood and lungs when compared to the surface of the nasopharynx, which suggested that the catabolism of BCAAs, the biosynthesis of phenylalanine and the conversion of UDP-sugars were possibly up-regulated by the highly virulent strains in the blood and lungs compared to the surface of the nasopharynx. In addition, the expression of *nplT* and *pezAT* was also found to be greater in the blood and lungs compared to the surface of the nasopharynx and expression of the major facilitator was greater in the blood than either the lungs or the surface of the nasopharynx. Therefore, it is quite clear that a significant proportion of the PPI-1 variable region in strains 1861 and 4496 was preferentially expressed in niches associated with disease, thus supporting a role for this region in IPD. The contribution of the region to virulence was confirmed by mutagenesis. However, due to the inability to genetically manipulate the serotype 1 isolates of this study (Chapter 6), mutagenesis was instead carried out on the easily transformable serotype 2 strain, D39. Competition between the mutants that were generated in Section 4.4.2 showed that the D39 mutant harbouring the PPI-1 variable region from strain 1861 was significantly more fit in the blood, lungs and nasopharyngeal tissue than the D39 mutant harbouring the PPI-1 variable region from strain 1. The consistency between the competition data of the D39 mutants and the virulence of strains 1 and 1861 strongly suggests that the PPI-1 variable region contributes to the heightened virulence of strains 1861 and 4496.

7.1.4 Identification of regions associated with hypervirulent serotype 1 isolates

Whilst the PPI-1 variable region of the highly virulent strains was shown to contribute to survival *in vivo*, the region may not be solely responsible for the heightened virulence of strains 1861 and 4496. Therefore, lists of additional ARs were compiled that were either associated with all strains that were capable of IPD or with only the highly virulent strains. Of particular interest was that the *psrP-secAY2A2* region is present in the intermediately virulent strains and not the non-invasive strains (Section 5.2.2). As the region has been suggested to be involved in adherence to lung epithelium and controversially suggested to be required for the virulence of TIGR4 (Orihuela, 2009; Blomberg *et al*., 2009; Shivshankar *et al*., 2009; Obert *et al*., 2006), it is possible that the inability of the non-invasive strains to cause IPD may be due to the absence of *psrP-secAY2A2*. However, the importance of *psrP-secAY2A2* in virulence was complicated by the fact that whilst strain 4496 possessed the region, strain 1861 did not. Therefore, either the region is not required for virulence, or strain 1861 was able to compensate for the absence of *psrP-secAY2A2* by possessing an alternative adherence factor. A potential alternative to PsrP is PblB, which mediates binding to human platelets by *S. mitis* and is encoded within a lysogenic phage (Bensing *et al*., 2001). Furthermore, PblB is present only in the highly virulent strains and not in either the non-invasive or intermediately virulent strains. In addition to PblB, the prophage also encodes an endolysin, presumably involved in the lysis of the host bacterium and phage release during the lytic cycle (reviewed in Loessner *et al*., 2005). Interestingly, the endolysin of the highly virulent strains exhibited greater expression in the blood and lungs of infected mice than on the surface of the nasopharynx. Increased endolysin-mediated lysis could indirectly lead to increased display of PblB on the surface of unlysed pneumococci as described for *S. mitis* (Mitchell *et al*., 2007). In addition, the role of *S. pneumoniae*-bound plasmin in the penetration of epithelial monolayers (Attali *et al*., 2008) suggests that greater PblB-mediated attachment to plasmin could facilitate greater migration into sterile sites and the development of IPD. Furthermore, the cell lysis triggered by the endolysin could also contribute directly to virulence via a similar mechanism to LytA (Section 1.3.1.2). Therefore, it is possible that the prophage that encodes PblB and the endolysin might not only compensate for the absence of PsrP by strain 1861, but could also contribute to the heightened virulence of both highly virulent strains.

The presence of a conjugative transposon (Tn*5253*) only in the highly virulent strains and not either the intermediately virulent or non-invasive strains could increase the resistance of strains 1861 and 4496 to environmental stresses. For example the region harbours *umuC*- and *umuD*-like genes, which are thought to be involved in the *E. coli* stress response (Munoz-Najar & Vijayakumar, 1999). In addition, the region encodes a putative TA system, homologous to PezAT and genes for resistance to tetracycline. Immediately upstream of Tn*5253* is *zmpD*, which encodes a putative zinc metalloproteinase. Other zinc metalloproteinases such as IgA1 protease and ZmpC have known functions in virulence (Section 1.5.3.1). However, the substrate of ZmpD remains to be determined.

Another potentially important virulence-associated AR encodes an ABC transporter and enzymes thought be involved in the degradation and transport of sialic acids. This region is present only in the highly virulent strains and not in either the intermediately virulent or non-invasive strains. Interestingly, *in vivo* expression analysis also suggested that components of this region were more active in the blood than in the lungs or on the nasopharyngeal mucosa. Therefore, the genes of this region may promote survival within the bloodstream and perhaps contribute to the severity of disease.

In addition, a 408-bp deletion and a late frameshift mutation (Section 5.5.2) were identified within *hylA* of the non-invasive strains, which could contribute to the reduced invasive potential of these strains if these nucleotide changes impact on the enzyme’s activity.

In summary, whilst a number of differences were identified between the genomes of the non-invasive, intermediately virulent and highly virulent strains, a number of key regions and genes, which include the PPI-1 variable region, *psrP-secAY2A2*, the PblB-associated prophage Tn*5253*, ZmpD and the sialic acid degradation/transport region, appeared to be the most likely to promote IPD and thus contribute to the differences observed in virulence between the serotype 1 isolates of this study.

7.1.5 Competence of serotype 1 isolates

Mutagenesis is an important tool for confirming the importance of potential virulence factors to IPD. However, a frustrating feature of the serotype 1 isolates used in this study was their apparent inability to undergo genetic transformation *in vitro*. Furthermore, whilst difficulty in genetically manipulating clinical isolates of *S. pneumoniae* is not uncommon (Pozzi *et al*., 1996), it is not clear whether these isolates are inherently non-transformable due to a naturally occurring defect in the competence system, or whether these isolates are only non-transformable in the *in vitro* conditions that were used. In this study, attempts were made to address both possibilities. Sequencing confirmed that strains 1861 and 4496 both encoded a CSP receptor that is compatible with CSP-1. In addition, different DNA donors, culture densities and media did not have a detectable effect on transformability. A number of alternative methods of genetic manipulation were also unsuccessfully trialled. Therefore, the possibility that a naturally occurring defect exists in the competence system of these strains was investigated. Gene expression comparisons showed that strain 4496 was able to respond to exogenous CSP-1 by increasing the expression of *comX*, which in turn triggered an increase in the expression of two late competence genes in a manner that was mostly consistent with D39. However, an interesting observation was the apparent CSP-1-independent over-expression of *comW* in strain 4496 when compared to D39. Initially it was thought that perhaps excessive *comW* expression could inhibit competence. However, if this was the case then given the role of ComW in the stabilisation of

ComX, it would be expected that ComX activity would be compromised. Instead, evidence of CSP-1-associated late competence gene expression suggested that ComX activity was unaffected. Therefore, perhaps a defect in a downstream component of the competence system had either directly or indirectly caused a lack of transformability, and resulted in over-expression of *comW* in 4496.

7.2 Conclusion

The diversity of the pneumococcal genome is thought to facilitate the pathogen's ability to adapt to different host environments and enable such high infection rates throughout the world. However, such adaptability appears to result in differences in invasive potential. Whilst a number of studies have attempted to identify ARs that determine the invasive potential of the host strain, this is the first study to perform comparisons on a group of serotype 1 isolates with well-defined virulence characteristics that appear to be similar in both mice and humans. In addition, the genes that were associated with heightened virulence in these strains might be responsible for the severity of disease caused by clonally-related serotype 1 isolates in sub-Saharan Africa. An understanding of the mechanism behind the progression from carriage to IPD is fundamental to understanding pneumococcal pathogenesis. Furthermore, by identifying factors responsible for the invasive potential of hypervirulent clones of *S. pneumoniae*, it might be possible to develop treatment strategies that selectively target these highly virulent clones during epidemics of severe IPD.

7.3 Future directions

As might be expected from large scale genetic/genomic comparisons, this study has identified many avenues for further research into the contribution of genetic variability to the invasive potential of *S. pneumoniae*. In particular, whilst the PPI-1 variable region of the highly virulent strains was shown to have a role in IPD, future work should focus on specific components of the region, by comparing the competitive fitness *in vivo* of *nplT*-, *pezAT*- and accessory region-deficient mutants. In addition, the various D39 PPI-1 mutants should also have the endogenous capsule locus replaced with the serotype 1 locus to assess whether the serotype 1 capsule either increases or decreases the importance of the PPI-1 variable region in virulence. Furthermore, the importance of the PPI-1 variable region to the invasive potential of a variety of serotypes and clones should be investigated. *In vitro* assays should also be used to confirm the activity of the enzymes encoded within the region. In addition, as more bioinformatic data become public, the functions of hypothetical proteins within the region could also be predicted and confirmed *in vitro*.

The mersacidin immunity system that is present in many different serotypes and clones, including the non-invasive and intermediately virulent serotype 1 isolates of this study, and the Spanish^{23F} ST81, clone should be investigated further. Firstly, *in vitro* growth comparisons should be used to confirm that the mersacidin immunity system is functional in these strains, and should be followed by investigations into possible mersacidin-mediated competition amongst the nasopharyngeal microflora. The findings of such studies may inform the development of mersacidin as a future antibiotic against methicillin-resistant *S. aureus* (Kruszewska *et al*., 2004).

The genomic comparisons of Chapter 5 unearthed enormous potential for future work. However, some key areas warrant particular priority. Initially the prevelance of key virulence-associated genes, such as *pblB* and the putative sialic acid utilisation region could be assessed in a selection of strains belonging to various serotypes using PCR and Southern blotting. In addition, *in vitro* and *in vivo* assays should confirm the contribution of the genes of interest to invasive potential. For example, particular attention should be paid to the role of PblB in adherence to human platelets. Further investigations should test to see if the invasive potential of a non-invasive or intermediately virulent strain is increased by the insertion of *pblB* and its associated endolysin into the genome. Similarly, it would be interesting to see if the sialic acid utilisation region provides a survival advantage within the blood to strains that normally lack the region. The ability of the pneumococcus to grow in the presence of sialic acid should also be addressed *in vitro* using assays that examine both planktonic growth and biofilm formation. Future work should also address the TA system in Tn*5253* that is homologous to PezAT. Whilst some evidence suggests that chromosomal TA systems rarely complement each other in *trans* (Fiebig *et al*., 2010), the ability of this system to compensate for the absence of PezAT should be investigated by mutagenesis.

The activity of HylA encoded by the non-invasive strains should be compared with the full-length enzyme *in vitro*. In addition, the full-length *hylA* in a strain such as D39 should be replaced with the strain 1 allele to assess the impact on virulence.

Allelic differences in *pspA*, *zmpB*, *iga*, *prtA*, *phtB* and *phtD* between the serotype 1 isolates should be investigated by sequencing the genes in all six strains to assess whether certain alleles of each gene are associated with a particular phenotype.

As discussed in Section 1.3.2, it has been suggested that long-term colonisation by *S. pneumoniae* requires localised deep invasion within the nasopharyngeal tissue. Therefore, the nasopharyngeal tissue could be a key niche for the transition from asymptomatic carriage to IPD. However, gene expression comparisons with pneumococci from the nasopharyngeal tissue have been hampered by microfloral contamination that interferes with the quantification of the internal reference gene (16S rRNA). Therefore, an alternative internal reference gene should be investigated and optimised to replace 16S rRNA in order to enable *S. pneumoniae*-specific quantification of target mRNAs within the nasopharyngeal tissue.

Further investigation into the competence of strain 4496 should include microarray analysis to compare global transcription patterns between strain 4496 and D39 in the presence and absence of CSP-1. This may allow the potentially defective component of the competence system to be identified. The analysis of strain 4496 will also shed light on parts of the competence system that are not well understood, such as the role of ComW. In addition, alternative options for the transformation of strain 4496 should also be investigated. For example transformation of strain 4496 in the presence of human lung epithelial cells should be attempted. Hammershcmidt *et al*. (2005) showed that immediately prior to invasion of lung cells, pneumococci become unencapsulated. Therefore, such a reduction in capsule might facilitate the import of DNA into the cell. In addition, genetic manipulation mediated by conjugation and transduction should also be attempted as these should operate independently of the competence system. Development of a method for genetic manipulation of the serotype 1 isolates of this study would enable more direct confirmation of the roles of candidate virulence genes in determining the invasive potential of *S. pneumoniae*.

References

Aaberge, I. S., Eng, J., Lermark, G. and Lovik, M. (1995). "Virulence of Streptococcus pneumoniae in mice: a standardized method for preparation and frozen storage of the experimental bacterial inoculum." Microb Pathog **18**(2): 141-52.

Adamou, J. E., Heinrichs, J. H., Erwin, A. L., Walsh, W., Gayle, T., Dormitzer, M., Dagan, R., Brewah, Y. A., Barren, P., Lathigra, R., Langermann, S., Koenig, S. and Johnson, S. (2001). "Identification and characterization of a novel family of pneumococcal proteins that are protective against sepsis." Infect Immun **69**(2): 949-58.

Aguiar, S. I., Serrano, I., Pinto, F. R., Melo-Cristino, J. and Ramirez, M. (2008). "The presence of the pilus locus is a clonal property among pneumococcal invasive isolates." BMC Microbiol **8**: 41.

Ahmad, H. and Chapnick, E. K. (1999). "Conjugated polysaccharide vaccines." Infect Dis Clin North Am **13**(1): 113-33, vii.

Alanee, S. R., McGee, L., Jackson, D., Chiou, C. C., Feldman, C., Morris, A. J., Ortqvist, A., Rello, J., Luna, C. M., Baddour, L. M., Ip, M., Yu, V. L. and Klugman, K. P. (2007). "Association of serotypes of Streptococcus pneumoniae with disease severity and outcome in adults: an international study." Clin Infect Dis **45**(1): 46-51.

Alcantara, R. B., Preheim, L. C. and Gentry-Nielsen, M. J. (2001). "Pneumolysin-induced complement depletion during experimental pneumococcal bacteremia." Infect Immun **69**(6): 3569-75.

Allegrucci, M., Hu, F. Z., Shen, K., Hayes, J., Ehrlich, G. D., Post, J. C. and Sauer, K. (2006). "Phenotypic characterization of Streptococcus pneumoniae biofilm development." J Bacteriol **188**(7): 2325-35.

Alloing, G., Martin, B., Granadel, C. and Claverys, J. P. (1998). "Development of competence in Streptococcus pneumonaie: pheromone autoinduction and control of quorum sensing by the oligopeptide permease." Mol Microbiol **29**(1): 75-83.

Alpern, E. R., Alessandrini, E. A., McGowan, K. L., Bell, L. M. and Shaw, K. N. (2001). "Serotype prevalence of occult pneumococcal bacteremia." Pediatrics **108**(2): E23.

Altena, K., Guder, A., Cramer, C. and Bierbaum, G. (2000). "Biosynthesis of the lantibiotic mersacidin: organization of a type B lantibiotic gene cluster." Appl Environ Microbiol **66**(6): 2565-71.

Andersson, B., Dahmen, J., Frejd, T., Leffler, H., Magnusson, G., Noori, G. and Eden, C. S. (1983). "Identification of an active disaccharide unit of a glycoconjugate receptor for pneumococci attaching to human pharyngeal epithelial cells." J Exp Med **158**(2): 559-70.

Anderton, J. M., Rajam, G., Romero-Steiner, S., Summer, S., Kowalczyk, A. P., Carlone, G. M., Sampson, J. S. and Ades, E. W. (2007). "E-cadherin is a receptor for the common protein pneumococcal surface adhesin A (PsaA) of Streptococcus pneumoniae." Microb Pathog **42**(5-6): 225-36.

Antonio, M., Hakeem, I., Awine, T., Secka, O., Sankareh, K., Nsekpong, D., Lahai, G., Akisanya, A., Egere, U., Enwere, G., Zaman, S. M., Hill, P. C., Corrah, T., Cutts, F., Greenwood, B. M. and Adegbola, R. A. (2008). "Seasonality and outbreak of a predominant Streptococcus pneumoniae serotype 1 clone from The Gambia: expansion of ST217 hypervirulent clonal complex in West Africa." BMC Microbiol **8**: 198.

Aricha, B., Fishov, I., Cohen, Z., Sikron, N., Pesakhov, S., Khozin-Goldberg, I., Dagan, R. and Porat, N. (2004). "Differences in membrane fluidity and fatty acid composition between phenotypic variants of Streptococcus pneumoniae." J Bacteriol **186**(14): 4638-44.

Arrecubieta, C., Garcia, E. and Lopez, R. (1995). "Sequence and transcriptional analysis of a DNA region involved in the production of capsular polysaccharide in Streptococcus pneumoniae type 3." Gene **167**(1-2): 1-7.

Attali, C., Durmort, C., Vernet, T. and Di Guilmi, A. M. (2008). "The interaction of Streptococcus pneumoniae with plasmin mediates transmigration across endothelial and epithelial monolayers by intercellular junction cleavage." Infect Immun **76**(11): 5350-6.

Austrian, R. (1981). "Some observations on the pneumococcus and on the current status of pneumococcal disease and its prevention." Rev Infect Dis **3 Suppl**: S1-17.

Avery, O. T. and Dubos, R. (1931). "The Protective Action of a Specific Enzyme against Type Iii Pneumococcus Infection in Mice." J Exp Med **54**(1): 73-89.

Avery, O. T., Macleod, C. M. and McCarty, M. (1944). "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii." J Exp Med **79**(2): 137-58.

Ayoubi, P., Kilic, A. O. and Vijayakumar, M. N. (1991). "Tn5253, the pneumococcal omega (cat tet) BM6001 element, is a composite structure of two conjugative transposons, Tn5251 and Tn5252." J Bacteriol **173**(5): 1617-22.

Bagnoli, F., Moschioni, M., Donati, C., Dimitrovska, V., Ferlenghi, I., Facciotti, C., Muzzi, A., Giusti, F., Emolo, C., Sinisi, A., Hilleringmann, M., Pansegrau, W., Censini, S., Rappuoli, R., Covacci, A., Masignani, V. and Barocchi, M. A. (2008). "A second pilus type in Streptococcus pneumoniae is prevalent in emerging serotypes and mediates adhesion to host cells." J Bacteriol **190**(15): 5480-92.

Balachandran, P., Hollingshead, S. K., Paton, J. C. and Briles, D. E. (2001). "The autolytic enzyme LytA of Streptococcus pneumoniae is not responsible for releasing pneumolysin." J Bacteriol **183**(10): 3108-16.

Barbosa, J. A., Smith, B. J., DeGori, R., Ooi, H. C., Marcuccio, S. M., Campi, E. M., Jackson, W. R., Brossmer, R., Sommer, M. and Lawrence, M. C. (2000). "Active site modulation in the N-acetylneuraminate lyase sub-family as revealed by the structure of the inhibitor-complexed Haemophilus influenzae enzyme." J Mol Biol **303**(3): 405-21.

Barocchi, M. A., Ries, J., Zogaj, X., Hemsley, C., Albiger, B., Kanth, A., Dahlberg, S., Fernebro, J., Moschioni, M., Masignani, V., Hultenby, K., Taddei, A. R., Beiter, K., Wartha, F., von Euler, A., Covacci, A., Holden, D. W., Normark, S., Rappuoli, R. and Henriques-Normark, B. (2006). "A pneumococcal pilus influences virulence and host inflammatory responses." Proc Natl Acad Sci U S A **103**(8): 2857-62.

Basset, A., Trzcinski, K., Hermos, C., O'Brien, K. L., Reid, R., Santosham, M., McAdam, A. J., Lipsitch, M. and Malley, R. (2007). "Association of the pneumococcal pilus with certain capsular serotypes but not with increased virulence." J Clin Microbiol **45**(6): 1684-9.

Bensing, B. A., Siboo, I. R. and Sullam, P. M. (2001). "Proteins PblA and PblB of Streptococcus mitis, which promote binding to human platelets, are encoded within a lysogenic bacteriophage." Infect Immun **69**(10): 6186-92.

Bergmann, S., Rohde, M. and Hammerschmidt, S. (2004). "Glyceraldehyde-3-phosphate dehydrogenase of Streptococcus pneumoniae is a surface-displayed plasminogen-binding protein." Infect Immun **72**(4): 2416-9.

Bergmann, S., Rohde, M., Preissner, K. T. and Hammerschmidt, S. (2005). "The nine residue plasminogen-binding motif of the pneumococcal enolase is the major cofactor of plasmin-mediated degradation of extracellular matrix, dissolution of fibrin and transmigration." Thromb Haemost **94**(2): 304-11.

Bermpohl, D., Halle, A., Freyer, D., Dagand, E., Braun, J. S., Bechmann, I., Schroder, N. W. and Weber, J. R. (2005). "Bacterial programmed cell death of cerebral endothelial cells involves dual death pathways." J Clin Invest **115**(6): 1607-15.

Berry, A. M., Yother, J., Briles, D. E., Hansman, D. and Paton, J. C. (1989a). "Reduced virulence of a defined pneumolysin-negative mutant of Streptococcus pneumoniae." Infect Immun **57**(7): 2037-42.

Berry, A. M., Lock, R. A., Hansman, D. and Paton, J. C. (1989b). "Contribution of autolysin to virulence of Streptococcus pneumoniae." Infect Immun **57**(8): 2324-30.

Berry, A. M., Lock, R. A., Thomas, S. M., Rajan, D. P., Hansman, D. and Paton, J. C. (1994). "Cloning and nucleotide sequence of the Streptococcus pneumoniae hyaluronidase gene and purification of the enzyme from recombinant Escherichia coli." Infect Immun **62**(3): 1101-8.

Berry, A. M., Alexander, J. E., Mitchell, T. J., Andrew, P. W., Hansman, D. and Paton, J. C. (1995). "Effect of defined point mutations in the pneumolysin gene on the virulence of Streptococcus pneumoniae." Infect Immun **63**(5): 1969-74.

Berry, A. M., Lock, R. A. and Paton, J. C. (1996). "Cloning and characterization of nanB, a second Streptococcus pneumoniae neuraminidase gene, and purification of the NanB enzyme from recombinant Escherichia coli." J Bacteriol **178**(16): 4854-60.

Berry, A. M. and Paton, J. C. (1996). "Sequence heterogeneity of PsaA, a 37-kilodalton putative adhesin essential for virulence of Streptococcus pneumoniae." Infect Immun **64**(12): 5255-62.

Berry, A. M. and Paton, J. C. (2000). "Additive attenuation of virulence of Streptococcus pneumoniae by mutation of the genes encoding pneumolysin and other putative pneumococcal virulence proteins." Infect Immun **68**(1): 133-40.

Bethe, G., Nau, R., Wellmer, A., Hakenbeck, R., Reinert, R. R., Heinz, H. P. and Zysk, G. (2001). "The cell wall-associated serine protease PrtA: a highly conserved virulence factor of Streptococcus pneumoniae." FEMS Microbiol Lett **205**(1): 99-104.

Blomberg, C., Dagerhamn, J., Dahlberg, S., Browall, S., Fernebro, J., Albiger, B., Morfeldt, E., Normark, S. and Henriques-Normark, B. (2009). "Pattern of accessory regions and invasive disease potential in Streptococcus pneumoniae." J Infect Dis **199**(7): 1032-42.

Bogaert, D., Engelen, M. N., Timmers-Reker, A. J., Elzenaar, K. P., Peerbooms, P. G., Coutinho, R. A., de Groot, R. and Hermans, P. W. (2001). "Pneumococcal carriage in children in The Netherlands: a molecular epidemiological study." J Clin Microbiol **39**(9): 3316-20.

Bogaert, D., De Groot, R. and Hermans, P. W. (2004). "Streptococcus pneumoniae colonisation: the key to pneumococcal disease." Lancet Infect Dis **4**(3): 144-54.

Bortoni, M. E., Terra, V. S., Hinds, J., Andrew, P. W. and Yesilkaya, H. (2009). "The pneumococcal response to oxidative stress includes a role for Rgg." Microbiology **155**(Pt 12): 4123-34.

Boulnois, G. J., Paton, J. C., Mitchell, T. J. and Andrew, P. W. (1991). "Structure and function of pneumolysin, the multifunctional, thiol-activated toxin of Streptococcus pneumoniae." Mol Microbiol **5**(11): 2611-6.

Braun, J. S., Novak, R., Gao, G., Murray, P. J. and Shenep, J. L. (1999). "Pneumolysin, a protein toxin of Streptococcus pneumoniae, induces nitric oxide production from macrophages." Infect Immun **67**(8): 3750-6.

Braun, J. S., Sublett, J. E., Freyer, D., Mitchell, T. J., Cleveland, J. L., Tuomanen, E. I. and Weber, J. R. (2002). "Pneumococcal pneumolysin and H(2)O(2) mediate brain cell apoptosis during meningitis." J Clin Invest **109**(1): 19-27.

Brehm, K., Ripio, M. T., Kreft, J. and Vazquez-Boland, J. A. (1999). "The bvr locus of Listeria monocytogenes mediates virulence gene repression by beta-glucosides." J Bacteriol **181**(16): 5024-32.

Briles, D. E., Nahm, M., Schroer, K., Davie, J., Baker, P., Kearney, J. and Barletta, R. (1981). "Antiphosphocholine antibodies found in normal mouse serum are protective against intravenous infection with type 3 streptococcus pneumoniae." J Exp Med **153**(3): 694-705.

Briles, D. E., Hollingshead, S. K., Paton, J. C., Ades, E. W., Novak, L., van Ginkel, F. W. and Benjamin, W. H., Jr. (2003). "Immunizations with pneumococcal surface protein A and pneumolysin are protective against pneumonia in a murine model of pulmonary infection with Streptococcus pneumoniae." J Infect Dis **188**(3): 339-48.

Briles, D. E., Novak, L., Hotomi, M., van Ginkel, F. W. and King, J. (2005). "Nasal colonization with Streptococcus pneumoniae includes subpopulations of surface and invasive pneumococci." Infect Immun **73**(10): 6945-51.

Brown, E. J. (1985). "Interaction of gram-positive microorganisms with complement." Curr Top Microbiol Immunol **121**: 159-87.

Brown, J. S., Gilliland, S. M. and Holden, D. W. (2001). "A Streptococcus pneumoniae pathogenicity island encoding an ABC transporter involved in iron uptake and virulence." Mol Microbiol **40**(3): 572-85.

Brown, J. S., Gilliland, S. M., Ruiz-Albert, J. and Holden, D. W. (2002). "Characterization of pit, a Streptococcus pneumoniae iron uptake ABC transporter." Infect Immun **70**(8): 4389-98.

Brown, J. S., Gilliland, S. M., Spratt, B. G. and Holden, D. W. (2004). "A locus contained within a variable region of pneumococcal pathogenicity island 1 contributes to virulence in mice." Infect Immun **72**(3): 1587-93.

Bruckner, R., Nuhn, M., Reichmann, P., Weber, B. and Hakenbeck, R. (2004). "Mosaic genes and mosaic chromosomes-genomic variation in Streptococcus pneumoniae." Int J Med Microbiol **294**(2-3): 157-68.

Brueggemann, A. B., Griffiths, D. T., Meats, E., Peto, T., Crook, D. W. and Spratt, B. G. (2003). "Clonal relationships between invasive and carriage Streptococcus pneumoniae and serotype- and clone-specific differences in invasive disease potential." J Infect Dis **187**(9): 1424-32.

Brueggemann, A. B. and Spratt, B. G. (2003). "Geographic distribution and clonal diversity of Streptococcus pneumoniae serotype 1 isolates." J Clin Microbiol **41**(11): 4966-70.

Brueggemann, A. B., Peto, T. E., Crook, D. W., Butler, J. C., Kristinsson, K. G. and Spratt, B. G. (2004). "Temporal and geographic stability of the serogroup-specific invasive disease potential of Streptococcus pneumoniae in children." J Infect Dis **190**(7): 1203-11.

Bruyn, G. A., Zegers, B. J. and van Furth, R. (1992). "Mechanisms of host defense against infection with Streptococcus pneumoniae." Clin Infect Dis **14**(1): 251-62.

Butler, J. C. (2004). Epidemiology of pneumococcal disease. The Pneumococcus. Tuomanen, E. Washington, ASM Press**:** 148-168.

Butler, J. R., McIntyre, P., MacIntyre, C. R., Gilmour, R., Howarth, A. L. and Sander, B. (2004). "The cost-effectiveness of pneumococcal conjugate vaccination in Australia." Vaccine **22**(9-10): 1138-49.

Byington, C. L., Hulten, K. G., Ampofo, K., Sheng, X., Pavia, A. T., Blaschke, A. J., Pettigrew, M., Korgenski, K., Daly, J. and Mason, E. O. (2010). "Molecular epidemiology of pediatric pneumococcal empyema from 2001 to 2007 in Utah." J Clin Microbiol **48**(2): 520-5.

Camilli, R., Pettini, E., Del Grosso, M., Pozzi, G., Pantosti, A. and Oggioni, M. R. (2006). "Zinc metalloproteinase genes in clinical isolates of Streptococcus pneumoniae: association of the full array with a clonal cluster comprising serotypes 8 and 11A." Microbiology **152**(Pt 2): 313-21.

Campbell, E. A., Choi, S. Y. and Masure, H. R. (1998). "A competence regulon in Streptococcus pneumoniae revealed by genomic analysis." Mol Microbiol **27**(5): 929-39.

Canvin, J. R., Marvin, A. P., Sivakumaran, M., Paton, J. C., Boulnois, G. J., Andrew, P. W. and Mitchell, T. J. (1995). "The role of pneumolysin and autolysin in the pathology of pneumonia and septicemia in mice infected with a type 2 pneumococcus." J Infect Dis **172**(1): 119-23.

CDC (2010). "Invasive pneumococcal disease in young children before licensure of 13-valent pneumococcal conjugate vaccine - United States, 2007." MMWR Morb Mortal Wkly Rep **59**(9): 253-7.

Chen, J. D. and Morrison, D. A. (1987). "Modulation of competence for genetic transformation in Streptococcus pneumoniae." J Gen Microbiol **133**(7): 1959-67.

Chen, C. J., Lin, C. L., Chen, Y. C., Wang, C. W., Chiu, C. H., Lin, T. Y. and Huang, Y. C. (2009). "Host and microbiologic factors associated with mortality in Taiwanese children with invasive pneumococcal diseases, 2001 to 2006." Diagn Microbiol Infect Dis **63**(2): 194-200.

Chetty, C. and Kreger, A. (1981). "Role of autolysin in generating the pneumococcal purpura-producing principle." Infect Immun **31**(1): 339-44.

Chiavolini, D., Memmi, G., Maggi, T., Iannelli, F., Pozzi, G. and Oggioni, M. R. (2003). "The three extra-cellular zinc metalloproteinases of Streptococcus pneumoniae have a different impact on virulence in mice." BMC Microbiol **3**: 14.

Claverys, J. P. and Havarstein, L. S. (2002). "Extracellular-peptide control of competence for genetic transformation in Streptococcus pneumoniae." Front Biosci **7**: d1798-814.

Claverys, J. P. and Havarstein, L. S. (2007). "Cannibalism and fratricide: mechanisms and raisons d'etre." Nat Rev Microbiol **5**(3): 219-29.

Claverys, J. P., Martin, B. and Polard, P. (2009). "The genetic transformation machinery: composition, localization, and mechanism." FEMS Microbiol Rev **33**(3): 643-56.

Coates, H. L., Morris, P. S., Leach, A. J. and Couzos, S. (2002). "Otitis media in Aboriginal children: tackling a major health problem." Med J Aust **177**(4): 177-8.

Cockeran, R., Theron, A. J., Steel, H. C., Matlola, N. M., Mitchell, T. J., Feldman, C. and Anderson, R. (2001). "Proinflammatory interactions of pneumolysin with human neutrophils." J Infect Dis **183**(4): 604-11.

Coffey, T. J., Enright, M. C., Daniels, M., Morona, J. K., Morona, R., Hryniewicz, W., Paton, J. C. and Spratt, B. G. (1998). "Recombinational exchanges at the capsular polysaccharide biosynthetic locus lead to frequent serotype changes among natural isolates of Streptococcus pneumoniae." Mol Microbiol **27**(1): 73-83.

Colino, J. and Snapper, C. M. (2003). "Two distinct mechanisms for induction of dendritic cell apoptosis in response to intact Streptococcus pneumoniae." J Immunol **171**(5): 2354-65.

Cotton, R. G. and Gibson, F. (1965). "The Biosynthesis of Phenylalanine and Tyrosine; Enzymes Converting Chorismic Acid into Prephenic Acid and Their Relationships to Prephenate Dehydratase and Prephenate Dehydrogenase." Biochim Biophys Acta **100**: 76-88.

Cronan, J. E., Jr. and Waldrop, G. L. (2002). "Multi-subunit acetyl-CoA carboxylases." Prog Lipid Res **41**(5): 407-35.

Croucher, N. J., Walker, D., Romero, P., Lennard, N., Paterson, G. K., Bason, N. C., Mitchell, A. M., Quail, M. A., Andrew, P. W., Parkhill, J., Bentley, S. D. and Mitchell, T. J. (2009). "Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone Streptococcus pneumoniaeSpain23F ST81." J Bacteriol **191**(5): 1480-9.

Cundell, D. R., Gerard, N. P., Gerard, C., Idanpaan-Heikkila, I. and Tuomanen, E. I. (1995a). "Streptococcus pneumoniae anchor to activated human cells by the receptor for platelet-activating factor." Nature **377**(6548): 435-8.

Cundell, D. R., Weiser, J. N., Shen, J., Young, A. and Tuomanen, E. I. (1995b). "Relationship between colonial morphology and adherence of Streptococcus pneumoniae." Infect Immun **63**(3): 757-61.

Dagan, R., Gradstein, S., Belmaker, I., Porat, N., Siton, Y., Weber, G., Janco, J. and Yagupsky, P. (2000). "An outbreak of Streptococcus pneumoniae serotype 1 in a closed community in southern Israel." Clin Infect Dis **30**(2): 319-21.

Dagerhamn, J., Blomberg, C., Browall, S., Sjostrom, K., Morfeldt, E. and Henriques-Normark, B. (2008). "Determination of accessory gene patterns predicts the same relatedness among strains of Streptococcus pneumoniae as sequencing of housekeeping genes does and represents a novel approach in molecular epidemiology." J Clin Microbiol **46**(3): 863-8.

Dalia, A. B., Standish, A. J. and Weiser, J. N. (2010). "Three surface exoglycosidases from Streptococcus pneumoniae, NanA, BgaA, and StrH, promote resistance to opsonophagocytic killing by human neutrophils." Infect Immun.

Dave, S., Brooks-Walter, A., Pangburn, M. K. and McDaniel, L. S. (2001). "PspC, a pneumococcal surface protein, binds human factor H." Infect Immun **69**(5): 3435-7.

Dawid, S., Roche, A. M. and Weiser, J. N. (2007). "The blp bacteriocins of Streptococcus pneumoniae mediate intraspecies competition both in vitro and in vivo." Infect Immun **75**(1): 443-51.

Ding, F., Tang, P., Hsu, M. H., Cui, P., Hu, S., Yu, J. and Chiu, C. H. (2009). "Genome evolution driven by host adaptations results in a more virulent and antimicrobial-resistant Streptococcus pneumoniae serotype 14." BMC Genomics **10**: 158.

Dintilhac, A., Alloing, G., Granadel, C. and Claverys, J. P. (1997). "Competence and virulence of Streptococcus pneumoniae: Adc and PsaA mutants exhibit a requirement for Zn and Mn resulting from inactivation of putative ABC metal permeases." Mol Microbiol **25**(4): 727-39.

Donlan, R. M., Piede, J. A., Heyes, C. D., Sanii, L., Murga, R., Edmonds, P., El-Sayed, I. and El-Sayed, M. A. (2004). "Model system for growing and quantifying Streptococcus pneumoniae biofilms in situ and in real time." Appl Environ Microbiol **70**(8): 4980-8.

Draper, L. A., Ross, R. P., Hill, C. and Cotter, P. D. (2008). "Lantibiotic immunity." Curr Protein Pept Sci **9**(1): 39-49.

Druml, W., Heinzel, G. and Kleinberger, G. (2001). "Amino acid kinetics in patients with sepsis." Am J Clin Nutr **73**(5): 908-13.

Embry, A., Hinojosa, E. and Orihuela, C. J. (2007). "Regions of Diversity 8, 9 and 13 contribute to Streptococcus pneumoniae virulence." BMC Microbiol **7**: 80.

Enright, M. C. and Spratt, B. G. (1998). "A multilocus sequence typing scheme for Streptococcus pneumoniae: identification of clones associated with serious invasive disease." Microbiology **144 (Pt 11)**: 3049-60.

Eskola, J., Kilpi, T., Palmu, A., Jokinen, J., Haapakoski, J., Herva, E., Takala, A., Kayhty, H., Karma, P., Kohberger, R., Siber, G. and Makela, P. H. (2001). "Efficacy of a pneumococcal conjugate vaccine against acute otitis media." N Engl J Med **344**(6): 403-9.

Feldman, C., Mitchell, T. J., Andrew, P. W., Boulnois, G. J., Read, R. C., Todd, H. C., Cole, P. J. and Wilson, R. (1990). "The effect of Streptococcus pneumoniae pneumolysin on human respiratory epithelium in vitro." Microb Pathog **9**(4): 275-84.

Feldman, C., Cockeran, R., Jedrzejas, M. J., Mitchell, T. J. and Anderson, R. (2007). "Hyaluronidase augments pneumolysin-mediated injury to human ciliated epithelium." Int J Infect Dis **11**(1): 11-5.

Felmingham, D., Canton, R. and Jenkins, S. G. (2007). "Regional trends in beta-lactam, macrolide, fluoroquinolone and telithromycin resistance among Streptococcus pneumoniae isolates 2001-2004." J Infect **55**(2): 111-8.

Fiebig, A., Castro Rojas, C. M., Siegal-Gaskins, D. and Crosson, S. (2010). "Interaction specificity, toxicity and regulation of a paralogous set of ParE/RelE-family toxin-antitoxin systems." Mol Microbiol **77**(1): 236-51.

Forrest, J. M., McIntyre, P. B. and Burgess, M. A. (2000). "Pneumococcal disease in Australia." Commun Dis Intell **24**(4): 89-92.

Ghaffar, F., Friedland, I. R. and McCracken, G. H., Jr. (1999). "Dynamics of nasopharyngeal colonization by Streptococcus pneumoniae." Pediatr Infect Dis J **18**(7): 638-46.

Giammarinaro, P., Sicard, M. and Gasc, A. M. (1999). "Genetic and physiological studies of the CiaH-CiaR two-component signal-transducing system involved in cefotaxime resistance and competence of Streptococcus pneumoniae." Microbiology **145 (Pt 8)**: 1859-69.

Giammarinaro, P. and Paton, J. C. (2002). "Role of RegM, a homologue of the catabolite repressor protein CcpA, in the virulence of Streptococcus pneumoniae." Infect Immun **70**(10): 5454-61.

Giefing, C., Meinke, A. L., Hanner, M., Henics, T., Bui, M. D., Gelbmann, D., Lundberg, U., Senn, B. M., Schunn, M., Habel, A., Henriques-Normark, B., Ortqvist, A., Kalin, M., von Gabain, A. and Nagy, E. (2008). "Discovery of a novel class of highly conserved vaccine antigens using genomic scale antigenic fingerprinting of pneumococcus with human antibodies." J Exp Med **205**(1): 117-31.

Giele, C. M., Keil, A. D., Lehmann, D. and Van Buynder, P. G. (2009). "Invasive pneumococcal disease in Western Australia: emergence of serotype 19A." Med J Aust **190**(3): 166.

Gleich, S., Morad, Y., Echague, R., Miller, J. R., Kornblum, J., Sampson, J. S. and Butler, J. C. (2000). "Streptococcus pneumoniae serotype 4 outbreak in a home for the aged: report and review of recent outbreaks." Infect Control Hosp Epidemiol **21**(11): 711-7.

Goldbart, A. D., Leibovitz, E., Porat, N., Givon-Lavi, N., Drukmann, I., Tal, A. and Greenberg, D. (2009). "Complicated community acquired pneumonia in children prior to the introduction of the pneumococcal conjugated vaccine." Scand J Infect Dis **41**(3): 182-7.

Gratten, M., Morey, F., Dixon, J., Manning, K., Torzillo, P., Matters, R., Erlich, J., Hanna, J., Asche, V. and Riley, I. (1993). "An outbreak of serotype 1 Streptococcus pneumoniae infection in central Australia." Med J Aust **158**(5): 340-2.

Gray, B. M., Converse, G. M., 3rd and Dillon, H. C., Jr. (1980). "Epidemiologic studies of Streptococcus pneumoniae in infants: acquisition, carriage, and infection during the first 24 months of life." J Infect Dis **142**(6): 923-33.

Greenwood, B. (1999). "The epidemiology of pneumococcal infection in children in the developing world." Philos Trans R Soc Lond B Biol Sci **354**(1384): 777-85.

Guan, R. and Mariuzza, R. A. (2007). "Peptidoglycan recognition proteins of the innate immune system." Trends Microbiol **15**(3): 127-34.

Guenzi, E., Gasc, A. M., Sicard, M. A. and Hakenbeck, R. (1994). "A two-component signal-transducing system is involved in competence and penicillin susceptibility in laboratory mutants of Streptococcus pneumoniae." Mol Microbiol **12**(3): 505-15.

Gupta, A., Khaw, F. M., Stokle, E. L., George, R. C., Pebody, R., Stansfield, R. E., Sheppard, C. L., Slack, M., Gorton, R. and Spencer, D. A. (2008). "Outbreak of Streptococcus pneumoniae serotype 1 pneumonia in a United Kingdom school." BMJ **337**: a2964.

Halfmann, A., Kovacs, M., Hakenbeck, R. and Bruckner, R. (2007). "Identification of the genes directly controlled by the response regulator CiaR in Streptococcus pneumoniae: five out of 15 promoters drive expression of small non-coding RNAs." Mol Microbiol **66**(1): 110-26.

Hall-Stoodley, L., Hu, F. Z., Gieseke, A., Nistico, L., Nguyen, D., Hayes, J., Forbes, M., Greenberg, D. P., Dice, B., Burrows, A., Wackym, P. A., Stoodley, P., Post, J. C., Ehrlich, G. D. and Kerschner, J. E. (2006). "Direct detection of bacterial biofilms on the middle-ear mucosa of children with chronic otitis media." Jama **296**(2): 202-11.

Hammerschmidt, S., Talay, S. R., Brandtzaeg, P. and Chhatwal, G. S. (1997). "SpsA, a novel pneumococcal surface protein with specific binding to secretory immunoglobulin A and secretory component." Mol Microbiol **25**(6): 1113-24.

Hammerschmidt, S., Wolff, S., Hocke, A., Rosseau, S., Muller, E. and Rohde, M. (2005). "Illustration of pneumococcal polysaccharide capsule during adherence and invasion of epithelial cells." Infect Immun **73**(8): 4653-67.

Hanage, W. P., Kaijalainen, T. H., Syrjanen, R. K., Auranen, K., Leinonen, M., Makela, P. H. and Spratt, B. G. (2005). "Invasiveness of serotypes and clones of Streptococcus pneumoniae among children in Finland." Infect Immun **73**(1): 431-5.

Hanna, J. N., Humphreys, J. L. and Murphy, D. M. (2008). "Invasive pneumococcal disease in Indigenous people in north Queensland: an update, 2005-2007." Med J Aust **189**(1): 43-6.

Hardy, G. G., Magee, A. D., Ventura, C. L., Caimano, M. J. and Yother, J. (2001). "Essential role for cellular phosphoglucomutase in virulence of type 3 Streptococcus pneumoniae." Infect Immun **69**(4): 2309-17.

Harvey, R. M. (2006). Genetic characterisation of Streptococcus pneumoniae type 1 isolates in relation to invasiveness. School of Molecular and Biomedical Science. Adelaide, University of Adelaide. **Bachelor of Science (Honours)**.

Hausdorff, W. P., Bryant, J., Paradiso, P. R. and Siber, G. R. (2000a). "Which pneumococcal serogroups cause the most invasive disease: implications for conjugate vaccine formulation and use, part I." Clin Infect Dis **30**(1): 100-21.

Hausdorff, W. P., Bryant, J., Kloek, C., Paradiso, P. R. and Siber, G. R. (2000b). "The contribution of specific pneumococcal serogroups to different disease manifestations: implications for conjugate vaccine formulation and use, part II." Clin Infect Dis **30**(1): 122-40.

Hausdorff, W. P., Feikin, D. R. and Klugman, K. P. (2005). "Epidemiological differences among pneumococcal serotypes." Lancet Infect Dis **5**(2): 83-93.

Hava, D. L. and Camilli, A. (2002). "Large-scale identification of serotype 4 Streptococcus pneumoniae virulence factors." Mol Microbiol **45**(5): 1389-406.

Havarstein, L. S., Coomaraswamy, G. and Morrison, D. A. (1995). "An unmodified heptadecapeptide pheromone induces competence for genetic transformation in Streptococcus pneumoniae." Proc Natl Acad Sci U S A **92**(24): 11140-4.

Havarstein, L. S., Gaustad, P., Nes, I. F. and Morrison, D. A. (1996). "Identification of the streptococcal competence-pheromone receptor." Mol Microbiol **21**(4): 863-9.

Havarstein, L. S., Martin, B., Johnsborg, O., Granadel, C. and Claverys, J. P. (2006). "New insights into the pneumococcal fratricide: relationship to clumping and identification of a novel immunity factor." Mol Microbiol **59**(4): 1297-307.

Henderson-Begg, S. K., Roberts, A. P. and Hall, L. M. (2009). "Diversity of putative Tn5253-like elements in Streptococcus pneumoniae." Int J Antimicrob Agents **33**(4): 364-7.

Henrichsen, J. (1995). "Six newly recognized types of Streptococcus pneumoniae." J Clin Microbiol **33**(10): 2759-62.

Henriques-Normark, B., Blomberg, C., Dagerhamn, J., Battig, P. and Normark, S. (2008). "The rise and fall of bacterial clones: Streptococcus pneumoniae." Nat Rev Microbiol **6**(11): 827-37.

Henrissat, B. (1991). "A classification of glycosyl hydrolases based on amino acid sequence similarities." Biochem J **280 (Pt 2)**: 309-16.

Hernandez-Bou, S., Garcia-Garcia, J. J., Esteva, C., Gene, A., Luaces, C. and Munoz Almagro, C. (2009). "Pediatric parapneumonic pleural effusion: epidemiology, clinical characteristics, and microbiological diagnosis." Pediatr Pulmonol **44**(12): 1192-200.

Hiller, N. L., Janto, B., Hogg, J. S., Boissy, R., Yu, S., Powell, E., Keefe, R., Ehrlich, N. E., Shen, K., Hayes, J., Barbadora, K., Klimke, W., Dernovoy, D., Tatusova, T., Parkhill, J., Bentley, S. D., Post, J. C., Ehrlich, G. D. and Hu, F. Z. (2007). "Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome." J Bacteriol **189**(22): 8186-95.

Hollands, A., Aziz, R. K., Kansal, R., Kotb, M., Nizet, V. and Walker, M. J. (2008). "A naturally occurring mutation in ropB suppresses SpeB expression and reduces M1T1 group A streptococcal systemic virulence." PLoS One **3**(12): e4102.

Hollingshead, S. K., Becker, R. and Briles, D. E. (2000). "Diversity of PspA: mosaic genes and evidence for past recombination in Streptococcus pneumoniae." Infect Immun **68**(10): 5889-900.

Holmes, A. R., McNab, R., Millsap, K. W., Rohde, M., Hammerschmidt, S., Mawdsley, J. L. and Jenkinson, H. F. (2001). "The pavA gene of Streptococcus pneumoniae encodes a fibronectin-binding protein that is essential for virulence." Mol Microbiol **41**(6): 1395-408.

Hondoh, H., Kuriki, T. and Matsuura, Y. (2003). "Three-dimensional structure and substrate binding of Bacillus stearothermophilus neopullulanase." J Mol Biol **326**(1): 177-88.

Hoskins, J., Alborn, W. E., Jr., Arnold, J., Blaszczak, L. C., Burgett, S., DeHoff, B. S., Estrem, S. T., Fritz, L., Fu, D. J., Fuller, W., Geringer, C., Gilmour, R., Glass, J. S., Khoja, H., Kraft, A. R., Lagace, R. E., LeBlanc, D. J., Lee, L. N., Lefkowitz, E. J., Lu, J., Matsushima, P., McAhren, S. M., McHenney, M., McLeaster, K., Mundy, C. W., Nicas, T. I., Norris, F. H., O'Gara, M., Peery, R. B., Robertson, G. T., Rockey, P., Sun, P. M., Winkler, M. E., Yang, Y., Young-Bellido, M., Zhao, G., Zook, C. A., Baltz, R. H., Jaskunas, S. R., Rosteck, P. R., Jr., Skatrud, P. L. and Glass, J. I. (2001). "Genome of the bacterium Streptococcus pneumoniae strain R6." J Bacteriol **183**(19): 5709-17.

Houldsworth, S., Andrew, P. W. and Mitchell, T. J. (1994). "Pneumolysin stimulates production of tumor necrosis factor alpha and interleukin-1 beta by human mononuclear phagocytes." Infect Immun **62**(4): 1501-3.

Hsieh, Y. C., Tsao, P. N., Chen, C. L., Lin, T. L., Lee, W. S., Shao, P. L., Lee, C. Y., Hsueh, P. R., Huang, L. M. and Wang, J. T. (2008). "Establishment of a young mouse model and identification of an allelic variation of zmpB in complicated pneumonia caused by Streptococcus pneumoniae." Crit Care Med **36**(4): 1248-55.

Hsu, K. K., Shea, K. M., Stevenson, A. E. and Pelton, S. I. (2010). "Changing serotypes causing childhood invasive pneumococcal disease: Massachusetts, 2001-2007." Pediatr Infect Dis J **29**(4): 289-93.

Huh, J. W., Shima, J. and Ochi, K. (1996). "ADP-ribosylation of proteins in Bacillus subtilis and its possible importance in sporulation." J Bacteriol **178**(16): 4935-41.

Hyams, C., Yuste, J., Bax, K., Camberlein, E., Weiser, J. N. and Brown, J. S. (2010). "Streptococcus pneumoniae resistance to complement-mediated immunity is dependent on the capsular serotype." Infect Immun **78**(2): 716-25.

Hyde, S. C., Emsley, P., Hartshorn, M. J., Mimmack, M. M., Gileadi, U., Pearce, S. R., Gallagher, M. P., Gill, D. R., Hubbard, R. E. and Higgins, C. F. (1990). "Structural model of ATP-binding proteins associated with cystic fibrosis, multidrug resistance and bacterial transport." Nature **346**(6282): 362-5.

Iannelli, F., Oggioni, M. R. and Pozzi, G. (2002). "Allelic variation in the highly polymorphic locus pspC of Streptococcus pneumoniae." Gene **284**(1-2): 63-71.

Iannelli, F. and Pozzi, G. (2004). "Method for introducing specific and unmarked mutations into the chromosome of Streptococcus pneumoniae." Mol Biotechnol **26**(1): 81-6.

Iannelli, F., Oggioni, M. R. and Pozzi, G. (2005). "Sensor domain of histidine kinase ComD confers competence pherotype specificity in Streptoccoccus pneumoniae." FEMS Microbiol Lett **252**(2): 321-6.

Ibrahim, Y. M., Kerr, A. R., McCluskey, J. and Mitchell, T. J. (2004). "Role of HtrA in the virulence and competence of Streptococcus pneumoniae." Infect Immun **72**(6): 3584-91.

Ishii, S., Yano, T. and Hayashi, H. (2006). "Expression and characterization of the peptidase domain of Streptococcus pneumoniae ComA, a bifunctional ATP-binding cassette transporter involved in quorum sensing pathway." J Biol Chem **281**(8): 4726-31.

Ishino, F., Jung, H. K., Ikeda, M., Doi, M., Wachi, M. and Matsuhashi, M. (1989). "New mutations fts-36, lts-33, and ftsW clustered in the mra region of the Escherichia coli chromosome induce thermosensitive cell growth and division." J Bacteriol **171**(10): 5523-30.

Iyer, R., Baliga, N. S. and Camilli, A. (2005). "Catabolite control protein A (CcpA) contributes to virulence and regulation of sugar metabolism in Streptococcus pneumoniae." J Bacteriol **187**(24): 8340-9.

Jedrzejas, M. J. (2000). "Structural and functional comparison of polysaccharide-degrading enzymes." Crit Rev Biochem Mol Biol **35**(3): 221-51.

Jedrzejas, M. J. (2004). "Extracellular virulence factors of Streptococcus pneumoniae." Front Biosci **9**: 891-914.

Jefferies, J. M., Johnston, C. H., Kirkham, L. A., Cowan, G. J., Ross, K. S., Smith, A., Clarke, S. C., Brueggemann, A. B., George, R. C., Pichon, B., Pluschke, G., Pfluger, V. and Mitchell, T. J. (2007). "Presence of nonhemolytic pneumolysin in serotypes of Streptococcus pneumoniae associated with disease outbreaks." J Infect Dis **196**(6): 936-44.

Johnsborg, O. and Havarstein, L. S. (2009). "Regulation of natural genetic transformation and acquisition of transforming DNA in Streptococcus pneumoniae." FEMS Microbiol Rev **33**(3): 627-42.

Johnson, M. K., Boese-Marrazzo, D. and Pierce, W. A., Jr. (1981). "Effects of pneumolysin on human polymorphonuclear leukocytes and platelets." Infect Immun **34**(1): 171-6.

Kadioglu, A., Gingles, N. A., Grattan, K., Kerr, A., Mitchell, T. J. and Andrew, P. W. (2000). "Host cellular immune response to pneumococcal lung infection in mice." Infect Immun **68**(2): 492-501.

Kadioglu, A. and Andrew, P. W. (2004). "The innate immune response to pneumococcal lung infection: the untold story." Trends Immunol **25**(3): 143-9.

Kazmierczak, K. M., Wayne, K. J., Rechtsteiner, A. and Winkler, M. E. (2009). "Roles of rel in stringent response, global regulation and virulence of serotype 2 Streptococcus pneumoniae D39." Mol Microbiol **72**(3): 590-611.

Kelly, T., Dillard, J. P. and Yother, J. (1994). "Effect of genetic switching of capsular type on virulence of Streptococcus pneumoniae." Infect Immun **62**(5): 1813-9.

Kenzel, S. and Henneke, P. (2006). "The innate immune system and its relevance to neonatal sepsis." Curr Opin Infect Dis **19**(3): 264-70.

Khoo, S. K., Loll, B., Chan, W. T., Shoeman, R. L., Ngoo, L., Yeo, C. C. and Meinhart, A. (2007). "Molecular and structural characterization of the PezAT chromosomal toxin-antitoxin system of the human pathogen Streptococcus pneumoniae." J Biol Chem **282**(27): 19606-18.

Kilian, M., Mestecky, J. and Schrohenloher, R. E. (1979). "Pathogenic species of the genus Haemophilus and Streptococcus pneumoniae produce immunoglobulin A1 protease." Infect Immun **26**(1): 143-9.

Kim, J. O. and Weiser, J. N. (1998). "Association of intrastrain phase variation in quantity of capsular polysaccharide and teichoic acid with the virulence of Streptococcus pneumoniae." J Infect Dis **177**(2): 368-77.

Kim, J. O., Romero-Steiner, S., Sorensen, U. B., Blom, J., Carvalho, M., Barnard, S., Carlone, G. and Weiser, J. N. (1999). "Relationship between cell surface carbohydrates and intrastrain variation on opsonophagocytosis of Streptococcus pneumoniae." Infect Immun **67**(5): 2327-33.

King, S. J., Hippe, K. R., Gould, J. M., Bae, D., Peterson, S., Cline, R. T., Fasching, C., Janoff, E. N. and Weiser, J. N. (2004). "Phase variable desialylation of host proteins that bind to Streptococcus pneumoniae in vivo and protect the airway." Mol Microbiol **54**(1): 159-71.

King, S. J., Whatmore, A. M. and Dowson, C. G. (2005). "NanA, a neuraminidase from Streptococcus pneumoniae, shows high levels of sequence diversity, at least in part through recombination with Streptococcus oralis." J Bacteriol **187**(15): 5376-86.

King, S. J., Hippe, K. R. and Weiser, J. N. (2006). "Deglycosylation of human glycoconjugates by the sequential activities of exoglycosidases expressed by Streptococcus pneumoniae." Mol Microbiol **59**(3): 961-74.

Kirkham, L. A., Jefferies, J. M., Kerr, A. R., Jing, Y., Clarke, S. C., Smith, A. and Mitchell, T. J. (2006). "Identification of invasive serotype 1 pneumococcal isolates that express nonhemolytic pneumolysin." J Clin Microbiol **44**(1): 151-9.

Kruszewska, D., Sahl, H. G., Bierbaum, G., Pag, U., Hynes, S. O. and Ljungh, A. (2004). "Mersacidin eradicates methicillin-resistant Staphylococcus aureus (MRSA) in a mouse rhinitis model." J Antimicrob Chemother **54**(3): 648-53.

Lacks, S. A. (1997). "Cloning and expression of pneumococcal genes in Streptococcus pneumoniae." Microb Drug Resist **3**(4): 327-37.

Lagos, R., Munoz, A., San Martin, O., Maldonado, A., Hormazabal, J. C., Blackwelder, W. C. and Levine, M. M. (2008). "Age- and serotype-specific pediatric invasive pneumococcal disease: insights from systematic surveillance in Santiago, Chile, 1994--2007." J Infect Dis **198**(12): 1809-17.

Lau, G. W., Haataja, S., Lonetto, M., Kensit, S. E., Marra, A., Bryant, A. P., McDevitt, D., Morrison, D. A. and Holden, D. W. (2001). "A functional genomic analysis of type 3 Streptococcus pneumoniae virulence." Mol Microbiol **40**(3): 555-71.

Law, C. J., Enkavi, G., Wang, D. N. and Tajkhorshid, E. (2009). "Structural basis of substrate selectivity in the glycerol-3-phosphate: phosphate antiporter GlpT." Biophys J **97**(5): 1346-53.

Lee, M. S. and Morrison, D. A. (1999). "Identification of a new regulator in Streptococcus pneumoniae linking quorum sensing to competence for genetic transformation." J Bacteriol **181**(16): 5004-16.

Lefrancois, J. and Sicard, A. M. (1997). "Electrotransformation of Streptococcus pneumoniae: evidence for restriction of DNA on entry." Microbiology **143 (Pt 2)**: 523-6.

Lefrancois, J., Samrakandi, M. M. and Sicard, A. M. (1998). "Electrotransformation and natural transformation of Streptococcus pneumoniae: requirement of DNA processing for recombination." Microbiology **144 (Pt 11)**: 3061-8.

Leimkugel, J., Adams Forgor, A., Gagneux, S., Pfluger, V., Flierl, C., Awine, E., Naegeli, M., Dangy, J. P., Smith, T., Hodgson, A. and Pluschke, G. (2005). "An outbreak of serotype 1 Streptococcus pneumoniae meningitis in northern Ghana with features that are characteristic of Neisseria meningitidis meningitis epidemics." J Infect Dis **192**(2): 192-9.

LeMessurier, K. S., Ogunniyi, A. D. and Paton, J. C. (2006). "Differential expression of key pneumococcal virulence genes in vivo." Microbiology **152**(Pt 2): 305-11.

Letto, J., Brosnan, M. E. and Brosnan, J. T. (1986). "Valine metabolism. Gluconeogenesis from 3-hydroxyisobutyrate." Biochem J **240**(3): 909-12.

Li, S., Kelly, S. J., Lamani, E., Ferraroni, M. and Jedrzejas, M. J. (2000). "Structural basis of hyaluronan degradation by Streptococcus pneumoniae hyaluronate lyase." EMBO J **19**(6): 1228-40.

Linder, T. E., Lim, D. J. and DeMaria, T. F. (1992). "Changes in the structure of the cell surface carbohydrates of the chinchilla tubotympanum following Streptococcus pneumoniae-induced otitis media." Microb Pathog **13**(4): 293-303.

Linder, T. E., Daniels, R. L., Lim, D. J. and DeMaria, T. F. (1994). "Effect of intranasal inoculation of Streptococcus pneumoniae on the structure of the surface carbohydrates of the chinchilla eustachian tube and middle ear mucosa." Microb Pathog **16**(6): 435-41.

Littmann, M., Albiger, B., Frentzen, A., Normark, S., Henriques-Normark, B. and Plant, L. (2009). "Streptococcus pneumoniae evades human dendritic cell surveillance by pneumolysin expression." EMBO Mol Med **1**(4): 211-22.

Livak, K. J. and Schmittgen, T. D. (2001). "Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method." Methods **25**(4): 402-8.

Lizcano, A., Chin, T., Sauer, K., Tuomanen, E. I. and Orihuela, C. J. (2010). "Early biofilm formation on microtiter plates is not correlated with the invasive disease potential of Streptococcus pneumoniae." Microb Pathog **48**(3-4): 124-30.

Lock, R. A., Zhang, Q. Y., Berry, A. M. and Paton, J. C. (1996). "Sequence variation in the Streptococcus pneumoniae pneumolysin gene affecting haemolytic activity and electrophoretic mobility of the toxin." Microb Pathog **21**(2): 71-83.

Loessner, M. J. (2005). "Bacteriophage endolysins--current state of research and applications." Curr Opin Microbiol **8**(4): 480-7.

Ludden, P. W. (1994). "Reversible ADP-ribosylation as a mechanism of enzyme regulation in procaryotes." Mol Cell Biochem **138**(1-2): 123-9.

Luo, P., Li, H. and Morrison, D. A. (2004). "Identification of ComW as a new component in the regulation of genetic transformation in Streptococcus pneumoniae." Mol Microbiol **54**(1): 172-83.

Lysenko, E. S., Ratner, A. J., Nelson, A. L. and Weiser, J. N. (2005). "The role of innate immune responses in the outcome of interspecies competition for colonization of mucosal surfaces." PLoS Pathog **1**(1): e1.

Mackenzie, G. A., Carapetis, J. R., Leach, A. J. and Morris, P. S. (2009). "Pneumococcal vaccination and otitis media in Australian Aboriginal infants: comparison of two birth cohorts before and after introduction of vaccination." BMC Pediatr **9**: 14.

MacLeod, C. M. and Kraus, M. R. (1950). "Relation of virulence of pneumococcal strains for mice to the quantity of capsular polysaccharide formed in vitro." J Exp Med **92**(1): 1-9.

Macrina, F. L., Evans, R. P., Tobian, J. A., Hartley, D. L., Clewell, D. B. and Jones, K. R. (1983). "Novel shuttle plasmid vehicles for Escherichia-Streptococcus transgeneric cloning." Gene **25**(1): 145-50.

Magee, A. D. and Yother, J. (2001). "Requirement for capsule in colonization by Streptococcus pneumoniae." Infect Immun **69**(6): 3755-61.

Mahdi, L. K., Ogunniyi, A. D., LeMessurier, K. S. and Paton, J. C. (2008). "Pneumococcal virulence gene expression and host cytokine profiles during pathogenesis of invasive disease." Infect Immun **76**(2): 646-57.

Malley, R., Henneke, P., Morse, S. C., Cieslewicz, M. J., Lipsitch, M., Thompson, C. M., Kurt-Jones, E., Paton, J. C., Wessels, M. R. and Golenbock, D. T. (2003). "Recognition of pneumolysin by Toll-like receptor 4 confers resistance to pneumococcal infection." Proc Natl Acad Sci U S A **100**(4): 1966-71.

Mangtani, P., Cutts, F. and Hall, A. J. (2003). "Efficacy of polysaccharide pneumococcal vaccine in adults in more developed countries: the state of the evidence." Lancet Infect Dis **3**(2): 71-8.

Margolis, E. and Levin, B. R. (2007). "Within-host evolution for the invasiveness of commensal bacteria: an experimental study of bacteremias resulting from Haemophilus influenzae nasal carriage." J Infect Dis **196**(7): 1068-75.

Margolis, E., Yates, A. and Levin, B. R. (2010). "The ecology of nasal colonization of Streptococcus pneumoniae, Haemophilus influenzae and Staphylococcus aureus: the role of competition and interactions with host's immune response." BMC Microbiol **10**: 59.

Marion, C., Limoli, D. H., Bobulsky, G. S., Abraham, J. L., Burnaugh, A. M. and King, S. J. (2009). "Identification of a pneumococcal glycosidase that modifies O-linked glycans." Infect Immun **77**(4): 1389-96.

Martin, B., Garcia, P., Castanie, M. P., Glise, B. and Claverys, J. P. (1995). "The recA gene of Streptococcus pneumoniae is part of a competence-induced operon and controls an SOS regulon." Dev Biol Stand **85**: 293-300.

Martner, A., Dahlgren, C., Paton, J. C. and Wold, A. E. (2008). "Pneumolysin released during Streptococcus pneumoniae autolysis is a potent activator of intracellular oxygen radical production in neutrophils." Infect Immun **76**(9): 4079-87.

Martner, A., Skovbjerg, S., Paton, J. C. and Wold, A. E. (2009). "Streptococcus pneumoniae autolysis prevents phagocytosis and production of phagocyte-activating cytokines." Infect Immun **77**(9): 3826-37.

Mascher, T., Zahner, D., Merai, M., Balmelle, N., de Saizieu, A. B. and Hakenbeck, R. (2003). "The Streptococcus pneumoniae cia regulon: CiaR target sites and transcription profile analysis." J Bacteriol **185**(1): 60-70.

Masepohl, B., Krey, R. and Klipp, W. (1993). "The draTG gene region of Rhodobacter capsulatus is required for post-translational regulation of both the molybdenum and the alternative nitrogenase." J Gen Microbiol **139**(11): 2667-75.

McAllister, L. J., Tseng, H. J., Ogunniyi, A. D., Jennings, M. P., McEwan, A. G. and Paton, J. C. (2004). "Molecular analysis of the psa permease complex of Streptococcus pneumoniae." Mol Microbiol **53**(3): 889-901.

McCormick, A. W., Whitney, C. G., Farley, M. M., Lynfield, R., Harrison, L. H., Bennett, N. M., Schaffner, W., Reingold, A., Hadler, J., Cieslak, P., Samore, M. H. and Lipsitch, M. (2003). "Geographic diversity and temporal trends of antimicrobial resistance in Streptococcus pneumoniae in the United States." Nat Med **9**(4): 424-30.

McCullers, J. A. and Tuomanen, E. I. (2001). "Molecular pathogenesis of pneumococcal pneumonia." Front Biosci **6**: D877-89.

McCullers, J. A. (2006). "Insights into the interaction between influenza virus and pneumococcus." Clin Microbiol Rev **19**(3): 571-82.

McDaniel, L. S., Yother, J., Vijayakumar, M., McGarry, L., Guild, W. R. and Briles, D. E. (1987). "Use of insertional inactivation to facilitate studies of biological properties of pneumococcal surface protein A (PspA)." J Exp Med **165**(2): 381-94.

McEllistrem, M. C., Adams, J. M., Patel, K., Mendelsohn, A. B., Kaplan, S. L., Bradley, J. S., Schutze, G. E., Kim, K. S., Mason, E. O. and Wald, E. R. (2005). "Acute otitis media due to penicillin-nonsusceptible Streptococcus pneumoniae before and after the introduction of the pneumococcal conjugate vaccine." Clin Infect Dis **40**(12): 1738-44.

McEllistrem, M. C., Ransford, J. V. and Khan, S. A. (2007). "Characterization of in vitro biofilm-associated pneumococcal phase variants of a clinically relevant serotype 3 clone." J Clin Microbiol **45**(1): 97-101.

McLeod, J. W. and Gordon, J. (1922). "Production of Hydrogen Peroxide by Bacteria." Biochem J **16**(4): 499-506.

Mehiri-Zghal, E., Decousser, J. W., Mahjoubi, W., Essalah, L., El Marzouk, N., Ghariani, A., Allouch, P. and Slim-Saidi, N. L. (2009). "Molecular epidemiology of a Streptococcus pneumoniae serotype 1 outbreak in a Tunisian jail." Diagn Microbiol Infect Dis **66**(2): 225-7.

Melin, M., Di Paolo, E., Tikkanen, L., Jarva, H., Neyt, C., Kayhty, H., Meri, S., Poolman, J. and Vakevainen, M. (2010). "Interaction of pneumococcal histidine triad proteins with human complement." Infect Immun **78**(5): 2089-98.

Mercat, A., Nguyen, J. and Dautzenberg, B. (1991). "An outbreak of pneumococcal pneumonia in two men's shelters." Chest **99**(1): 147-51.

Meyers, L. A., Levin, B. R., Richardson, A. R. and Stojiljkovic, I. (2003). "Epidemiology, hypermutation, within-host evolution and the virulence of Neisseria meningitidis." Proc Biol Sci **270**(1525): 1667-77.

Mitchell, T. J., Andrew, P. W., Saunders, F. K., Smith, A. N. and Boulnois, G. J. (1991). "Complement activation and antibody binding by pneumolysin via a region of the toxin homologous to a human acute-phase protein." Mol Microbiol **5**(8): 1883-8.

Mitchell, L., Smith, S. H., Braun, J. S., Herzog, K. H., Weber, J. R. and Tuomanen, E. I. (2004). "Dual phases of apoptosis in pneumococcal meningitis." J Infect Dis **190**(11): 2039-46.

Mitchell, J., Siboo, I. R., Takamatsu, D., Chambers, H. F. and Sullam, P. M. (2007). "Mechanism of cell surface expression of the Streptococcus mitis platelet binding proteins PblA and PblB." Mol Microbiol **64**(3): 844-57.

Mitchell, A. M. and Mitchell, T. J. (2010). "Streptococcus pneumoniae: virulence factors and variation." Clin Microbiol Infect **16**(5): 411-8.

Mond, J. J., Vos, Q., Lees, A. and Snapper, C. M. (1995). "T cell independent antigens." Curr Opin Immunol **7**(3): 349-54.

Morens, D. M., Taubenberger, J. K. and Fauci, A. S. (2008). "Predominant role of bacterial pneumonia as a cause of death in pandemic influenza: implications for pandemic influenza preparedness." J Infect Dis **198**(7): 962-70.

Morgan, P. J., Hyman, S. C., Rowe, A. J., Mitchell, T. J., Andrew, P. W. and Saibil, H. R. (1995). "Subunit organisation and symmetry of pore-forming, oligomeric pneumolysin." FEBS Lett **371**(1): 77-80.

Morris, P. S., Leach, A. J., Silberberg, P., Mellon, G., Wilson, C., Hamilton, E. and Beissbarth, J. (2005). "Otitis media in young Aboriginal children from remote communities in Northern and Central Australia: a cross-sectional survey." BMC Pediatr **5**: 27.

Moschioni, M., Donati, C., Muzzi, A., Masignani, V., Censini, S., Hanage, W. P., Bishop, C. J., Reis, J. N., Normark, S., Henriques-Normark, B., Covacci, A., Rappuoli, R. and Barocchi, M. A. (2008). "Streptococcus pneumoniae contains 3 rlrA pilus variants that are clonally related." J Infect Dis **197**(6): 888-96.

Moxon, E. R. and Murphy, P. A. (1978). "Haemophilus influenzae bacteremia and meningitis resulting from survival of a single organism." Proc Natl Acad Sci U S A **75**(3): 1534-6.

Moxon, R., Bayliss, C. and Hood, D. (2006). "Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation." Annu Rev Genet **40**: 307-33.

Munoz-Najar, U. and Vijayakumar, M. N. (1999). "An operon that confers UV resistance by evoking the SOS mutagenic response in streptococcal conjugative transposon Tn5252." J Bacteriol **181**(9): 2782-8.

Muramatsu, H., Tachikui, H., Ushida, H., Song, X., Qiu, Y., Yamamoto, S. and Muramatsu, T. (2001). "Molecular cloning and expression of endo-beta-N-acetylglucosaminidase D, which acts on the core structure of complex type asparagine-linked oligosaccharides." J Biochem **129**(6): 923-8.

Murin, R., Schaer, A., Kowtharapu, B. S., Verleysdonk, S. and Hamprecht, B. (2008). "Expression of 3-hydroxyisobutyrate dehydrogenase in cultured neural cells." J Neurochem **105**(4): 1176-86.

Musher, D. M., Groover, J. E., Watson, D. A., Rodriguez-Barradas, M. C. and Baughn, R. E. (1998). "IgG responses to protein-conjugated pneumococcal capsular polysaccharides in persons who are genetically incapable of responding to unconjugated polysaccharides." Clin Infect Dis **27**(6): 1487-90.

Nelson, A. L., Roche, A. M., Gould, J. M., Chim, K., Ratner, A. J. and Weiser, J. N. (2007). "Capsule enhances pneumococcal colonization by limiting mucus-mediated clearance." Infect Immun **75**(1): 83-90.

Nesin, M., Ramirez, M. and Tomasz, A. (1998). "Capsular transformation of a multidrug-resistant Streptococcus pneumoniae in vivo." J Infect Dis **177**(3): 707-13.

Nunes, S., Sa-Leao, R., Pereira, L. C. and Lencastre, H. (2008). "Emergence of a serotype 1 Streptococcus pneumoniae lineage colonising healthy children in Portugal in the seven-valent conjugate vaccination era." Clin Microbiol Infect **14**(1): 82-4.

Obert, C., Sublett, J., Kaushal, D., Hinojosa, E., Barton, T., Tuomanen, E. I. and Orihuela, C. J. (2006). "Identification of a Candidate Streptococcus pneumoniae core genome and regions of diversity correlated with invasive pneumococcal disease." Infect Immun **74**(8): 4766-77.

O'Brien, K. L., Wolfson, L. J., Watt, J. P., Henkle, E., Deloria-Knoll, M., McCall, N., Lee, E., Mulholland, K., Levine, O. S. and Cherian, T. (2009). "Burden of disease caused by Streptococcus pneumoniae in children younger than 5 years: global estimates." Lancet **374**(9693): 893-902.

Oggioni, M. R., Memmi, G., Maggi, T., Chiavolini, D., Iannelli, F. and Pozzi, G. (2003). "Pneumococcal zinc metalloproteinase ZmpC cleaves human matrix metalloproteinase 9 and is a virulence factor in experimental pneumonia." Mol Microbiol **49**(3): 795-805.

Oggioni, M. R., Trappetti, C., Kadioglu, A., Cassone, M., Iannelli, F., Ricci, S., Andrew, P. W. and Pozzi, G. (2006). "Switch from planktonic to sessile life: a major event in pneumococcal pathogenesis." Mol Microbiol **61**(5): 1196-210.

Ogunniyi, A. D., Giammarinaro, P. and Paton, J. C. (2002). "The genes encoding virulence-associated proteins and the capsule of Streptococcus pneumoniae are upregulated and differentially expressed in vivo." Microbiology **148**(Pt 7): 2045-53.

Ogunniyi, A. D., LeMessurier, K. S., Graham, R. M., Watt, J. M., Briles, D. E., Stroeher, U. H. and Paton, J. C. (2007). "Contributions of pneumolysin, pneumococcal surface protein A (PspA), and PspC to pathogenicity of Streptococcus pneumoniae D39 in a mouse model." Infect Immun **75**(4): 1843-51.

Ogunniyi, A. D., Grabowicz, M., Mahdi, L. K., Cook, J., Gordon, D. L., Sadlon, T. A. and Paton, J. C. (2009). "Pneumococcal histidine triad proteins are regulated by the Zn2+-dependent repressor AdcR and inhibit complement deposition through the recruitment of complement factor H." Faseb J **23**(3): 731-8.

Orihuela, C. J. (2009). "Role played by psrP-secY2A2 (accessory region 34) in the invasive disease potential of Streptococcus pneumoniae." J Infect Dis **200**(7): 1180-1; author reply 1181-2.

Overweg, K., Pericone, C. D., Verhoef, G. G., Weiser, J. N., Meiring, H. D., De Jong, A. P., De Groot, R. and Hermans, P. W. (2000). "Differential protein expression in phenotypic variants of Streptococcus pneumoniae." Infect Immun **68**(8): 4604-10.

Pallen, M. J. and Wren, B. W. (2007). "Bacterial pathogenomics." Nature **449**(7164): 835-42.

Pao, S. S., Paulsen, I. T. and Saier, M. H., Jr. (1998). "Major facilitator superfamily." Microbiol Mol Biol Rev **62**(1): 1-34.

Parker, D., Soong, G., Planet, P., Brower, J., Ratner, A. J. and Prince, A. (2009). "The NanA neuraminidase of Streptococcus pneumoniae is involved in biofilm formation." Infect Immun **77**(9): 3722-30.

Paton, J. C., Rowan-Kelly, B. and Ferrante, A. (1984). "Activation of human complement by the pneumococcal toxin pneumolysin." Infect Immun **43**(3): 1085-7.

Pericone, C. D., Overweg, K., Hermans, P. W. and Weiser, J. N. (2000). "Inhibitory and bactericidal effects of hydrogen peroxide production by Streptococcus pneumoniae on other inhabitants of the upper respiratory tract." Infect Immun **68**(7): 3990-7.

Pestova, E. V., Havarstein, L. S. and Morrison, D. A. (1996). "Regulation of competence for genetic transformation in Streptococcus pneumoniae by an auto-induced peptide pheromone and a two-component regulatory system." Mol Microbiol **21**(4): 853-62.

Pestova, E. V. and Morrison, D. A. (1998). "Isolation and characterization of three Streptococcus pneumoniae transformation-specific loci by use of a lacZ reporter insertion vector." J Bacteriol **180**(10): 2701-10.

Peterson, S. N., Sung, C. K., Cline, R., Desai, B. V., Snesrud, E. C., Luo, P., Walling, J., Li, H., Mintz, M., Tsegaye, G., Burr, P. C., Do, Y., Ahn, S., Gilbert, J., Fleischmann, R. D. and Morrison, D. A. (2004). "Identification of competence pheromone responsive genes in Streptococcus pneumoniae by use of DNA microarrays." Mol Microbiol **51**(4): 1051-70.

Pettigrew, M. M., Fennie, K. P., York, M. P., Daniels, J. and Ghaffar, F. (2006). "Variation in the presence of neuraminidase genes among Streptococcus pneumoniae isolates with identical sequence types." Infect Immun **74**(6): 3360-5.

Pilishvili, T., Lexau, C., Farley, M. M., Hadler, J., Harrison, L. H., Bennett, N. M., Reingold, A., Thomas, A., Schaffner, W., Craig, A. S., Smith, P. J., Beall, B. W., Whitney, C. G. and Moore, M. R. (2010). "Sustained reductions in invasive pneumococcal disease in the era of conjugate vaccine." J Infect Dis **201**(1): 32-41.

Plumbridge, J. and Vimr, E. (1999). "Convergent pathways for utilization of the amino sugars N-acetylglucosamine, N-acetylmannosamine, and N-acetylneuraminic acid by Escherichia coli." J Bacteriol **181**(1): 47-54.

Polissi, A., Pontiggia, A., Feger, G., Altieri, M., Mottl, H., Ferrari, L. and Simon, D. (1998). "Large-scale identification of virulence genes from Streptococcus pneumoniae." Infect Immun **66**(12): 5620-9.

Pozzi, G., Masala, L., Iannelli, F., Manganelli, R., Havarstein, L. S., Piccoli, L., Simon, D. and Morrison, D. A. (1996). "Competence for genetic transformation in encapsulated strains of Streptococcus pneumoniae: two allelic variants of the peptide pheromone." J Bacteriol **178**(20): 6087-90.

Prudhomme, M., Attaiech, L., Sanchez, G., Martin, B. and Claverys, J. P. (2006). "Antibiotic stress induces genetic transformability in the human pathogen Streptococcus pneumoniae." Science **313**(5783): 89-92.

Quentin, Y., Fichant, G. and Denizot, F. (1999). "Inventory, assembly and analysis of Bacillus subtilis ABC transport systems." J Mol Biol **287**(3): 467-84.

Raymond, J., Le Thomas, I., Moulin, F., Commeau, A., Gendrel, D. and Berche, P. (2000). "Sequential colonization by Streptococcus pneumoniae of healthy children living in an orphanage." J Infect Dis **181**(6): 1983-8.

Reichmann, P., Konig, A., Linares, J., Alcaide, F., Tenover, F. C., McDougal, L., Swidsinski, S. and Hakenbeck, R. (1997). "A global gene pool for high-level cephalosporin resistance in commensal Streptococcus species and Streptococcus pneumoniae." J Infect Dis **176**(4): 1001-12.

Ren, B., McCrory, M. A., Pass, C., Bullard, D. C., Ballantyne, C. M., Xu, Y., Briles, D. E. and Szalai, A. J. (2004a). "The virulence function of Streptococcus pneumoniae surface protein A involves inhibition of complement activation and impairment of complement receptor-mediated protection." J Immunol **173**(12): 7506-12.

Ren, B., Szalai, A. J., Hollingshead, S. K. and Briles, D. E. (2004b). "Effects of PspA and antibodies to PspA on activation and deposition of complement on the pneumococcal surface." Infect Immun **72**(1): 114-22.

Rigden, D. J. and Jedrzejas, M. J. (2003). "Genome-based identification of a carbohydrate binding module in Streptococcus pneumoniae hyaluronate lyase." Proteins **52**(2): 203-11.

Ring, A., Weiser, J. N. and Tuomanen, E. I. (1998). "Pneumococcal trafficking across the blood-brain barrier. Molecular analysis of a novel bidirectional pathway." J Clin Invest **102**(2): 347-60.

Robinson, W. G. and Coon, M. J. (1957). "The purification and properties of beta-hydroxyisobutyric dehydrogenase." J Biol Chem **225**(1): 511-21.

Roche, P. W., Krause, V., Cook, H., Barralet, J., Coleman, D., Sweeny, A., Fielding, J., Giele, C., Gilmour, R., Holland, R., Kampen, R., Brown, M., Gilbert, L., Hogg, G. and Murphy, D. (2008). "Invasive pneumococcal disease in Australia, 2006." Commun Dis Intell **32**(1): 18-30.

Rodgers, G. L., Arguedas, A., Cohen, R. and Dagan, R. (2009). "Global serotype distribution among Streptococcus pneumoniae isolates causing otitis media in children: potential implications for pneumococcal conjugate vaccines." Vaccine **27**(29): 3802-10.

Romero, P., Garcia, E. and Mitchell, T. J. (2009a). "Development of a prophage typing system and analysis of prophage carriage in Streptococcus pneumoniae." Appl Environ Microbiol **75**(6): 1642-9.

Romero, P., Croucher, N. J., Hiller, N. L., Hu, F. Z., Ehrlich, G. D., Bentley, S. D., Garcia, E. and Mitchell, T. J. (2009b). "Comparative genomic analysis of ten Streptococcus pneumoniae temperate bacteriophages." J Bacteriol **191**(15): 4854-62.

Romero-Steiner, S., Pilishvili, T., Sampson, J. S., Johnson, S. E., Stinson, A., Carlone, G. M. and Ades, E. W. (2003). "Inhibition of pneumococcal adherence to human nasopharyngeal epithelial cells by anti-PsaA antibodies." Clin Diagn Lab Immunol **10**(2): 246-51.

Romero-Steiner, S., Caba, J., Rajam, G., Langley, T., Floyd, A., Johnson, S. E., Sampson, J. S., Carlone, G. M. and Ades, E. (2006). "Adherence of recombinant pneumococcal surface adhesin A (rPsaA)-coated particles to human nasopharyngeal epithelial cells for the evaluation of anti-PsaA functional antibodies." Vaccine **24**(16): 3224-31.

Rosenow, C., Ryan, P., Weiser, J. N., Johnson, S., Fontan, P., Ortqvist, A. and Masure, H. R. (1997). "Contribution of novel choline-binding proteins to adherence, colonization and immunogenicity of Streptococcus pneumoniae." Mol Microbiol **25**(5): 819-29.

Rothstein, J., Heazlewood, R. and Fraser, M. (2007). "Health of Aboriginal and Torres Strait Islander children in remote Far North Queensland: findings of the Paediatric Outreach Service." Med J Aust **186**(10): 519-21.

Sakane, F., Imai, S., Kai, M., Yasuda, S. and Kanoh, H. (2007). "Diacylglycerol kinases: why so many of them?" Biochim Biophys Acta **1771**(7): 793-806.

Sa-Leao, R., Nunes, S., Brito-Avo, A., Frazao, N., Simoes, A. S., Crisostomo, M. I., Paulo, A. C., Saldanha, J., Santos-Sanches, I. and de Lencastre, H. (2009). "Changes in pneumococcal serotypes and antibiotypes carried by vaccinated and unvaccinated day-care centre attendees in Portugal, a country with widespread use of the seven-valent pneumococcal conjugate vaccine." Clin Microbiol Infect **15**(11): 1002-7.

Saluja, S. K. and Weiser, J. N. (1995). "The genetic basis of colony opacity in Streptococcus pneumoniae: evidence for the effect of box elements on the frequency of phenotypic variation." Mol Microbiol **16**(2): 215-27.

Sandgren, A., Sjostrom, K., Olsson-Liljequist, B., Christensson, B., Samuelsson, A., Kronvall, G. and Henriques Normark, B. (2004). "Effect of clonal and serotype-specific properties on the invasive capacity of Streptococcus pneumoniae." J Infect Dis **189**(5): 785-96.

Sandgren, A., Albiger, B., Orihuela, C. J., Tuomanen, E., Normark, S. and Henriques-Normark, B. (2005). "Virulence in mice of pneumococcal clonal types with known invasive disease potential in humans." J Infect Dis **192**(5): 791-800.

Schmeck, B., Gross, R., N'Guessan, P. D., Hocke, A. C., Hammerschmidt, S., Mitchell, T. J., Rosseau, S., Suttorp, N. and Hippenstiel, S. (2004). "Streptococcus pneumoniae-induced caspase 6-dependent apoptosis in lung epithelium." Infect Immun **72**(9): 4940-7.

Schmitz, S., Hoffmann, A., Szekat, C., Rudd, B. and Bierbaum, G. (2006). "The lantibiotic mersacidin is an autoinducing peptide." Appl Environ Microbiol **72**(11): 7270-7.

Severi, E., Muller, A., Potts, J. R., Leech, A., Williamson, D., Wilson, K. S. and Thomas, G. H. (2008). "Sialic acid mutarotation is catalyzed by the Escherichia coli beta-propeller protein YjhT." J Biol Chem **283**(8): 4841-9.

Shakhnovich, E. A., King, S. J. and Weiser, J. N. (2002). "Neuraminidase expressed by Streptococcus pneumoniae desialylates the lipopolysaccharide of Neisseria meningitidis and Haemophilus influenzae: a paradigm for interbacterial competition among pathogens of the human respiratory tract." Infect Immun **70**(12): 7161-4.

Shaper, M., Hollingshead, S. K., Benjamin, W. H., Jr. and Briles, D. E. (2004). "PspA protects Streptococcus pneumoniae from killing by apolactoferrin, and antibody to PspA enhances killing of pneumococci by apolactoferrin [corrected]." Infect Immun **72**(9): 5031-40.

Shelburne, S. A., 3rd, Keith, D. B., Davenport, M. T., Beres, S. B., Carroll, R. K. and Musser, J. M. (2009). "Contribution of AmyA, an extracellular alpha-glucan degrading enzyme, to group A streptococcal host-pathogen interaction." Mol Microbiol **74**(1): 159-74.

Shiio, I. and Sugimoto, S. (1976). "Altered prephenate dehydratase in phenylalanine-excreting mutants of Brevibacterium flavum." J Biochem **79**(1): 173-83.

Shivshankar, P., Sanchez, C., Rose, L. F. and Orihuela, C. J. (2009). "The Streptococcus pneumoniae adhesin PsrP binds to Keratin 10 on lung cells." Mol Microbiol **73**(4): 663-79.

Shoemaker, N. B., Smith, M. D. and Guild, W. R. (1979). "Organization and transfer of heterologous chloramphenicol and tetracycline resistance genes in pneumococcus." J Bacteriol **139**(2): 432-41.

Siber, G. R., Klugman, K. P. and Makela, P. H. (2008). Pneumococcal Vaccines: The Impact of Conjugate Vaccine. Washington, DC, ASM press.

Silva, N. A., McCluskey, J., Jefferies, J. M., Hinds, J., Smith, A., Clarke, S. C., Mitchell, T. J. and Paterson, G. K. (2006). "Genomic diversity between strains of the same serotype and multilocus sequence type among pneumococcal clinical isolates." Infect Immun **74**(6): 3513-8.

Simell, B., Korkeila, M., Pursiainen, H., Kilpi, T. M. and Kayhty, H. (2001). "Pneumococcal carriage and otitis media induce salivary antibodies to pneumococcal surface adhesin a, pneumolysin, and pneumococcal surface protein a in children." J Infect Dis **183**(6): 887-96.

Sjostrom, K., Spindler, C., Ortqvist, A., Kalin, M., Sandgren, A., Kuhlmann-Berenzon, S. and Henriques-Normark, B. (2006). "Clonal and capsular types decide whether pneumococci will act as a primary or opportunistic pathogen." Clin Infect Dis **42**(4): 451-9.

Sjostrom, K., Blomberg, C., Fernebro, J., Dagerhamn, J., Morfeldt, E., Barocchi, M. A., Browall, S., Moschioni, M., Andersson, M., Henriques, F., Albiger, B., Rappuoli, R., Normark, S. and Henriques-Normark, B. (2007). "Clonal success of piliated penicillin nonsusceptible pneumococci." Proc Natl Acad Sci U S A **104**(31): 12907-12.

Smith-Vaughan, H., Marsh, R., Mackenzie, G., Fisher, J., Morris, P. S., Hare, K., McCallum, G., Binks, M., Murphy, D., Lum, G., Cook, H., Krause, V., Jacups, S. and Leach, A. J. (2009). "Age-specific cluster of cases of serotype 1 Streptococcus pneumoniae carriage in remote indigenous communities in Australia." Clin Vaccine Immunol **16**(2): 218-21.

Srivastava, A., Henneke, P., Visintin, A., Morse, S. C., Martin, V., Watkins, C., Paton, J. C., Wessels, M. R., Golenbock, D. T. and Malley, R. (2005). "The apoptotic response to pneumolysin is Toll-like receptor 4 dependent and protects against pneumococcal disease." Infect Immun **73**(10): 6479-87.

Steenhoff, A. P., Shah, S. S., Ratner, A. J., Patil, S. M. and McGowan, K. L. (2006). "Emergence of vaccine-related pneumococcal serotypes as a cause of bacteremia." Clin Infect Dis **42**(7): 907-14.

Steinfort, C., Wilson, R., Mitchell, T., Feldman, C., Rutman, A., Todd, H., Sykes, D., Walker, J., Saunders, K. and Andrew, P. W. (1989). "Effect of Streptococcus pneumoniae on human respiratory epithelium in vitro." Infect Immun **57**(7): 2006-13.

Stroeher, U. H., Kidd, S. P., Stafford, S. L., Jennings, M. P., Paton, J. C. and McEwan, A. G. (2007). "A pneumococcal MerR-like regulator and S-nitrosoglutathione reductase are required for systemic virulence." J Infect Dis **196**(12): 1820-6.

Sung, C. K. and Morrison, D. A. (2005). "Two distinct functions of ComW in stabilization and activation of the alternative sigma factor ComX in Streptococcus pneumoniae." J Bacteriol **187**(9): 3052-61.

Swiatlo, E., Champlin, F. R., Holman, S. C., Wilson, W. W. and Watt, J. M. (2002). "Contribution of choline-binding proteins to cell surface properties of Streptococcus pneumoniae." Infect Immun **70**(1): 412-5.

Syrjanen, R. K., Kilpi, T. M., Kaijalainen, T. H., Herva, E. E. and Takala, A. K. (2001). "Nasopharyngeal carriage of Streptococcus pneumoniae in Finnish children younger than 2 years old." J Infect Dis **184**(4): 451-9.

Szekeres, S., Dauti, M., Wilde, C., Mazel, D. and Rowe-Magnus, D. A. (2007). "Chromosomal toxin-antitoxin loci can diminish large-scale genome reductions in the absence of selection." Mol Microbiol **63**(6): 1588-605.

Takata, H., Kuriki, T., Okada, S., Takesada, Y., Iizuka, M., Minamiura, N. and Imanaka, T. (1992). "Action of neopullulanase. Neopullulanase catalyzes both hydrolysis and transglycosylation at alpha-(1----4)- and alpha-(1----6)-glucosidic linkages." J Biol Chem **267**(26): 18447-52.

Tettelin, H., Nelson, K. E., Paulsen, I. T., Eisen, J. A., Read, T. D., Peterson, S., Heidelberg, J., DeBoy, R. T., Haft, D. H., Dodson, R. J., Durkin, A. S., Gwinn, M., Kolonay, J. F., Nelson, W. C., Peterson, J. D., Umayam, L. A., White, O., Salzberg, S. L., Lewis, M. R., Radune, D., Holtzapple, E., Khouri, H., Wolf, A. M., Utterback, T. R., Hansen, C. L., McDonald, L. A., Feldblyum, T. V., Angiuoli, S., Dickinson, T., Hickey, E. K., Holt, I. E., Loftus, B. J., Yang, F., Smith, H. O., Venter, J. C., Dougherty, B. A., Morrison, D. A., Hollingshead, S. K. and Fraser, C. M. (2001). "Complete genome sequence of a virulent isolate of Streptococcus pneumoniae." Science **293**(5529): 498-506.

Tomasz, A. and Hotchkiss, R. D. (1964). "Regulation of the Transformability of Pheumococcal Cultures by Macromolecular Cell Products." Proc Natl Acad Sci U S A **51**: 480-7.

Tong, H. H., McIver, M. A., Fisher, L. M. and DeMaria, T. F. (1999). "Effect of lacto-N-neotetraose, asialoganglioside-GM1 and neuraminidase on adherence of otitis media-associated serotypes of Streptococcus pneumoniae to chinchilla tracheal epithelium." Microb Pathog **26**(2): 111-9.

Trombe, M. C., Clave, C. and Manias, J. M. (1992). "Calcium regulation of growth and differentiation in Streptococcus pneumoniae." J Gen Microbiol **138**(1): 77-84.

Tseng, H. J., McEwan, A. G., Paton, J. C. and Jennings, M. P. (2002). "Virulence of Streptococcus pneumoniae: PsaA mutants are hypersensitive to oxidative stress." Infect Immun **70**(3): 1635-9.

Tu, A. H., Fulgham, R. L., McCrory, M. A., Briles, D. E. and Szalai, A. J. (1999). "Pneumococcal surface protein A inhibits complement activation by Streptococcus pneumoniae." Infect Immun **67**(9): 4720-4.

Tuomanen, E., Tomasz, A., Hengstler, B. and Zak, O. (1985). "The relative role of bacterial cell wall and capsule in the induction of inflammation in pneumococcal meningitis." J Infect Dis **151**(3): 535-40.

Uchiyama, S., Carlin, A. F., Khosravi, A., Weiman, S., Banerjee, A., Quach, D., Hightower, G., Mitchell, T. J., Doran, K. S. and Nizet, V. (2009). "The surface-anchored NanA protein promotes pneumococcal brain endothelial cell invasion." J Exp Med **206**(9): 1845-52.

Umemoto, J., Bhavanandan, V. P. and Davidson, E. A. (1977). "Purification and properties of an endo-alpha-N-acetyl-D-galactosaminidase from Diplococcus pneumoniae." J Biol Chem **252**(23): 8609-14.

Urban, C. F., Lourido, S. and Zychlinsky, A. (2006). "How do microbes evade neutrophil killing?" Cell Microbiol **8**(11): 1687-96.

Van Melderen, L. and Saavedra De Bast, M. (2009). "Bacterial toxin-antitoxin systems: more than selfish entities?" PLoS Genet **5**(3): e1000437.

Van der Poll, T. and Opal, S. M. (2009). "Pathogenesis, treatment, and prevention of pneumococcal pneumonia." Lancet **374**(9700): 1543-56.

Ventura, C. L., Cartee, R. T., Forsee, W. T. and Yother, J. (2006). "Control of capsular polysaccharide chain length by UDP-sugar substrate concentrations in Streptococcus pneumoniae." Mol Microbiol **61**(3): 723-33.

Waite, R. D., Struthers, J. K. and Dowson, C. G. (2001). "Spontaneous sequence duplication within an open reading frame of the pneumococcal type 3 capsule locus causes high-frequency phase variation." Mol Microbiol **42**(5): 1223-32.

Waite, R. D., Penfold, D. W., Struthers, J. K. and Dowson, C. G. (2003). "Spontaneous sequence duplications within capsule genes cap8E and tts control phase variation in Streptococcus pneumoniae serotypes 8 and 37." Microbiology **149**(Pt 2): 497-504.

Wani, J. H., Gilbert, J. V., Plaut, A. G. and Weiser, J. N. (1996). "Identification, cloning, and sequencing of the immunoglobulin A1 protease gene of Streptococcus pneumoniae." Infect Immun **64**(10): 3967-74.

Wartha, F., Beiter, K., Normark, S. and Henriques-Normark, B. (2007). "Neutrophil extracellular traps: casting the NET over pathogenesis." Curr Opin Microbiol **10**(1): 52-6.

Watson, D. A. and Musher, D. M. (1990). "Interruption of capsule production in Streptococcus pneumonia serotype 3 by insertion of transposon Tn916." Infect Immun **58**(9): 3135-8.

Watson, D. A., Musher, D. M., Jacobson, J. W. and Verhoef, J. (1993). "A brief history of the pneumococcus in biomedical research: a panoply of scientific discovery." Clin Infect Dis **17**(5): 913-24.

Weinberger, D. M., Trzcinski, K., Lu, Y. J., Bogaert, D., Brandes, A., Galagan, J., Anderson, P. W., Malley, R. and Lipsitch, M. (2009). "Pneumococcal capsular polysaccharide structure predicts serotype prevalence." PLoS Pathog **5**(6): e1000476.

Weiser, J. N., Austrian, R., Sreenivasan, P. K. and Masure, H. R. (1994). "Phase variation in pneumococcal opacity: relationship between colonial morphology and nasopharyngeal colonization." Infect Immun **62**(6): 2582-9.

Weiser, J. N., Markiewicz, Z., Tuomanen, E. I. and Wani, J. H. (1996). "Relationship between phase variation in colony morphology, intrastrain variation in cell wall physiology, and nasopharyngeal colonization by Streptococcus pneumoniae." Infect Immun **64**(6): 2240-5.

Whatmore, A. M., Barcus, V. A. and Dowson, C. G. (1999). "Genetic diversity of the streptococcal competence (com) gene locus." J Bacteriol **181**(10): 3144-54.

Willey, J. M. and van der Donk, W. A. (2007). "Lantibiotics: peptides of diverse structure and function." Annu Rev Microbiol **61**: 477-501.

Wilson, D. B. and Hogness, D. S. (1969). "The enzymes of the galactose operon in Escherichia coli. II. The subunits of uridine diphosphogalactose 4-epimerase." J Biol Chem **244**(8): 2132-6.

Winkelstein, J. A. (1984). "Complement and the host's defense against the pneumococcus." Crit Rev Microbiol **11**(3): 187-208.

Woehle, D. L., Lueddecke, B. A. and Ludden, P. W. (1990). "ATP-dependent and NAD-dependent modification of glutamine synthetase from Rhodospirillum rubrum in vitro." J Biol Chem **265**(23): 13741-9.

Yaro, S., Lourd, M., Traore, Y., Njanpop-Lafourcade, B. M., Sawadogo, A., Sangare, L., Hien, A., Ouedraogo, M. S., Sanou, O., Parent du Chatelet, I., Koeck, J. L. and Gessner, B. D. (2006). "Epidemiological and molecular characteristics of a highly lethal pneumococcal meningitis epidemic in Burkina Faso." Clin Infect Dis **43**(6): 693-700.

Yesilkaya, H., Manco, S., Kadioglu, A., Terra, V. S. and Andrew, P. W. (2008). "The ability to utilize mucin affects the regulation of virulence gene expression in Streptococcus pneumoniae." FEMS Microbiol Lett **278**(2): 231-5.

Young, I., Wang, I. and Roof, W. D. (2000). "Phages will out: strategies of host cell lysis." Trends Microbiol **8**(3): 120-8.

Zahner, D. and Hakenbeck, R. (2000). "The Streptococcus pneumoniae beta-galactosidase is a surface protein." J Bacteriol **182**(20): 5919-21.

Zapun, A., Contreras-Martel, C. and Vernet, T. (2008). "Penicillin-binding proteins and beta-lactam resistance." FEMS Microbiol Rev **32**(2): 361-85.

Zemlickova, H., Melter, O. and Urbaskova, P. (2006). "Epidemiological relationships among penicillin non-susceptible Streptococcus pneumoniae strains recovered in the Czech Republic." J Med Microbiol **55**(Pt 4): 437-42.

Zhang, J. R., Mostov, K. E., Lamm, M. E., Nanno, M., Shimida, S., Ohwaki, M. and Tuomanen, E. (2000). "The polymeric immunoglobulin receptor translocates pneumococci across human nasopharyngeal epithelial cells." Cell **102**(6): 827-37.

Appendix

A.1 Chapter 5 supplementary material

A.1.1 Explanatory notes

Additional data from the genomic comparisons of Section 5.4 is contained within this appendix. Tables A.1 and A.2 include the BLASTx and BLASTn search results of the contigs greater than 300-bp in size that were assembled from ‘boneyard’ sequences from strain 1 (Table A.1) and strain 1861 (Table A.2). Genes homologous to those returned from the BLAST results are also included in Tables A.1 and A.2 and were identified using the KEGG database (Section 2.6.2). Where a high-scoring hit was not returned from the KEGG database, the strain or species with the highest-scoring hit from the NCBI database is included in the ORF description in brackets. In addition, ORF descriptions are colour coded according to the criteria of Table A.i.

Table A.i Colour coding for ORF descriptions of BLAST results

Colour key for gene description	
Virulence factors	🟧
Transporters	🟨
Metabolism-associated	🟩
Regulation of transcription	🟨
Phage-associated	🟥
Transposase/transposon-associated	🟦
Hypothetical proteins	⬜
Other types of genes	🟪

In addition, the BLASTx and BLASTn search results of the complete list of discrepancies between the consensus sequences of the sequenced strains and the reference genome (P1031) are included for strain 1 (Table A.3) and strain 1861 (Table A.4). For discrepancies less than 150-bp in size, 150-bp of P1031 sequence including and flanking the discrepancy was used for the BLAST searches. Therefore, the data of Tables A.3 and A.4 for discrepancies less than 150-bp in size would require verification in order to confirm the location of the discrepancy relative to the genes returned from the BLAST searches. In addition, *S. pneumoniae* ORFs of the KEGG database (Section 2.6.2) homologous to those identified by the BLAST searches are also included in Tables A.3 and A.4. The ORF descriptions are colour coded according to that of Table A.i.

Table A.1 BLAST results of strain 1 boneyard

Contig	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Contig length
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
1318	Hypothetical	-	-	1353	-	-	1377	-	0827	-	-	3139bp
	ABC transporter, permease, LplC	-	-	-	-	-	-	-	0826	-	-	
	ABC transporter, substrate binding protein, LplB	-	-	-	-	-	-	-	0825	-	-	
	N-acetylmannosamine-6-phosphate epimerase	1685	1166	1702	12220	1579	1724	1623	0824	1593	1658	
1280	Hypothetical	-	-	-	-	-	-	-	0833	-	-	2789bp
	Cytidine deaminase	-	-	-	-	-	-	-	-	-	-	
	Diadenosine tetraphosphate	-	-	-	-	-	-	-	0830	-	-	
	Dihydrolipoamide dehydrogenase	-	-	-	-	-	-	-	0829	-	-	
1440	Collagen adhesion protein	-	-	-	-	-	-	1056	-	-	-	2540bp
1531	Sugar ABC transporter, substrate-binding protein	-	-	-	-	-	-	-	0828	-	-	2515bp
1714	Phosphonate monoester hydrolase, putative choline sulfatase	-	-	-	18200	1708	1861	-	1924	-	1785	2656bp
	PTS, Cellobiose-specific IIc component	-	-	1807	18210	1709	1862	-	-	-	1786	
1712	Hypothetical	-	-	-	-	-	-	1059	-	-	-	1999bp
	SrtG1, membrane cysteine transpeptidases (TCH8431/19A)	-	-	-	-	-	-	-	-	-	-	
2028	ZmpB	0664	0577	0684	10590	1074	0723	0688	0759	0605	0620	1928bp
1759	Putative bacteriocin, Lactococcin group	-	0595	0704	20090	-	0742	0707	0778	0625	1952	1883bp
	M50 family peptidase	0694	0603	-	-	-	-	-	-	-	0648	
1481	Drug efflux ABC transporter, ATP-binding/permease	1342	1177	-	12310	-	1382	-	1479	-	-	1855bp
	ABC transporter, ATP-binding	1341	1176	-	12300	-	1381	-	1478	-	-	
1416	PlcR transcriptional regulator	+	-	-	-	-	-	-	-	-	-	1798bp
1215	ZmpB	0664	0577	0684	10590	1074	0723	0688	0759	0605	0620	1684bp
2026	PTS cellobiose specific, IIc component	0310	0283	-	-	-	0373	0360	-	-	0318	1562bp
1247	Sialidase nanA-like (*Erysipelothrix rhusiopathiae*)	-	-	-	-	-	-	-	-	-	-	1558bp
1379	Hypothetical, Putative DNA replication protein	1136	-	-	-	-	-	-	-	-	-	1541bp
	Hypothetical	1135	-	-	-	-	-	0227	0294	-	-	
	Hypothetical	1133	-	-	-	-	-	-	-	-	-	
	Hypothetical, phage-related	1132	-	-	-	-	-	-	0298	-	-	
1294	Hypothetical	1140	-	-	-	-	-	-	-	-	-	1488bp
	Hypothetical	1139	-	-	-	-	-	-	-	-	-	

Contig	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Contig length
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
1170	IgA1 protease	1154	1018	1140	10580	1073	1207	1181	1229	1053	1143	1481bp
1369	Substilin-like serine protease	-	-	-	17970	1683	-	-	-	-	1762	1471bp
1276	PTS cellobiose specific, IIa component	0309	0282	-	-	-	0372	0359	-	-	0317	1436bp
	PTS cellobiose specific, IIa component	0308	0281	-	-	-	0371	0358	-	-	0316	
	PTS cellobiose specific, IIb component	0307	0280	-	-	-	0370	0357	-	-	-	
1351	Protease II (oligopeptidase)	1343	1178	-	12320	-	1383	-	-	-	-	1434bp
	Hypothetical (protein kinase domain)	1344	1179	-	12330	-	1384	-	-	-	-	
1459	Tryptophan synthase β	-	-	-	17960	1682	-	-	-	-	1761	1406bp
	Substilin-like serine protease	-	-	-	17970	1683	-	1708	-	-	1762	
1408	Lantibiotic mersacidin transport system (CDC3059-06)	-	-	-	-	-	-	-	-	-	-	1432bp
1771	Phosphosugar-binding transcriptional regulator, RipR	-	-	-	12240	-	-	-	-	-	-	1650bp
	Hypothetical	-	-	-	12250	-	-	-	0822	-	-	
1784	Cell wall surface anchor protein, SrtG2	-	-	-	-	-	-	1061	-	-	-	1362bp
1386	Hypothetical	0304	0278	-	-	-	0368	0355	-	-	0313	
	PTS cellobiose specific, IIb component	0305	0279	-	-	-	0369	0356	-	-	0314	
	PTS cellobiose specific, IIb component	0306	0280	-	-	-	0370	0357	-	-	0315	
1653	Hypothetical	1344	1179	-	12330	-	1384	-	-	-	-	1283bp
1179	Hypothetical (*S. thermophilus*)	-	-	-	-	-	-	-	-	-	-	1260bp
	UBA/THIF-type NAD/FAD binding protein	-	-	-	-	-	-	-	-	-	-	
1391	IS1381, transposase OrfA	-	+	+	-	+	+	+	-	+	+	1252bp
	PcpA, cell surface choline binding protein	2136	-	-	21690	2161	2262	2148	2328	2074	2105	
1273	ABC-2 type transport system permease	0698	0606	-	-	-	-	-	-	-	0651	1211bp
	ABC transporter, ATP-binding, unknown substrate	-	-	-	-	-	-	-	-	-	0650	
1283	Phage integrase family	1129	-	-	-	-	-	0222	0289	-	-	1205bp
1495	Probable ABC transporter, permease	-	-		09840	1001			-	-	-	1182bp
	Bacitracin transport, ATP-binding protein, BcrA				09830	1000						
1303	ABC-type multidrug transport system, ATPase and permease (*S. thermophilus*)	-	-	-	-	-	-	-	-	-	-	1175bp
1159	6-phospho-β-glucosidase, bglA	0303	0277	0594	05220	0275	0367	0354	0677	0526	0312	1088bp
1404	Prolyl oligopeptidase family	1333	1181	1148	13120	-	-	-	1475	1288	1337	974bp
2064	HesA/MoeB/ThiF family	0696	0604	-	-	-	-	-	-	-	-	949bp
	Bacitracin transport, ATP-binding protein, BcrA	-	-	-	-	-	-	-	-	-	0650	
1716	ABC-type phosphate transport system, ATPase component	0826	-	-	07470	0764	-	-	-	-	0767	947bp
1763	Hypothetical, possibly permease	-	-	-	12290	-	1380	-	-	-	-	938bp

Contig	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Contig length
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
1105	Hypothetical	-	-	-	-	-	-	-	0219	0112	-	889bp
1523	ATPase of AAA+ class	-	-	-	17980	1684	-	1709	-	-	1763	878bp
1292	Argininosuccinate lyase	-	0111	-	-	0138	0178	-	0216	-	-	859bp
1317	Acetyltransferase, GNAT family	-	0697	-	07220	0739	0839	-	-	-	-	859bp
1259	Peptide-2 ABC transporter, ATP-binding/membrane spanning (*S. thermophilus*)	-	-	-	-	-	-	-	-	-	-	847bp
1287	Transporter, major facilitator family, LabT	0145	0148	0215	01540	0174	0218	0191	0256	0147	1596	844bp
1304	Putative membrane protein (23-BS72)	-	-	-	-	-	-	-	-	-	-	831bp
1169	Hypothetical	0531	-	-	-	-	0589	-	0637	-	-	827bp
1398	lanthione biosynthesis protein (*Clostridium cellulovorans*)	-	-	-	-	-	-	-	-	-	-	797bp
1262	Hypothetical	1142	-	-	15540	-	-	-	0094	-	-	796bp
1264	IgA1 protease	1154	1018	1140	10580	1073	1207	1181	1229	1053	1143	792bp
2106	PTS Iibc, mannitol-specific	0333	0303	0367	03050	0326	0396	0383	0443	0302	0336	775bp
2702	Acetyltransferase, GNAT family	-	-	-	-	-	2134	-	-	-	-	772bp
1409	Lantibiotic mersacidin modifying enzyme	-	-	-	-	0997	-	-	-	-	-	761bp
776	Mannitol-1-phosphate 5-dehydrogenase	0397	0363	-	03730	00386	0470	-	0507	0365	0397	736bp
1269	Transcriptional regulator mtlR	0395	0361	-	03730	0384	0468	-	0505	0363	0395	712bp
3007	Signal peptidase I	-	-	-	-	-	-	1058	-	-	-	694bp
1609	GTP-binding protein - DNA replication protein DnaC	1137	-	-	-	-	-	-	-	-	-	691bp
1574	PcpA, cell surface choline binding protein	2136	1965	-	21690	2161	2262	2148	2328	2074	2105	667bp
1687	ZmpB	0664	0577	0684	05990	1074	0723	0688	0759	0605	0620	656bp
1959	Hypothetical	-	-	-	-	-	-	0229	0296	-	-	647bp
	Replication protein (*CDC1873-00*)	-	-	-	-	-	-	-	-	-	-	
1341	PTS mannitol specific Iibc, mltA	0394	0360	-	03700	0383	0467	-	0504	0362	0394	637bp
1090	Argininosuccinate synthase, argG	-	0110	-	-	-	-	-	0215	-	-	623bp
1506	Drug efflux ABC transporter ATP-binding/permease	-	1177	-	-	-	-	-	-	-	-	622bp
	Prolyl oligopeptidase family protein, ptrB	1343	1178	-	12320	-	1383	-	-	-	-	
1551	HesA/MoeB/ThiF family	0694	0603	-	-	-	-	-	-	-	0648	613bp
1301	PTS IIb component	0306	0280	-	-	-	0370	-	-	-	0315	590bp
1724	Transporter, truncation	-	-	-	-	0140	0180	0151	-	-	0111	587bp
1414	Phosphosugar-binding transcriptional regulator, RpiR	-	-	-	12240	-	-	-	-	-	-	587bp
1693	Putative bacteriocin, Lactococcin domain	-	-	-	20090	1981	2060	-	2096	1856	1952	585bp
1853	IgA1 protease	1154	1018	1140	10580	1073	1207	1181	1229	1053	1143	583bp

Contig	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Contig length
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
1851	IgA1 protease	1154	1018	1140	10580	1073	1207	1181	1229	1053	1143	580bp
1940	PcpA, cell surface choline binding protein	2136	1965	-	21690	2161	2262	2148	2328	2074	2105	564bp
1308	ATPase of AAA+ class	-	-	-	17980	1684	-	1709	-	-	1763	543bp
1755	Bacterial extracellular solute-binding Dihydrolipoamide dehydrogenase	- -	- -	- -	- -	- -	- -	- -	0828 0829	- -	- -	537bp
1594	Transcriptional regulator mtlR	0395	0361	-	03710	0384	0468	-	0505	0363	0395	536bp
1450	Transposase	-	-	-	+	-	-	-	+	-	+	532bp
1079	PTS, lactose specific IIBC components	1185	1047	1228	10860	1103	1237	1039	1305	1082	1111	529bp
1453	Transcriptional regulator	0395	0361	-	03710	0384	0468	-	0505	0363	0395	513bp
1652	Hypothetical	1340	1175	-	12290	-	1380	-	1477	-	1454	303bp
1428	Type I restriction modification, S subunit	-	-	-	-	0828	0927	1309	-	-	-	462bp
1330	Mannitol-1-phosphate 5-dehydrogenase, mltD	0397	0363	-	03730	0386	0470	-	0507	0365	0397	461bp
1752	Putative lantibiotic synthetase	1344	1179	-	12330	-	1384	-	-	-	-	458bp
1719	Type I restriction modification, S subunit	0508	0453	0530	04590	0474	-	0542	0616	0460	0483	457bp
1691	IgA1 protease	1154	1018	1140	10580	1073	1207	1181	1229	1053	1143	454bp
1785	Drug efflux ABC transporter, ATP-binding/permease protein	1342	1177	-	12310	-	1382	-	1479	-	-	433bp
1338	Valyl-tRNA Synthetase	0568	0494	0585	05140	0529	0633	0599	0666	0516	0531	433bp
1750	Hypothetical Hypothetical	- 0398	0364 -	- -	03750 -	0387 -	0471 -	- -	0508 -	0366 -	- 0398	425bp
1380	Hypothetical	0108	0112	-	01707	0139	0179	0150	0208	0109	0110	418bp
1266	PTS, lactose specific IIBC components	1185	1047	1228	10860	1103	1237	1039	1305	1082	1111	414bp
1310	PTS, lactose specific IIBC components	1185	1047	1228	10860	1103	1237	1039	1305	1082	1111	413bp
1545	NanA	-	1504	1709	16920	1586	1731	1630	1797	1600	1665	401bp
1628	Conserved domain protein (CDC1087-00)	-	-	-	-	-	-	-	-	-	-	397bp
1306	Lantibiotic biosynthesis protein (*B. cereus*)	-	-	-	-	-	-	-	-	-	-	395bp
	DUF181 protein of unknown function (*Geobacillus thermoglucosidasius*)	-	-	-	-	-	-	-	-	-	-	
1458	Mannitol-specific enzyme IIA, mtlF Mannitol-1-phosphate, 5-dehydrogenase	0396 0397	0362 0363	- -	03720 03730	0385 0386	0469 0470	- -	0506 0507	0364 0365	0396 0397	395bp
1309	Putative transcriptional regulator	1131	-	-	-	-	-	-	-	-	-	394bp
1498	IS1239-truncation degenerate transposase	-	-	-	-	-	-	-	-	-	-	393bp
1577	Substilisin-like serine protease	-	-	-	17970	1683	-	1708	-	-	1762	393bp
2178	IS1381, transposase OrfA	-	-	-	-	-	-	-	-	-	-	391bp

Contig	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Contig length
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
1304	PTS, cellobiose IIC component, celD	0310	0283	-	-	-	0373	0360	-	-	0318	388bp
	PTS, IIA component	0309	0282	-	-	-	0372	0359	-	-	0317	
1181	Choline binding protein F	0391	0357	0415	03650	0379	0464	0421	0501	0343	0373	383bp
1789	Choline binding protein G	0390	-	0430	03640	0378	-	0436	-	0356	0387	376bp
1662	Hypothetical (*S. thermophilus*)	-	-	-	-	-	-	-	-	-	-	374bp
1352	Ribose 5-phosphate isomerase A	0828	0723	0834	07490	0766	0864	1374	0926	0749	0770	372bp
1825	Conserved hypothetical protein	-	-	-	-	-	-	-	-	-	-	368bp
1984	Predicted Cro/CI family transcriptional regulator	-	-		-	-	-	-	-	-	-	368bp
2125	IgA1 protease	1154	1018	1140	10580	1073	1207	1181	1229	1053	1143	364bp
1323	6-phospho-β-glucosidase, bglA	0303	0277	-	-	-	0367	0354	-	-	0312	363bp
1426	Lantibiotic biosynthesis protein (*S. thermophilus*)	-	-	-	-	-	-	-	-	-	-	358bp
2369	Hypothetical	1140	-	-	-	-	-	-	-	-	-	357bp
	Hypothetical	1141	-	-	-	-	-	-	-	-	-	
2073	IgA1 protease	1154	1018	1140	10580	1073	1207	1181	1229	1053	1143	353bp
1460	PTS, mannitol-specific IIBC components	0394	0360	-	03700	0383	0467	-	0504	0362	0394	352bp
2238	Ferrochelatase, hemH	1009	0895	-	09340	0949	1048	1062	1111	0935	0985	350bp
1078	PspC	2190	2017	2242	22240	2217	2316	2208	2388	0121	2158	364bp
1819	Hypothetical	1338	-	-	-	-	-	-	-	-	-	346bp
	Hypothetical	1339	-	-	12280	-	1379	-	1476	-	-	
1469	Transcriptional regulator, mtlR	0395	0361	-	03710	0384	0468	-	0505	0363	0395	342bp
1855	GTP-binding protein	1137	-	-	-	-	-	-	-	-	-	337bp
1097	Conserved hypothetical protein	1914	1717	1943	19380	1909	1995	1864	2057	1826	1889	335bp
1242	Hypothetical	-	-	-	09850	1002	-	-	-	-	-	334bp
1451	Conserved hypothetical protein (CDC0288-04)	-	-	-	-	-	-	-	-	-	-	330bp
1436	β-galactosidase precursor	0648	0562	0665	05830	0596	0707	0670	0741	0588	0603	330bp
1639	Branched-chain amino acid permease	0146	-	0216	01550	0175	-	0193	0258	0148	0151	330bp
1884	Esterase superfamily	0882	0778	-	-	-	0920	1319	0987	0807	0831	326bp
2285	phtA precursor	1175	1038	1218	10770	1093	1227	1049	1104	1073	1122	325bp
1630	Bifunctional purine synthesis, PurH	0050	0057	0115	00670	0082	0117	0089	0157	0056	0052	322bp
1598	Conserved hypothetical protein	-	-	-	-	-	-	-	0841	-	-	315bp
1733	Transcription antiterminator, BglG family	0306	0280	-	-	-	0370	0357	-	-	0315	313bp
2232	Hypothetical	1136	1405	1598	15920	1481	1616	1515	0295	1501	1561	313bp
1710	PTS, mannitol-specific IIBC, mltA	0394	0360	-	03700	0383	0467	-	0504	0362	0394	311bp

Table A.2 BLAST results of strain 1861 boneyard

Contig	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Contig length
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
506	RecT Phage protein (CDC0288-04)	-	-	-	-	-	-	-	-	-	-	3131bp
	Phage protein	-	-	-	-	1884	-	-	-	-	-	
	Putative single strand binding protein	-	-	-	-	1881	-	-	-	-	-	
124	CI repressor phage	-	-	-	-	1901	-	0029	-	-	-	1766bp
	Hypothetical (19-BS75)	-	-	-	-	-	-	-	-	-	-	
683	Hypothetical	-	-	-	12610	-	-	-	1264	-	1294	1615bp
	Hypothetical	-	-	-	12600	-	-	-	1265	-	1293	
126	Phage portal protein (6-BS73)	-	-	-	-	-	-	-	-	-	-	1519bp
	Phage terminase, large subunit (14-BS69)	-	-	-	-	-	-	-	-	-	-	
755	ABC transporter, ATP-binding	1341	1176	-	12300	-	1381	-	1478	-	1455	1221bp
302	Hypothetical (14-BS69)	-	-	-	-	-	-	-	-	-	-	1197bp
	Phage protein (14-BS69)	-	-	-	-	-	-	-	-	-	-	
	Phage protein (14-BS69)	-	-	-	-	-	-	-	-	-	-	
308	Putative phage integrase	-	-	-	-	1903	-	-	-	-	-	1181bp
246	Phage terminase, large subunit (6-BS73)	-	-	-	-	-	-	-	-	-	-	1081bp
	Phage terminase, small subunit (6-BS73)	-	-	-	-	-	-	-	-	-	-	
478	Prophage λBa01, membrane protein	-	-	-	-	1852	-	-	-	-	-	955bp
290	Hypothetical	-	-	-	-	1853	-	-	-	-	-	931bp
	Hypothetical	-	-	-	-	1854	-	-	-	-	-	
572	Phage portal protein, Spp1 family (23-BS72)	-	-	-	-	-	-	-	-	-	-	843bp
1870	Putative replication initiation protein Rep(RC)	-	-	-	12580	-	-	-	1267	-	1291	793bp
1268	Chloramphenicol acetyltransferase	-	-	-	12950	-	-	-	1266	-	1292	748bp
264	Phage putative head morphogenesis protein, Spp1 gp7 family (23-BS72)	-	-	-	-	-	-	-	-	-	-	405bp
405	Phage scaffold protein	-	-	-	-	1863	-	-	-	-	-	697bp
	Phage protein (23-BS72)	-	-	-	-	-	-	-	-	-	-	697bp
303	Hypothetical (14-BS69)	-	-	-	-	-	-	-	-	-	-	667bp
	Phage antirepressor	-	-	-	-	-	-	-	0075	-	-	667bp
239	PblB	-	-	0074	-	1850	0074	-	0062	-	-	665bp

Contig	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Contig length
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
201	Carbohydrate-binding domain family (6-BS73)	-	-	-	-	-	-	-	-	-	-	662bp
280	Hypothetical	-	-	-	15560	-	-	-	-	-	-	660bp
	Hypothetical	-	-	-	-	-	-	-	-	-	-	660bp
1140	Hypothetical	1340	1175	-	12290	-	1380	-	1477	-	1454	650bp
451	Prophage λBa01, membrane protein	-	-	-	-	1852	-	-	-	-	-	649bp
378	Hypothetical	-	-	-	-	1858	-	-	-	-	-	606bp
	Hypothetical (9-BS68)	-	-	-	-	-	-	-	-	-	-	606bp
671	NanA	-	1504	1709	16920	1586	1731	1630	1797	1600	1665	586bp
434	Phage-related protein-like protein, TMP repeat family	-	-	-	-	1852	-	-	-	-	-	560bp
267	Hypothetical	1340	1175	-	12290	-	1380	-	1477	-	1454	515bp
607	Hypothetical	-	-	-	-	1851	-	-	-	-	-	511bp
	Hypothetical, some homology to PblB	-	-	0074	-	1850	0074	-	0062	-	-	511bp
316	Phage protein (*S. pyogenes*)	-	-	-	-	-	-	-	-	-	-	498bp
461	Hypothetical (14-BS69)	-	-	-	-	-	-	-	-	-	-	492bp
	Lj965 prophage major tail protein	-	-	-	-	1855	-	-	-	-	-	492bp
470	Phage antirepressor protein	-	-	-	-	-	-	-	0075	-	-	491bp
341	PblB	-	-	0074	-	1850	0074	-	0062	-	-	485bp
344	Hypothetical, phage related (14-BS69)	-	-	-	-	-	-	-	-	-	-	485bp
356	Hypothetical (14-BS69)	-	-	-	-	-	-	-	-	-	-	483bp
	Phage protein (14-BS69)	-	-	-	-	-	-	-	-	-	-	483bp
333	PblB	-	-	0074	-	1850	0074	-	0062	-	-	483bp
279	Metal dependent phosphydrolase HD region	-	-	-	-	1866	-	-	-	-	-	481bp
243	Phage scaffold protein	-	-	-	-	1863	-	-	-	-	-	475bp
	Possible carbohydrate-binding family v/xii (9-BS68)	-	-	-	-	-	-	-	-	-	-	475bp
475	Hypothetical, phage related (19-BS75)	-	-	-	-	-	-	-	-	-	-	471bp
244	Hypothetical, putative conserved structural	-	-	-	-	1857	-	-	-	-	-	458bp
162	Phage protein (*S. pyogenes*)	-	-	-	-	-	-	-	-	-	-	440bp
164	Phage-related protein, TMP repeat family, membrane protein	-	-	-	-	1852	-	-	-	-	-	429bp
982	35s rRNA, 18s rRNA, 5.8s rRNA, 25s rRNA, 5s rRNA (*Saccharomyces cerevisiae*)	-	-	-	-	-	-	-	-	-	-	428bp
325	Putative phage integrase (14-BS69)	-	-	-	-	-	-	-	-	-	-	425bp
260	Hypothetical (14-BS69)	-	-	-	-	-	-	-	-	-	-	407bp
	Transcriptional regulator (*S. pyogenes*)	-	-	-	-	-	-	-	-	-	-	407bp

Contig	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Contig length
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
282	Prophage λBa01, membrane protein	-	-	-	-	1852	-	-	-	-	-	406bp
	Hypothetical	-	-	-	-	1851	-	-	-	-	-	406bp
271	Hypothetical (*S. suis*)	-	-	-	-	-	-	-	-	-	-	390bp
	Putative transcriptional regulator	-	-	-	-	1900	-	-	-	-	-	390bp
352	Prophage λBa01, membrane protein	-	-	-	-	1852	-	-	-	-	-	383bp
821	PezA	1050	0930	1188	12630	0987	1128	-	1262	-	1296	380bp
493	Hypothetical	1340	1175	-	12290	-	1380	-	1477	-	1454	365bp
223	Hypothetical (probably phage related)	-	-	-	-	1846	-	-	-	-	-	355bp
361	Prophage λBa01, membrane protein	-	-	-	-	1852	-	-	-	-	-	344bp
2109	Hypothetical (*S. aureus*)	-	-	-	-	-	-	-	-	-	-	343bp
995	Hypothetical	-	-	-	12580	-	-	-	1267	-	1291	341bp
507	Hypothetical	-	-	-	-	1898	-	-	-	-	-	332bp
752	Hypothetical (*S. mitis*)	-	-	-	-	-	-	-	-	-	-	330bp
454	PblB	-	-	0074	-	1850	0074	-	0062	-	-	318bp
359	Phage protein	-	-	0038	-	1899	-	-	-	-	-	318bp
335	Phage head morphogenesis protein, Spp1 gp7 family	-	-	-	-	1868	-	-	-	-	-	311bp
262	Hypothetical	-	-	-	-	1887	-	-	-	-	-	311bp
629	Phage head morphogenesis protein, Spp1 gp7 family	-	-	-	-	1868	-	-	-	-	-	311bp
311	Phage protein	-	-	-	-	1873	-	-	-	-	-	308bp
299	Hypothetical (14-BS69)	-	-	-	-	-	-	-	-	-	-	304bp
353	Phage terminase, large subunit (6-BS73)	-	-	-	-	-	-	-	-	-	-	303bp

Table A.3 BLAST results of strain 1 discrepancies

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
1	rRNA - 5s	-	-	-	-	-	-	-	-	0019	-	1bpΔ
2	rRNA - 23s	+	+	+	+	+	+	+	+	+	+	Poor assembly
	rRNA - 5s	+	+	+	+	+	+	+	+	+	+	Poor assembly
3	IS630-Spn1 Transposase	+	+	+	+	+	+	+	+	+	+	2 bpΔ, Poor assembly
4	IS630-Spn1 Transposase	+	+	+	+	+	+	+	+	+	+	Poor assembly
5	IS1167 Transposase	+	+	-	+	+	+	-	+	+	+	20bpΔ
	IS1167 Transposase	+	+	-	+	+	+	-	+	+	+	
6	Integrase	-	-	0028	-	-	0028	-	0026	-	-	33,844bpΔ
	Plasmid addiction system poinsin protein	-	-	0029	-	-	0029	-	0028	-	-	
	Conserved hypothetical	-	-	0032	00290	0034	-	-	-	-	-	
	Phage transcriptional regulator, Cro/CI family	-	-	0033	00300	0035	-	-	-	-	-	
	Phage protein	-	-	0034	00310	0036	0031	-	-	-	-	
	gp15	-	-	0038	-	-	-	-	-	-	-	
	Phage protein	-	-	0039	-	-	0035	-	0034	-	-	
	Hypothetical	-	-	0040	-	-	0036	-	0035	-	-	
	Hypothetical	-	-	0041	-	-	0037	-	0036	-	-	
	gp21	-	-	0043	-	-	0039	-	0038	-	-	
	gp24	-	-	0045	-	-	0041	-	0040	-	-	
	Hypothetical, phage related	-	-	0049	-	-	0045	-	0091	-	-	
	DNA N-4 cytosine methyl transferase	-	-	0050	-	-	0046	-	0092	-	-	
	Hypothetical	-	-	0053	-	-	0050	-	-	-	-	
	Hypothetical phage protein	-	-	0054	-	-	0051	-	-	-	-	
	Prophage ISa2 site-specific recombinase	-	-	0056	-	-	0054	-	0046	-	-	
	Phage-related hypothetical	-	-	0058	-	-	-	-	-	-	-	
	Phage-related terminase, small subunit	-	-	0059	-	-	-	-	-	-	-	
	Phage-related terminase	-	-	0060	-	-	-	-	-	-	-	
	Phage portal protein	-	-	0062	-	-	-	-	-	-	-	
	Capsid protein	-	-	0063	-	-	-	-	-	-	-	
	Phage maturation protease	-	-	0064	-	-	-	-	-	-	-	
	Prophage pi2 protein 39	-	-	0070	-	-	-	-	-	-	-	
	Unknown phage protein	-	-	0072	-	-	-	-	0060	-	-	

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
	Hypothetical	-	-	0073	-	-	-	-	0061	-	-	
	PblB	-	-	0074	-	1850	0074	-	0062	-	-	
	PblB	-	-	0075	-	1850	0074	-	0062	-	-	
	PblB	-	-	0076	-	1850	0074	-	0062	-	-	
	Phage holin 4 superfamily	-	-	0081	-	-	0077	-	0065	-	-	
	Prophage ISa2, holin LLH superfamily	-	-	0082	-	-	0078	-	0066	-	-	
9	Hypothetical	0029	0035	0093	-	0053	0089	0067	0128	0035	-	4bpΔ
	Competence induced protein Ccs16	0030	0036	-	-	0054	-	0068	-	-	-	
10	Ribose-phosphate pyrophosphokinase	0031	0037	0094	14440	0055	0091	0069	0130	0037	0030	15bpΔ
11	Ribose-phosphate pyrophosphokinase	0031	0037	0094	14440	0055	0091	0069	0130	0037	0030	17bpΔ
12	Hypothetical	0033	0039	0096	00500	0057	0093	0071	0132	0039	0033	Poor assembly
13	Membrane protein	0034	0040	0098	00520	0058	0094	0072	0134	0040	0035	220bpΔ
14	Fatty acid/synthesis protein plsX	0037	0043	0101	00550	0061	0097	0075	0137	0043	0039	27bpΔ
15	Bifunctional purine synthesis, PurH	0050	0057	0115	00670	0082	0117	0089	0157	0056	0052	677bpΔ
16	Bifunctional purine synthesis, PurH	0050	0057	0115	00670	0082	0117	0089	0157	0056	0052	2bpΔ
17	Bifunctional purine synthesis, PurH	0050	0057	0115	00670	0082	0117	0089	0157	0056	0052	7bpΔ
18	Phosphoribosylamine-glycine ligase	0051	0058	0116	00680	0083	0118	0090	0158	0057	0053	13bpΔ
19	Phosphoribosylamine-glycine ligase	0051	0058	0116	00680	0083	0118	0090	0158	0057	0053	44bpΔ
20	Phosphoribosylaminoimidazole carboxylase catalytic subunit	0053	0059	0118	00690	0084	0120	0091	0160	0058	0055	77bpΔ
21	Phosphoribosylaminoimidazole carboxylase catalytic subunit	0053	0059	0118	00690	0084	0120	0091	0160	0058	0055	1bpΔ, 73bpΔ
22	Phosphoribosylaminoimidazole carboxylase catalytic subunit	0053	0059	0118	00690	0084	0120	0091	0160	0058	0055	139bpΔ
23	Phosphorylase, Pnp/Udp family	0075	0074	0136	00870	0101	0138	0109	0177	0078	0072	99bpΔ
24	Cell wall surface anchor family	0082	0080	0144	-	0108	0145	0116	0184	0083	0080	Poor assembly
25	Cell wall surface anchor family	0082	0080	0144	-	0108	0145	0116	0184	0083	0080	1bpΔ
26	Cell wall surface anchor family	0082	0080	0144	-	0108	0145	0116	0184	0083	0080	1bpΔ
28	Cell wall surface anchor family	0082	0080	0144	-	0108	0145	0116	0184	0083	0080	Poor assembly
30	ABC transporter, permease	0091	0089	0152	01030	0116	0152	0123	0192	0090	0088	19bpΔ
30	ABC transporter, substrate binding	0092	0090	0153	01040	0117	0153	0124	0193	0091	0089	19bpΔ
31	Hypothetical	0093	0093	0156	01070	0120	0156	0131	0196	0094	0092	1bpΔ
32	Hypothetical	0093	0093	0156	01070	0120	0156	0131	0196	0094	0092	8bpΔ, 29bpΔ
33	Bacteriocin	0109	0106	0168	01202	0132	0172	0143	0211	-	0105	183bpΔ
34	Helix-turn-helix domain	0114	0112	0173	17070	0139	0179	0150	0208	0109	0104	92bpΔ

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
38	Hypothetical	0115	0116	0174	-	0143	0183	0149	0218	0116	0112	384bpΔ
39	Hypothetical	-	0123	0181	-	0145	0194	0160	0227	0118	0118	14bpΔ
40	Putative membrane protein	-	0124	0182	01270	0146	0195	0161	0228	0111	0119	12bpΔ
42	PspA	0117	0126	0185	-	-	0197	0163	0232	0121	0120	47bpΔ
44	PspA	0117	0126	0185	-	-	0197	0163	0232	0121	0120	85bpΔ
45	PspA	0117	0126	0185	-	-	0197	0163	0232	0121	0120	52bpΔ
46	PspA	0117	0126	0185	-	-	0197	0163	0232	0121	0120	40bpΔ
47	PspA	0117	0126	0185	-	-	0197	0163	0232	0121	0120	2bpΔ
48	PspA	0117	0126	0185	-	-	0197	0163	0232	0121	0120	18bpΔ
49	PspA	0117	0126	0185	-	-	0197	0163	0232	0121	0120	Poor assembly
50	tRNA	0118	0127	0186	01320	0151	0198	0165	0234	0123	0121	32bpΔ
51	Metallo-b-lactamase superfamily	0121	0130	0190	01350	0155	0201	0169	0237	0126	0124	98bpΔ
52	Competence induced protein Ccs1	0123	-	-	-	-	-	0171	0239	-	0127	2bpΔ
53	Glycoprotein endopeptidase, M22 peptidase	0127	0134	0195	01390	0159	0205	0174	0242	0130	0130	18bpΔ
54	Glycoprotein endopeptidase, M22 peptidase	0127	0134	0195	01390	0159	0205	0174	0242	0130	0130	Poor assembly
55	Glycoprotein endopeptidase, M22 peptidase	0129	0136	0197	01410	0161	0207	0176	0244	0132	0132	93bpΔ
56	Glycosyl transferase, group 2 family	0136	0139	-	01460	0166	-	0181	0249	0137	0137	305bpΔ
57	ABC transporter, substrate binding	0148	0150	0217	01570	0177	0222	0195	0260	0150	0152	3bpΔ
58	Sensor histidine kinase, hk07	0155	0157	0224	01640	0184	0229	0202	0267	0157	-	56bpΔ
59	Hypothetical	0160	0162	0229	01690	0189	0234	0208	0274	0162	0186	2bpΔ
60	Response regulator, LytR/AlgR family	0161	0163	0230	01700	0190	0235	0209	0275	0163	0187	65bpΔ
	Hypothetical	0162	-	-	-	-	-	-	-	-	-	
61	Dihydrofolate folylpolyglutamate synthetase	0197	0183	0249	-	0209	0255	0245	0313	0184	0208	109bpΔ
62	Anaerobic ribonucleoside-triphosphate reductase activating protein	0205	0190	0257	01960	0216	0262	0253	0320	0192	0215	Poor assembly
63	Anaerobic ribonucleoside-triphosphate reductase activating protein	0205	0190	0257	01960	0216	0262	0253	0320	0192	0215	3bpΔ
64	Anaerobic ribonucleoside-triphosphate reductase activating protein	0205	0190	0257	01960	0216	0262	0253	0320	0192	0215	3bpΔ
66	Hypothetical	0244	-	-	-	-	-	-	-	-	-	245bpΔ
67	Leucyl-tRNA Synthetase	0254	0238	0309	02440	0266	0316	0302	0372	0241	0265	22bpΔ, 36bpΔ
68	Leucyl-tRNA Synthetase	0254	0238	0309	02440	0266	0316	0302	0372	0241	0265	26bpΔ
69	Leucyl-tRNA Synthetase	0254	0238	0309	02440	0266	0316	0302	0372	0241	0265	8bpΔ
70	Leucyl-tRNA Synthetase	0254	0238	0309	02440	0266	0316	0302	0372	0241	0265	28bpΔ
71	Holliday junction DNA helicase RuvB	0259	0241	0309	02470	0269	0319	0306	0376	0244	0269	89bpΔ

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
72	Holliday junction DNA helicase RuvB	0259	0241	0309	02470	0269	0319	0306	0376	0244	0269	57bpΔ
74	Holliday junction DNA helicase RuvB	0259	0241	0309	02470	0269	0319	0306	0376	0244	0269	21bpΔ
	Hypothetical	0260	0242	0310	02480	0270	0320	-	0377	0245	0270	
75	DNA polymerase III Subunit a	0274	0254	0323	02600	0282	0332	0319	0390	0257	0285	14bpΔ
76	DNA polymerase III Subunit a	0274	0254	0323	02600	0282	0332	0319	0390	0257	0285	72bpΔ
77	DNA polymerase III Subunit a	0274	0254	0323	02600	0282	0332	0319	0390	0257	0285	7bpΔ
78	DNA polymerase III Subunit a	0274	0254	0323	02600	0282	0332	0319	0390	0257	0285	7bpΔ
79	DNA polymerase III Subunit a	0274	2054	0323	02600	0282	0332	0319	0390	0257	0285	99bpΔ
80	Xanthine/uracil permease family	0287	2067	0337	02760	0296	0346	0333	0403	0272	0300	70bpΔ
81	CAAX amino terminal protease family	0288	2068	0338	02770	0297	0347	0334	0404	0273	0301	54bpΔ
82	CAAX amino terminal protease family	0288	0268	0338	02770	0297	0347	0334	0404	0273	0301	18bpΔ
	Dihydropteroate synthase	0289	0269	0339	02780	0298	0348	0335	0406	0274	0302	
83	Dihydropteroate synthase	0289	0269	0339	02780	0298	0348	0335	0406	0274	0302	Poor assembly
84	Dihydropteroate synthase	0289	0269	0339	02780	0298	0348	0335	0406	0274	0302	11bpΔ
85	Integrase/recombinase, phage integrase family	-	0261	0330	02690	0289	0340	0326	0397	0281	0292	483bpΔ
86	Hyaluronate lyase, HylA	0314	0287	0350	02890	0310	0379	0364	0426	0285	0322	395bpΔ
88	s-adenosyl-methyltransferase, MraW	0334	0304	0368	03060	0327	0397	0384	0444	0303	0337	Poor assembly
92	Hypothetical, Putative ATP-dependent Zn protease	0341	0310	0376	03140	0333	0405	0390	0451	0310	0344	43bpΔ
93	IS630-Spn1 Transposase orf 2	+	+						+	+	+	3bpΔ
94	Transposase	+	+						+	+	+	Poor assembly
95	Transposase	+	+						+	+	+	Poor assembly
96	Transposase	+	+						+	+	+	90bpΔ
97	Oligopeptide ABC transporter, Oligo-peptide binding protein, AliA	0366	0334	0405	03390	0353	0438	0410	0474	0332	0363	3bpΔ
98	Endo-a-N-acetylgalactosaminidase, cell wall surface anchor family	0368	0335	0406	03400	0354	0439	0411	0475	0333	0364	98bpΔ, 4bpΔ, 96bpΔ
99	Endo-a-N-acetylgalactosaminidase, cell wall surface anchor family	0368	0335	0406	03400	0354	0439	0411	0475	0333	0364	6bpΔ
100	Endo-a-N-acetylgalactosaminidase, cell wall surface anchor family	0368	0335	0406	03400	0354	0439	0411	0475	0333	0364	7bpΔ
101	CbpF	0377	0345	0415	03650	0363	0448	0420	0484	0343	0373	6bpΔ
102	Phosphomevalonate kinase	0382	0347	0420	03540	0369	0453	0427	0489	0347	0377	2bpΔ
103	Phosphomevalonate kinase	0382	0347	0420	03540	0369	0453	0427	0489	0347	0377	3bpΔ
104	Sensor histidine kinase, hk03	0386	0351	0424	03580	0373	0457	0431	0493	0351	0381	11bpΔ
105	DNA alkylation repair enzyme	-	-	-	-	0375	0459	-	-	0353	0384	Poor assembly
	DNA alkylation repair enzyme	-	-	-	-	-	-	-	-	0354	0385	

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
107	Isopentenyl pyrophosphate isomerase	-	-	0431	-	-	-	0437	-	-	-	3bpΔ, 376bpΔ
108	Signal peptidase I, ABC transporter, Substrate-binding	-	-	0717	-	0755	-	-	-	0639	0401	29bpΔ
	Helicase, RecD/TraA family	0401	0366	0433	03770	0389	0473	0439	0510	0368	0402	
109	Signal peptidase I, lepB	0402	0367	0434	03790	0390	0474	0440	0511	0639	0401	4bpΔ
109	Ribonuclease HIII, rnhC -RNaseH	0403	0368	0435	03800	0391	0475	0441	0512	0370	0402	4bpΔ
110	Conserved hypothetical	0429	0391	0459	04030	0414	0499	0465	0537	0394	0426	7bpΔ
	Hypothetical	0430	0392	0460	04040	-	-	-	0538	0395	-	76bpΔ
112	L-iditol 2-dehydrogenase	-	-	0466	-	-	-	0472	-	-	-	22bpΔ
113	Threonine dehydratase	0454	0414	0486	04270	0439	0525	0492	0562	0418	0449	80bpΔ
114	Threonine dehydratase	0454	0414	0486	04270	0439	0525	0492	0562	0418	0449	56bpΔ
115	Transposase	+	+	+	+	+	+	+	+	+	+	12bpΔ
116	Transposase	+	+	+	+	+	+	+	+	+	+	9bpΔ
120	Conserved hypothetical	-	-	-	-	-	-	-	-	0427	-	157bpΔ
121	ABC transporter, ATP-binding	0483	0434	0509	04410	0455	0547	0525	0596	0439	0465	63bpΔ
122	Conserved hypothetical	0490	0439	0514	04460	0460	0522	0530	0601	0444	0470	207bpΔ
122	Potassium transporter, peripheral membrane component	0491	0440	0515	04470	0461	0553	0531	0602	0445	0471	207bpΔ
123	IS1380-Spn1, Transposase	+	+	+	+	+	+	+	+	+	+	1,700bpΔ
124	Type I site-specific deoxyribonuclease chains	0505	0453	0527	04610	0476	0566	0543	0614	0460	0483	1,454bpΔ
	Integrase/recombinase, phage integrase family	0506	0452	0528	04600	0475	-	-	0615	-	-	
125	BlpT protein, fusion	0524	0466	0558	04670	0490	0581	0558	0630	0474	0496	58bpΔ
126	Response regulator, BlpR	0526	0468	0548	04780	0492	0583	0561	0632	0476	0498	31bpΔ
127	Histidine Kinase, BlpH	0527	0469	0549	04790	0493	0584	0562	0633	0477	0499	3bpΔ
132	Immunity protein, BlpY	0545	0473	0564	04940	0507	0607	0577	0644	0493	0510	18bpΔ
133	Immunity protein, BlpY	0545	0473	0564	04940	0507	0607	0577	0644	0493	0510	68bpΔ
134	Conserved domain protein, CAAX amino terminal protease family	0547	0475	0566	04950	0509	0609	0579	0646	0495	0512	35bpΔ
135	Conserved Hypothetical	0559	0485	0576	05050	0519	0622	0589	0656	0506	-	106bpΔ
135	Conserved hypothetical	-	0486	-	-	-	0623	0590	0657	0507	-	106bpΔ
136	Hemerythrin HHE cation binding domain subfamily	0562	0488	0579	05080	0522	0627	0593	0660	0509	0509	28bpΔ
137	Hemerythrin HHE cation binding domain subfamily	0562	0488	0579	05080	0522	0627	0593	0660	0509	0509	32bpΔ
138	Hemerythrin HHE cation binding domain subfamily	0562	0488	0579	05080	0522	0627	0593	0660	0509	0509	33bpΔ
141	Acetyltransferase, GNAT family	0566	0492	0583	05120	0527	0631	0597	0664	0514	0529	79bpΔ
142	Valyl-tRNA synthetase	0566	0492	0583	05120	0527	0631	0597	0664	0514	0529	79bpΔ
143	Valyl-tRNA synthetase	0566	0492	0583	05120	0527	0631	0597	0664	0514	0529	14bpΔ
144	Valyl-tRNA synthetase	0566	0492	0583	05120	0527	0631	0597	0664	0514	0529	39bpΔ
145	Valyl-tRNA synthetase	0566	0492	0583	05120	0527	0631	0597	0664	0514	0529	105bpΔ
146	Valyl-tRNA synthetase	0566	0492	0583	05120	0527	0631	0597	0664	0514	0529	2bpΔ

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
146	Conserved hypothetical	-	0495	0586	05150	0530	0634	-	-	0517	-	2bpΔ
147	Conserved hypothetical	-	0495	0586	05150	0530	0634	-	-	0517	-	21bpΔ
148	Conserved hypothetical	-	0495	0586	05150	0530	0634	-	-	0517	-	169bpΔ
149	IS3-Spn1, transposase	+	+	+	+	+	+	+	+	+	+	8bpΔ
150	IS3-Spn1, transposase	+	+	+	+	+	+	+	+	+	+	5bpΔ
151	IS3-Spn1, transposase	+	+	+	+	+	+	+	+	+	+	Poor assembly
152	Nitroreductase family protein	0574	-	-	05190	0532	-	-	0673	-	0538	2,656bpΔ
152	DEAD/DEAH Box helicase	0575	-	-	05190	0532	-	-	0673	-	0541	2,656bpΔ
153	Phenylalanyl-tRNA synthetase, b subunit	0581	0506	0597	05250	0536	0645	0611	0680	0529	0547	5bpΔ
154	Phenylalanyl-tRNA synthetase, b subunit	0581	0506	0597	05250	0536	0645	0611	0680	0529	0547	40bpΔ
155	Phenylalanyl-tRNA synthetase, b subunit	0581	0506	0597	05250	0536	0645	0611	0680	0529	0547	63bpΔ
156	Phenylalanyl-tRNA synthetase, b subunit	0581	0506	0597	05250	0536	0645	0611	0680	0529	0547	30bpΔ
157	Phenylalanyl-tRNA synthetase, b subunit	0581	0506	0597	05250	0536	0645	0611	0680	0529	0547	1bpΔ
	Endonuclease/exonuclease/phosphatase family	0582	0507	0598	05260	0537	0646	0612	0681	0530	0548	
159	Transcriptional regulator	-	-	0601	05280	0539	0649	0614	0683	-	0550	129bpΔ
161	Cycteinyl-tRNA synthetase	0591	0515	0608	05340	0545	0655	0620	0689	0539	0556	9bpΔ
162	Ribonuclease III family	0592	0516	0609	05350	0546	0656	0621	0690	0540	0557	6bpΔ
163	Hypothetical	0595	0518	-	05380	0549	0658	0642	0692	0542	0559	562bpΔ
164	Transposase	-	0520	0613	03120	0551	0660	0625	-	-	-	45bpΔ, Poor assembly
165	Excinuclease ABC subunit C	0618	0538	0632	05570	0569	0679	0643	0711	0561	0578	30bpΔ
166	Excinuclease ABC subunit C	0618	0538	0632	05570	0569	0679	0643	0711	0561	0578	abutting
167	Excinuclease ABC subunit C	0618	0538	0632	05570	0569	0679	0643	0711	0561	0	6bpΔ
168	Excinuclease ABC subunit C	0618	0538	0632	05570	0569	0679	0643	0711	0561	0578	81bpΔ
169	Excinuclease ABC subunit C	0618	0538	0632	05570	0569	0679	0643	0711	0561	0578	7bpΔ, 6bpΔ
169	Ser/Thr protein phosphatase family, metallophosphoesterase	0619	0539	0633	05580	-	0680	0644	0712	0562	0579	7bpΔ, 6bpΔ
170	Ser/Thr protein phosphatase family, metallophosphoesterase	0619	0539	0633	05580	-	0680	0644	0712	0562	0579	51bpΔ
171	Ser/Thr protein phosphatase family, metallophosphoesterase	0619	0539	0633	05580	-	0680	0644	0712	0562	0579	53bpΔ
172	Nitroreductase family	0622	0541	0636	05610	0573	0682	0646	0714	-	0583	144bpΔ
173	Dipeptidase, PepV, Peptidase M28	0623	0542	0637	05620	0574	0683	0647	0715	0566	0584	12bpΔ
174	M42 peptidase	0627	0547	0642	05660	0579	0688	0652	0720	0571	0588	24bpΔ
	HIT family protein, histidine triad domain	0628	0548	0643	05670	0580	0689	0653	0721	0572	0589	
175	HIT family protein, histidine triad domain	0628	0548	0643	05670	0580	0689	0653	0721	0572	0589	4bpΔ
176	HIT family protein, histidine triad domain	0628	0548	0643	05670	0580	0689	0653	0721	0572	0589	81bpΔ
177	IS1380 Transposase, OrfB	+	+	+	+	+	+	+	+	+	+	1bpΔ, poor assembly
178	D-alanyl-D-alanine carboxypeptidase	0629	0544	0639	05640	0576	0685	0649	0717	0573	0586	10bpΔ, 71bpΔ
183	Cell wall associated serine protease, subtilase family, PrtA	0641	0558	0657	05790	0592	0702	0666	0733	0584	0599	Poor assembly
184	Cell wall associated serine protease, subtilase family, PrtA	0641	0558	0657	05790	0592	0702	0666	0733	0584	0599	55bpΔ

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
185	Cell wall associated serine protease, subtilase family, PrtA	0641	0558	0657	05790	0592	0702	0666	0733	0584	0599	148bpΔ
186	Cell wall associated serine protease, subtilase family, PrtA	0641	0558	0657	05790	0592	0702	0666	0733	0584	0599	28bpΔ
187	Cell wall associated serine protease, subtilase family, PrtA	0641	0558	0657	05790	0592	0702	0666	0733	0584	0599	9bpΔ
188	Cell wall associated serine protease, subtilase family, PrtA	0641	0558	0657	05790	0592	0702	0666	0733	0584	0599	92bpΔ
189	Cell wall associated serine protease, subtilase family, PrtA	0641	0558	0657	05790	0592	0702	0666	0733	0584	0599	118bpΔ
190	Cell wall associated serine protease, subtilase family, PrtA	0641	0558	0657	05790	0592	0702	0666	0733	0584	0599	11bpΔ
191	Cell wall associated serine protease, subtilase family, PrtA	0641	0558	0657	05790	0592	0702	0666	0733	0584	0599	2bpΔ
192	Cell wall associated serine protease, subtilase family, PrtA	0641	0558	0657	05790	0592	0702	0666	0733	0584	0599	11bpΔ
200	PTS system IIA component	0645	0559	0661	05800	0593	0703	0667	0738	0585	0600	Poor assembly
201	PTS system IIB	0646	0560	0662	05810	0594	0704	0668	0739	0586	0601	Poor assembly
202	b-galactosidase	0648	0562	0665	05830	0596	0707	0670	0741	0588	0603	198bpΔ
203	b-galactosidase	0648	0562	0665	05830	0596	0707	0670	0741	0588	0603	260bpΔ
204	b-galactosidase	0648	0562	0665	05830	0596	0707	0670	0741	0588	0603	2bpΔ
205	b-galactosidase	0648	0562	0665	05830	0596	0707	0670	0741	0588	0603	11bpΔ
206	Hypothetical, transposase-related	-	-	0667	05850	0598	0709	0673	0743	-	-	Poor assembly
207	Hypothetical, transposase-related	-	-	0667	05850	0598	0709	0673	0743	-	-	8bpΔ
208	ZmpB	0664	0577	0684	05990	1074	0723	0688	0759	0605	0620	166bpΔ
209	ZmpB	0664	0577	0684	05990	1074	0723	0688	0759	0605	0620	2,322bpΔ
210	Conserved hypothetical	0666	-	0686	-	-	0725	0690	0761	0607	0622	29bpΔ
211	Choline binding protein, pneumococcal surface protein	0667	0579	-	06030	-	0726	0691	0762	0608	0623	11bpΔ
212	Glucokinase	0668	0580	0689	06040	0619	0727	0692	0763	0609	0624	3bpΔ
213	ABC-type amino acid transport/signal transduction system, substrate binding	-	-	0717	-	-	0755	-	0792	0638	0658	10bpΔ
214	ABC transporter, ATP-binding, glutamine transport	0709	0616	-	06340	0649	-	-	-	0641	0659	3bpΔ
215	ABC transporter membrane-spanning permease, glutamine transport	0710	0617	0722	06360	0651	0758	0727	0797	-	0660	4bpΔ
	ABC transporter membrane-spanning permease, glutamine transport	0711	0618	0722	06360	0651	0758	0727	0797	0645	-	
217	Lysyl-tRNA Synthetase	-	0690	0727	-	-	0792	-	0803	-	0665	8bpΔ
218	ABC transporter, ATP-binding protein, cobalt/nickel transport	0720	0626	0731	06450	0660	0767	0736	0807	0653	0669	1bpΔ
219	ABC transporter, ATP-binding protein, cobalt/nickel transport	0720	0626	0731	06450	0660	0767	0736	0807	0653	0669	11bpΔ
220	IS630-spnII, transposase	+	+	+	+	+	+	+	-	0665	-	1bpΔ
221	Transposase	+	+	+	+	+	+	+	+	+	+	1bpΔ, 70bpΔ
222	Transposase	+	711	+	+	+	+	+	+	+	+	12bpΔ
222	Transposase	+	712	+	+	+	+	+	+	+	+	12bpΔ
223	Transposase	+	+	+	+	+	+	+	+	+	+	6bpΔ
224	Sodium dependent transporter	0737	0642	-	-	0676	-	0753	0838	0669	0687	Poor assembly
225	Sodium dependent transporter	0737	0642	-	-	0676	-	0753	0838	0669	0687	1,896bpΔ
	MerR Family transcriptional regulator, regulator of pmrA	0739	-	0750	06630	0678	-	-	-	0670	0689	
	MuT/Nudix family protein	0740	0644	0751	06640	0679	0786	0755	-	0671	0690	

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
228	Hydrolase a,b fold family	0777	0677	0785	06980	0714	0819	1424	0875	0714	0725	19bpΔ
229	Glutathione-disulfide reductase	0784	0685	0793	07100	0726	0827	1416	0883	0714	0714	Poor assembly
230	Membrane-fusion protein, HlyD family secretion protein	0785	0686	0794	07110	0727	0828	1415	0884	0715	0734	7bpΔ
230	Macrolide export, ATP-binding/permease	0786	0687	0795	07120	0728	0829	1414	0885	0716	0735	7bpΔ
231	3-ketoacyl-acyl-carrier protein reductase	-	+	+	+	+	+	+	+	+	+	6bpΔ
232	Conserved hypothetical	0792	0694	0801	07190	0736	0836	1407	0894	0722	0714	110bpΔ
233	DNA gyrase, b-subunit	0806	-	0815	07350	0752	0851	1393	0908	-	0755	23bpΔ
235	IS630-spnII, transposase	+	+	+	+	+	+	+	+	+	+	1bpΔ, Poor assembly
236	IS630-spnII, transposase	+	+	+	+	+	+	+	+	+	+	43bpΔ
237	IS630-spnII, transposase	+	+	+	+	+	+	+	+	+	+	1bpΔ
238	IS630-spnII, transposase	+	+	+	+	+	+	+	+	+	+	1bpΔ
239	IS630-spnII, transposase	+	+	+	+	+	+	+	+	+	+	Poor assembly
240	IS3-Spn1, transposase	+	+	+	+	+	+	+	+	+	+	3bpΔ
241	IS3-Spn1, transposase	+	+	+	+	+	+	+	+	+	+	10bpΔ
242	IS3-Spn1, transposase	+	+	+	+	+	+	+	+	+	+	Poor assembly
243	IS3-Spn1, transposase	+	+	+	+	+	+	+	+	+	+	10bpΔ
244	IS3-Spn1, transposase	+	+	+	+	+	+	+	+	+	+	4bpΔ
245	Amino acid ABC transporter, ATP-binding protein	0824	0720	0831	07450	0762	0861	1377	0923	0746	0765	70bpΔ
246	Amino acid ABC transporter, ATP-binding protein	0824	0720	0831	07450	0762	0861	1377	0923	0746	0765	Poor assembly
247	5, 10-methylene-tetrahydrofolate dehydrogenase, FolD	0825	0721	0832	07460	0763	0862	1376	0924	0747	0766	109bpΔ
248	5, 10-methylene-tetrahydrofolate dehydrogenase, FolD	0825	0721	0832	07460	0763	0862	1376	0924	0747	0766	7bpΔ
248	YjeF-like protein	-	0722	0833	07480	0765	0863	1375	0925	0748	0768	7bpΔ
249	YjeF-like protein	-	0722	0833	07480	0765	0863	1375	0925	0748	0768	205bpΔ
250	YjeF-like protein	-	0722	0833	07480	0765	0863	1375	0925	0748	0768	58bpΔ
251	YjeF-like protein	-	0722	0833	07480	0765	0863	1375	0925	0748	0768	434bpΔ
251	Ribose-phosphate isomerase A	0828	0723	0834	07490	0766	0864	1374	0926	0749	0770	434bpΔ
252	Ribose-phosphate isomerase A	0828	0723	0834	07490	0766	0864	1374	0926	0749	0770	4bpΔ, 18bpΔ
253	Ribose-phosphate isomerase A	0828	0723	0834	07490	0766	0864	1374	0926	0749	0770	22bpΔ
	Phosphopentomutase	0829	0724	0835	07500	0767	0865	1373	0927	0750	0771	
254	Phosphopentomutase	0829	0724	0835	07500	0767	0865	1373	0927	0750	0771	28bpΔ
255	Phosphopentomutase	0829	0724	0835	07500	0767	0865	1373	0927	0750	0771	36bpΔ
256	Phosphopentomutase	0829	0724	0835	07500	0767	0865	1373	0927	0750	0771	17bpΔ
258	Hypothetical	0830	-	0836	07510	0768	0866	1372	0928	0751	0772	374bpΔ
	Purine nucleoside phosphorylase, family 2 protein	0831	0726	0837	07520	0769	0868	1371	0929	0752	0773	
259	Purine nucleoside phosphorylase, family 2 protein	0831	0726	0837	07520	0769	0868	1371	0929	0752	0773	60bpΔ
260	Purine nucleoside phosphorylase, family 2 protein	0831	0726	0837	07520	0769	0868	1371	0929	0752	0773	53bpΔ
261	Purine nucleoside phosphorylase, family 2 protein	0831	0726	0837	07520	0769	0868	1371	0929	0752	0773	36bpΔ

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
262	Methyl transferase small domain	0841	0735	0847	07620	0778	0877	1362	0940	0761	0783	123bpΔ
263	Methyl transferase small domain	0841	0735	0847	07620	0778	0877	1362	0940	0761	0783	2bpΔ
264	IS3-Spn1, transposase	+	+	+	+	+	+	+	+	+	+	2bpΔ, 6bpΔ
266	PTS system, fructose specific IIABC components	0877	0773	0885	07990	0818	0915	1324	0981	0801	0826	30bpΔ
267	SpoE family protein	0878	0774	0886	08010	0820	0916	1323	0983	0803	0827	89bpΔ
267	Hypothetical	0879	0775	0887	08020	0821	0917	1322	0984	0804	0828	89bpΔ
268	Aminotransferase, class V	0880	0776	0888	08030	0822	0918	1321	0985	0805	0829	17bpΔ
269	Aminotransferase, class V	0880	0776	0888	08030	0822	0918	1321	0985	0805	0829	98bpΔ
270	Aminotransferase, class V	0880	0776	0888	08030	0822	0918	1321	0985	0805	0829	20bpΔ
271	Probable thiamine biosynthesis tRNA modification protein, thiI	0881	0777	0889	08040	0823	0919	1320	0986	0806	0830	29bpΔ
272	Probable thiamine biosynthesis tRNA modification protein, thiI	0881	0777	0889	08040	0823	0919	1320	0986	0806	0830	194bpΔ
273	Type I restriction modification system, S subunit	0887	0783	0894	08090	0832	0925	1313	0993	0813	0841	435bpΔ
274	Type I restriction modification system, S subunit	0887	0783	0894	08090	0832	0925	1313	0993	0813	0841	440bpΔ
276	DNA polymerase III Subunit a	0895	0788	0903	08170	0836	0934	1304	1002	0821	0871	2bpΔ
277	DNA polymerase III Subunit a	0895	0788	0903	08170	0836	0934	1304	1002	0821	0871	165bpΔ
278	DNA polymerase III Subunit a	0895	0788	0903	08170	0836	0934	1304	1002	0821	0871	2bpΔ
279	DNA polymerase III Subunit a	0895	0788	0903	08170	0836	0934	1304	1002	0821	0871	15bpΔ
282	S1 RNA binding domain	0908	0802	0915	08290	0847	0946	1293	1014	0833	0885	40bpΔ
283	S1 RNA binding domain	0908	0802	0915	08290	0847	0946	1293	1014	0833	0885	129bpΔ
284	Zinc-dependent metalloprotease	0909	-	0916	08300	0848	0947	1292	1015	-	0885	35bpΔ
287	Saccharopine dehydrogenase	0919	0812	0927	08430	0860	0958	1280	1028	0845	0896	18bpΔ
288	Saccharopine dehydrogenase	0919	0812	0927	08430	0860	0958	1280	1028	0845	0896	71bpΔ
289	Carboxynorspermidine decarboxylase	0920	0813	0928	08440	0861	0959	1279	1029	0846	0897	3bpΔ
290	Cof family	0923	0816	0931	08470	0864	0962	1276	1032	0849	0900	56bpΔ
291	g-glutamyl phosphate reductase, proA	0932	0823	0939	08550	0872	0971	1267	1041	0859	0907	24bpΔ
292	g-glutamyl phosphate reductase, proA	0932	0823	0939	08550	0872	0971	1267	1041	0859	0907	23bpΔ
293	g-glutamyl phosphate reductase, proA	0932	0823	0939	08550	0872	0971	1267	1041	0859	0907	59bpΔ
294	Tetrapyrrole methylase family	0938	0828	0944	08600	0877	0976	1262	1046	0864	0912	1bpΔ
295	Degenerate transposase	+	+	+	+	+	+	+	-	+	+	14bpΔ
296	tRNA uridine 5-carboxymethylaminomethyl modification enzyme	0943	0833	0950	08650	0882	0982	1261	1047	0870	0917	61bpΔ
297	DNA internalisation-related competence protein ComEC/Rec2	0955	0844	0962	08800	0896	0995	1247	1055	0878	0931	21bpΔ
298	Hypothetical	0956	-	-	-	-	0996	-	1056	0879	0932	56bpΔ
298	ABC transporter, ATP-binding protein	0957	0845	0963	08820	0897	0997	1246	1057	0880	0933	56bpΔ
299	ABC transporter, ATP-binding protein	0957	0845	0963	08820	0897	0997	1246	1057	0880	0933	56bpΔ
300	Ribosome recycling factor Hypothetical	0958 -	0846 -	0964 -	08830 -	0898 0899	0998 -	1245 -	1058 1059	0881 0882	0934 -	46bpΔ
301	Endo-b-N-acetylglucosaminidase, LytB	0965	0853	0971	08900	0906	1005	1238	1067	0889	0941	9bpΔ, 6bpΔ

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
302	Putative metalloprotease	0967	0855	0973	08920	0908	1007	1236	1068	0891	0943	7bpΔ
303	Formamidopyrimidine-DNA glycosylase	0970	0858	0976	08950	0911	1010	1233	1071	0970	0946	4bpΔ
304	Dephospho-CoA kinase	0971	0859	0977	08960	0912	1011	1232	1072	0971	0947	5bpΔ
305	Transposase related	+	+	+	+	+	+	+	+	+	+	8bpΔ
306	Transposase related	+	+	+	+	+	+	+	+	+	+	4bpΔ, 1bpΔ
307	Transposase related	+	+	+	+	+	+	+	+	+	+	13bpΔ
309	Pneumococcal histidine triad protein B, PhtB	1003	0889	1009	09290	0944	1043	1198	1104	0928	0977	2bpΔ
310	Pneumococcal histidine triad protein B, PhtB	1003	0889	1009	09290	0944	1043	1198	1104	0928	0977	52bpΔ
311	Pneumococcal histidine triad protein B, PhtB	1003	0889	1009	09290	0944	1043	1198	1104	0928	0977	7bpΔ
312	Pneumococcal histidine triad protein B, PhtB	1003	0889	1009	09290	0944	1043	1198	1104	0928	0977	20bpΔ
313	Pneumococcal histidine triad protein B, PhtB	1003	0889	1009	09290	0944	1043	1198	1104	0928	0977	20bpΔ
314	Pneumococcal histidine triad protein B, PhtB	1003	0889	1009	09290	0944	1043	1198	1104	0928	0977	2bpΔ
315	Pneumococcal histidine triad protein B, PhtB	1003	0889	1009	09290	0944	1043	1198	1104	0928	0977	3bpΔ
316	Ferrochelatase, hemH	1009	0895	-	09340	0949	1048	1062	1111	0935	0985	439bpΔ
317	tRNA modification GTPase	1016	0902	1022	09400	0956	1055	1069	1119	0942	0995	70bpΔ
318	Thiopene & furan oxidation protein, ThdF	1016	0902	1022	09400	0956	1055	1069	1119	0942	0995	25bpΔ
319	Thymidine kinase	1018	0904	1024	09420	0958	1057	1071	1121	0944	0997	1bpΔ
320	Thymidine kinase	1018	0904	1024	09420	0958	1057	1071	1121	0944	0997	516bpΔ
	Acetyltransferase, GNAT family	1019	0905	1025	-	-	1058	-	1122	-	-	
321	Peptide chain release factor 1, PrfA	1020	0906	1026	09430	0959	1059	1072	1123	0945	0998	115bpΔ
322	Serine hydroxymethyltransferase	1024	0910	1030	09470	0963	1063	1076	1127	0949	1002	12bpΔ
323	Serine hydroxymethyltransferase	1024	0910	1030	09470	0963	1063	1076	1127	0949	1002	13bpΔ
324	Serine hydroxymethyltransferase	1024	0910	1030	09470	0963	1063	1076	1127	0949	1002	16bpΔ
325	Hypothetical	1027	0913	1033	09500	0966	1066	1079	1130	0952	1005	19bpΔ
326	23s rRNA (uracil-5)-methyltransferase, RumA	1029	0914	1034	09510	0967	1067	1080	1131	0953	1007	88bpΔ
327	Neopullulanase	-	0927	-	09660	0983	1125	-	1147	0972	1026	6,317bpΔ
	Hypothetical	1047	0928	1051	09670	0984	1126	1104	1148	-	1027	
	PezA, TA system	1050	0930	1053	09700	-	1128	-	-	-	1029	
	PezT, TA system	1051	0931	1054	09710	0988	1129	-	-	-	1030	
	Hypothetical	1052	0932	1055	09720	0989	1130	-	-	-	1031	
	Tn*5252* ORF 10 protein	1054	0934	1056	09740	0991	-	-	-	-	1033	
	Tn*5252* relaxase	1056	0936	-	-	-	1135	-	-	0973	1035	
328	Tn*5252* relaxase	1056	-	-	-	-	-	-	-	0973	1035	45bpΔ
329	Tn*5252* relaxase	1056	-	-	-	-	-	-	-	0974	1035	28bpΔ
330	Tn*5252* , relaxase	-	0938	-	09780	0995	1135	-	-	0975	-	9,505bpΔ
	Rgg/GadR/MutR family transcriptional regulator	-	0939	-	09790	0996	-	-	-	0976	-	
	3-hydroxyisobutyrate dehydrogenase	-	-	1062	-	-	-	-	-	0977	-	

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
	Hypothetical	-	-	1063	-	-	-	-	-	0978	-	
	Prephenate dehydratase	-	-	1064	-	-	-	-	-	0979	-	
	Hypothetical	-	-	1065	-	-	-	-	-	0980	-	
	Hypothetical	-	-	1066	-	-	-	-	-	0981	-	
	UDP-glucose-4-epimerase	-	-	1067	-	-	-	-	-	0982	-	
	Biotin carboxylase	-	-	1068	-	-	-	-	-	0983	-	
	Transporter, Major facilitator superfamily	-	0950	-	-	-	-	-	-	0985	-	
332	Degenerate transposase	+	+	+	+	+	+	+	+	+	+	1bpΔ, 10bpΔ
333	Transposase	+	+	+	+	+	+	+	+	+	+	25bpΔ
334	Transposase	+	+	+	+	+	+	+	+	+	+	6bpΔ
335	Transposase	+	+	+	+	+	+	+	+	+	+	41bpΔ
336	ABC transporter, ATP-binding protein	1114	0998	1119	10350	1052	1185	1160	1207	1032	1167	28bpΔ
338	1,4-a-glucan branching enzyme	1121	1005	1126	10140	1058	1191	1166	1214	1039	1159	Poor assembly
339	1,4-a-glucan branching enzyme	1121	1005	1126	10140	1058	1191	1166	1214	1039	1159	5bpΔ
342	IS630-Spn1, transposase	+	+	+	+	+	+	+	+	+	+	3bpΔ
343	Exonuclease, RexA	1151	1015	1137	10540	1070	1204	1178	1225	1050	1145	71bpΔ
344	IgA1 protease	1154	1018	1140	10580	1073	1207	1181	1229	1053	1143	63,754bpΔ
	ZmpD	-	-	1141	10590	1074	-	-	-	1054	1142	
	Replication initiator protein A, N-terminus, RepA	-	-	1144	13160	-	-	-	1232	1292	1340	
	Type II DNA modification methyltransferase	1336	-	1145	-	-	1071	-	1233	1291	1339	
	Tn*5253* , conserved hypothetical	1349	1183	1146	13140	1251	1072	-	1234	1290	1338	
	Caax amino protease	1346	1180	1149	13110	1248	1075	-	1237	1287	1336	
	Tn916, Transposase	-	-	1150	-	-	-	1931	1402	-	1334	
	RNA polymerase σ-70 region 4 family protein	-	-	1152	13070	-	-	1929	1406	-	1331	
	TetM	-	-	1155	13050	-	-	1919	1409	1231	171	
	Tn*916* , Hypothetical	-	-	1157	13040	-	-	1917	1411	1232	169	
	NLP/P60 family	-	-	1158	13030	-	-	1916	1412	1233	168	
	Conjugative transposon membrane protein	-	-	1159	13020	-	-	1915	1413	1234	167	
	Conjugative transposon protein	-	-	1160	13010	-	-	1914	1414	1235	166	
	Conjugative transposon membrane protein	-	-	1161	-	-	-	1913	1415	1236	165	
	Tn*5251* , Cro/CI family transcriptional regulator	-	-	1164	12970	-	-	1910	1422	1242	162	
	Tn*916* , FtsK/SpoIIIE family	-	-	1165	12960	-	-	1909	1423	1243	161	
	Tn*5253* , hypothetical	-	-	1170	12900	-	1076	-	1241	1281	1314	
	TraG/TraD family protein	-	-	1171	012890	-	1077	-	1242	1280	1313	
	Tn*5252* , Orf23	-	-	1173	12870	-	1079	-	1244	1278	1311	
	Type IV sec pathway, VirB4 component	-	-	1175	12850	-	-	-	1246	1276	1309	
	M23 peptidase	-	-	1176	12840	-	1083	-	1247	1275	1308	

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
	Hypothetical	-	-	1177	12830	-	-	-	1248	1274	1307	
	SNF2 family protein	-	-	1179	-	-	-	-	-	1272	1305	
	Parvulin-like peptidyl-prolyl isomerase	-	-	1183	12680	-	-	-	1257	1267	1301	
	DNA primase	-	-	1184	12670	-	-	-	1258	1266	1300	
	UmuD/MucA homolog, trasncriptional regulator	-	-	1190	-	-	-	-	-	-	-	
	UmuC/mucB homolog	-	-	1191	-	-	-	-	-	-	-	
	Tn*5252* , relaxase	-	-	1196	12430	-	-	-	-	1250	1276	
	Integrase	-	-	1198	-	-	-	-	-	-	-	
345	Integrase/recombinase, phage integrase family	1159	1023	1202	10630	1078	1211	1185	1277	1060	1138	12bpΔ
	Lipoate-protein ligase	1160	1024	1203	10640	1079	1212	1186	-	1061	1137	
346	IS1380, transposase	+	-	+	+	+	+	+	+	+	+	1,702bpΔ
347	PTS system IIA component	1172	-	-	-	-	-	-	-	-	-	9bpΔ
348	Dihydroxyacetone kinase, phosphotransfer	-	-	-	-	-	-	-	1291	-	-	9bpΔ
349	PhtD precursor	1174	1037	1217	10770	1093	1226	1049	1104	1073	1122	1bpΔ
350	PhtD precursor	1174	1037	1217	10770	1093	1226	1049	1104	1073	1122	13bpΔ
351	PhtD precursor	1174	1037	1217	10770	1093	1226	1049	1104	1073	1122	35bpΔ
352	PhtD precursor	1174	1037	1217	10770	1093	1226	1049	1104	1073	1122	19bpΔ
353	Hypothetical	1183	1045	1226	-	1101	1235	1041	1303	1080	1113	62bpΔ
354	PTS system, lactose-specific IIBC component	1185	1047	1228	10860	1103	1237	1039	1305	1082	1111	5bpΔ
355	PTS system, lactose-specific IIC component	1185	1047	1228	10860	1103	1237	1039	1305	1082	1111	1,234bpΔ
356	PTS system, lactose-specific IIA component	1198	0559	1239	10980	0593	0703	1028	0738	0585	1101	28bpΔ, 1bpΔ
357	Hypothetical	1213	1072	1252	11110	1129	1261	1015	1329	1106	1087	11bpΔ
358	Hypothetical	1213	1072	1252	11110	1129	1261	1015	1329	1106	1087	11bpΔ
359	O-acetylhomoserine sulfhydrylase	-	1073	-	-	-	1262	-	-	1107	1086	46bpΔ
360	Ion transporter, cation channel family, potassium	-	-	1259	11180	1136	1268	1006	1336	1113	1080	78bpΔ
361	Ion transporter, cation channel family, potassium	-	-	1259	11180	1136	1268	1006	1336	1113	1080	83bpΔ
362	Excinuclease ABC, B subunit, uvrB	1238	1096	1276	11340	1152	1301	0990	1353	1129	1064	93bpΔ
363	Excinuclease ABC, B subunit, uvrB	1238	1096	1276	11340	1152	1301	0990	1353	1129	1064	24bpΔ
364	Excinuclease ABC, B subunit, uvrB	1238	1096	1276	11340	1152	1301	0990	1353	1129	1064	4bpΔ
365	Excinuclease ABC, B subunit, uvrB	1238	1096	1276	11340	1152	1301	0990	1353	1129	1064	94bpΔ
366	Signal recognition particle-docking protein, FtsY	1244	1101	1281	11390	1157	1306	0985	1359	1134	1059	82bpΔ
367	Chromosome segregation protein, smc	1247	1104	1284	11420	1160	1309	0982	1362	1137	1056	77bpΔ
368	Chromosome segregation protein, smc	1247	1104	1284	11420	1160	1309	0982	1362	1137	1056	8bpΔ
369	Chromosome segregation protein, smc	1247	1104	1284	11420	1160	1309	0982	1362	1137	1056	3bpΔ
370	Chromosome segregation protein, smc	1247	1104	1284	11420	1160	1309	0982	1362	1137	1056	42bpΔ
371	Chromosome segregation protein, smc	1247	1104	1284	11420	1160	1309	0982	1362	1137	1056	30bpΔ
372	Chromosome segregation protein, smc	1247	1104	1284	11420	1160	1309	0982	1362	1137	1056	24bpΔ

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
373	Conserved hypothetical	1250	1108	1288	11450	1164	1313	0978	1366	1141	1053	102bpΔ
374	Conserved hypothetical	1250	1108	1288	11450	1164	1313	0978	1366	1141	1053	25bpΔ
375	Conserved hypothetical	1250	1108	1288	11450	1164	1313	0978	1366	1141	1053	24bpΔ
377	ABC transporter, ATP-binding	1282	1137	1321	11770	1197	1345	0945	1398	1176	1245	3bpΔ
378	ABC transporter, ATP-binding	1282	1137	1321	11770	1197	1345	0945	1398	1176	1245	10bpΔ
379	Uracil permease, pyrP	1286	1141	1325	11810	1201	1351	0941	1427	1180	1249	9bpΔ
380	HD superfmaily phosphohydrolase	1290	1145	1329	11850	1205	1355	0937	1431	1184	1253	17bpΔ
381	HD superfmaily phosphohydrolase	1290	1145	1329	11850	1205	1355	0937	1431	1184	1253	5bpΔ
382	Phosphatase, YidA	1291	1146	1330	11860	1206	1356	1291	1432	1185	1254	11bpΔ
383	High-affinity Fe2+/Pb2+ permease	1300	1155	1340	-	-	-	-	1442	1194	1267	1,696bpΔ
	High-affinity Fe2+/Pb2+ permease	-	1156	-	-	-	-	-	-	-	1268	
	Transposase	+	+	1343	+	+	+	+	+	+	+	
384	NAD (p)-specific glutamate dehydrogenase	1306	1158	1344	11970	1219	1369	0923	1446	1197	1272	17bpΔ
385	IS1380-Spn1, Transposase	+	-	+	+	+	+	+	+	+	+	2,225bpΔ
386	N-acteylneuraminate lyase	1329	-	1350	12210	1245	1374	1615	1471	1222	1648	18,026bpΔ
	Cytidine deaminase, EC 3.5.4.5	-	1164	1351	-	-	1371	-	-	-	-	
	Phoatidylglycerophoatase A, EC 3.1.3.27	-	1165	1352	-	-	1376	-	-	-	-	
	Glycoside hydrolase family protein, EC 3.2.1.26	-	1166	1353	-	-	1377	-	0827	-	-	
	ABC transporter, ATP-binding protein, oligopeptide transport	1888	1167	1354	-	-	1378	-	-	-	-	
	Oligopeptide transport, membrane spanning permease	1889	1168	1355	-	-	-	-	-	-	-	
	Oligopeptide transport, membrane spanning permease	1890	1169	1356	-	-	-	-	-	-	-	
	ABC transporter, substrate binding		1170	1357	-	-	-	-	-	-	-	
	Kelch-like protein	-	1171	1358	-	-	-	-	-	-	-	
	N-acetylmannosamine-6-phoate 2-epimerase, EC 5.1.3.9	-	1172	1359	-	-	-	-	-	-	-	
	Tn5253 bacteriocine	-	1174	1361	12910	-	-	-	-	1282	1315	
	Hypothetical	-	-	1362	12920	-	-	-	1239	-	1316	
	Tn5253, hypothetical	-	-	1363	12930	-	-	-	1238	1285	1317	
	CAAX amino protease	1346	1180	1364	12350	1248	1386	-	1237	1287	1336	
	Arsenate reductase	1348	1182	1366	12370	1250	1388	-	1235	1289	-	
	Hypothetical	1349	1183	1367	12380	1251	1389	-	1234	1290	1338	
	Methyltransferase	-	-	-	-	-	-	-	1233	1291	1339	
	Replication initiation protein A, N-terminus	-	-	-	12390	-	1390	-	1232	1292	1340	
387	ABC transporter, ATP-binding permease	1357	1191	1376	13220	1257	1396	0918	1488	1299	1346	55bpΔ
390	ABC-2 type transport system permease	1380	1213	1399	13450	1279	1418	0894	1511	1320	1369	11bpΔ
391	ABC-2 type transport system permease	1380	1213	1399	13450	1279	1418	0894	1511	1320	1369	5bpΔ
392	ABC-2 type transport system permease	1380	1213	1399	13450	1279	1418	0894	1511	1320	1369	9bpΔ
393	α-amylase	1382	1215	1401	13470	1281	1420	0892	1513	1322	1371	24bpΔ

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
394	Alanyl-tRNA synthetase	1383	1216	1402	13480	1282	1421	0891	1514	1323	1372	21bpΔ
395	Alanyl-tRNA synthetase	1383	1216	1402	13480	1282	1421	0891	1514	1323	1372	37bpΔ
396	GntR family transcriptional regulator	1393	1225	1412	13570	1292	1432	0881	1524	1334	-	190bpΔ
397	Peptidase U32	+	+	+	+	+	+	+	+	+	+	3bpΔ
398	Peptidase U32	+	+	+	+	+	+	+	+	+	+	30bpΔ
399	Peptidase U32	+	+	+	+	+	+	+	+	+	+	43bpΔ
400	ABC transporter, permease, cobalt/nickel transport ABC transporter, permease, cobalt/nickle transport	1436 1437	1265 1266	1457 1458	14010 14020	1335 1336	1477 1478	0837 0838	- -	- -	1424 1425	177bpΔ
402	Transposase	-	1268	+	-	-	-	-	-	1203	-	Δs over 123bp
403	Transposase	-	1269	+	-	-	-	-	-	1204	-	139bpΔ
404	Transposase	-	1270	1462	-	1340	-	-	-	2157	1429	298bpΔ
406	Transposase	-	1270	1462	-	1340	-	-	-	2157	1429	18bpΔ
407	IS66 family element, orf2	1442	1272	-	14080	1342	-	-	913	1206	2179	Poor assembly
408	Poly-gamma-glutamate synthesis protein	1453	1282	-	-	-	-	-	-	1381	1441	Poor assembly
409	Peptide deformylase, defB	1456	1285	1476	14200	1355	1497	0819	1572	1382	1443	23bpΔ
410	Oxidoreductase	1471	1301	1492	14360	1371	1513	0804	1587	1399	1460	19bpΔ
411	Oxidoreductase	1471	1301	1492	14360	1371	1513	0804	1587	1399	1460	44bpΔ
412	Oxidoreductase	1471	1301	1492	14360	1371	1513	0804	1587	1399	1460	Poor assembly
413	Hypothetical	1473	1303	1494	14380	1373	1515	0802	1589	1401	1462	46bpΔ
415	Transposase	+	+	+	+	+	+	+	+	+	+	1bpΔ
416	IS1239, transposase	+	+	+	+	+	+	+	+	+	+	Poor assembly
418	ABC transporter, substrate binding, oligopeptide transport	1527	1357	1550	14910	1432	1567	1466	1641	1454	1514	23bpΔ
419	UDP-N-acetylmuramoyalanyl-D-glutamate 1,6-diaminopimelate ligase (MurE)	1530	1359	1552	14930	1434	1569	1468	1643	1456	1517	43bpΔ
420	Cell shape determining protein	1548	1380	1573	15150	1456	1591	1490	1664	1476	1535	19bpΔ
421	ABC transporter, ATP-binding protein	1553	1385	1578	15200	1461	1596	1495	1669	1481	1540	Poor assembly
422	ABC transporter, ATP-binding protein	1553	1385	1578	15200	1461	1596	1495	1669	1481	1540	31bpΔ, 2bpΔ
423	ABC transporter, ATP-binding protein	1553	1385	1578	15200	1461	1596	1495	1669	1481	1540	16bpΔ, 68bpΔ
424	Oxidoreductase, pyridine nucleotide-disulfide, class I	1588	1415	1609	16030	1493	1630	1526	1700	1512	1570	13bpΔ
425	Mur ligase family	1589	1416	1610	16040	1494	1631	1527	1701	1513	1571	8bpΔ
427	Transposase	+	+	+	+	+	+	+	+	+	+	5bpΔ
428	FAD dependent oxidoreductase	1608	1433	1630	16230	1515	1649	1549	1722	1532	1590	2bpΔ
429	Bcl-2 family	1610	1435	1632	16250	1517	1651	1551	1724	1534	1592	233bpΔ
430	Hypothetical	1612	-	1633	-	-	1652	1553	-	-	1594	1,545bpΔ
434	Transketolase, cytidylate kinase D-allulose-6-phoate 3-epimerase PTS system IIC component, ribulose PTS system, IIB component	1615 1616 1617 1618	- - - -	1637 1638 1639 -	- - - -	- - - -	1657 1658 1659 1660	1555 1556 1557 -	1729 1730 1731 -	- - - -	- - - -	7,287bpΔ

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
	PTS system, IIA component	1619	-	1641	-	-	1661	1559	1733	-	-	
	PTS system, nitrogen regulatory component IIA	1620	-	1642	-	-	1662	1560	1734	-	-	
	Transcription antiterminator, BlgIG family	1621	-	1643	-	-	1663	1561	1735	-	-	
435	Cation-transporting ATPase, E1-E2 protein	1623	1436	1644	16260	1518	1664	1562	1736	1535	1598	39bpΔ
436	Cation-transporting ATPase, E1-E2 protein	1623	1436	1644	16260	1518	1664	1562	1736	1535	1598	6bpΔ, Poor assembly
437	Ribosomal protein S15	1626	1439	1647	16290	1521	1667	1565	1739	1538	1601	78bpΔ
438	Isoleucyl-tRNA synthetase	1659	1472	1677	16600	1554	1699	1598	1767	1658	1631	8Δs over 1,515bp
439	MutT/Nudix family	1669	1482	1687	16700	1564	1709	1608	1777	1578	-	591bpΔ
440	UDP-Nacetylmuramoylanalanyl-D-glutamyl-2,6-diaminopimelate--D-alanyl ligase (mutT)	1670	1483	1688	16710	1565	1710	1609	1778	1579	1642	38bpΔ
441	UDP-Nacetylmuramoylanalanyl-D-glutamyl-2,6-diaminopimelate--D-alanyl ligase (mutT)	1670	1483	1688	16710	1565	1710	1609	1778	1579	1642	184bpΔ
442	UDP-Nacetylmuramoylanalanyl-D-glutamyl-2,6-diaminopimelate--D-alanyl ligase (mutT)	1670	1483	1688	16710	1565	1710	1609	1778	1579	1642	26bpΔ
443	Glucokinase (ROK family)	1675	1488	1693	16760	1570	1715	1614	1783	1584	1647	45bpΔ
444	N-acetylneuraminate lyase	1676	-	1694	16770	1571	1716	1615	1784	1585	1648	65bpΔ
445	N-acetylneuraminate lyase	1676	-	1694	16770	1571	1716	1615	1784	1585	1648	47bpΔ
446	Conserved hypothetical	1677	1490	1695	16780	1572	1717	-	1785	1586	1649	2Δs over 166bp
447	Hypothetical	1679	1491	1696	16790	1573	1718	1617	1786	1587	1651	28bpΔ
448	NanA	-	1504	1709	16920	1586	1731	1630	1797	1600	1665	421bpΔ
449	NanA	-	1504	1709	16920	1586	1731	1630	1797	1600	1665	523bpΔ
451	Hypothetical	+	+	+	+	+	+	+	+	+	+	60bpΔ
453	Alanine racemase	1698	1508	1714	16970	1591	1737	1636	1802	-	1670	Poor assembly
454	Hypothetical	1703	-	1719	17200	1596	-	1614	1807	1609	1675	3bpΔ
456	Hypothetical	1707	-	1723	17080	1602	1745	1645	1815	1613	1679	38bpΔ
457	Nitroreductase	1710	1520	1726	17110	1605	-	1648	1818	1616	1682	74bpΔ
460	3-hydroxy-3-methylglutanyl-CoA reductase	1726	1536	-	17270	1621	1765	1664	1835	1631	1698	11bpΔ
461	IS1384, transposase orfA/orfB	+	+	+	+	+	+	+	+	+	+	Poor assembly
462	rRNAmethyltransferase, RsmB	1734	1544	1752	17370	1630	1774	1672	1843	1640	1707	21bpΔ
463	Primosome assembly protein, PriA	1736	1546	1754	17390	1632	1776	1674	1845	1642	1709	Poor assembly
464	Hypothetical	1742	1552	1759	17490	1638	1781	1680	1851	1647	1716	Poor assembly
466	Conserved hypothetical	1779	1750	1778	17910	1677	1824	-	1895	1666	1757	221bpΔ
467	ABC-2 type transporter	-	-	1779	-	-	-	-	-	-	-	2,876bpΔ
	Nod factor export ATP-binding protein I	-	-	1780	-	-	-	-	-	-	-	
	Transcriptional regulator, ArsR family	-	-	1781	-	-	-	-	-	-	-	
468	Oligopeptidase F	1780	1571						1897	1668	1758	38bpΔ
469	Ribosomal protein methyltransferase	1781	1572	1783	17940	1680	1827	1704	1898	1669	1759	Δs over 1,514bp
	Ribosomal protein L11 methyltransferase, prmA	1782	1573	1784	17950	1681	1828	1705	1899	-	1760	
	MutT/Nudix family, 7,8-dihydro-8-oxoguanine-triphosphatase	1783	1574	1785	-	-	1829	-	1900	1672	-	
470	Peptidase, M50 family	1784	1575	1786	18000	1830	1830	1711	1901	1673	1765	7bpΔ

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
472	Acetyltransferase, GNAT family	1807	1592	1814	18270	1715	1868	1730	1931	1703	1792	12bpΔ
473	Tryptophan synthetase, b subunit	1813	1597	1820	18320	1720	1873	1735	1934	1708	1797	13bpΔ
476	Hypothetical	1827	1610	1832	18450	1733	1887	1747	1947	1719	1809	3bpΔ
478	Thioesterase family, phenylacetic acid degredation	1851	1632	1851	18660	1756	1906	1768	1967	1736	1825	28bpΔ
479	Galactokinase, galactose metabolism	1853	1634	1853	18680	1758	1908	1770	1969	1738	1827	15bpΔ
480	Transcriptional regulator, TetR	1858	1639	1858	18730	1763	1913	1775	1974	1744	1832	66bpΔ
481	Hypothetical	-	-	-	-	1765	1915	1777	1976	-	-	13bpΔ
482	IS1380-Spn1 transposase	+	-	+	+	+	+	+	+	+	+	1,703bpΔ
483	Choline transporter	1860	1642	1861	18750	1766	1916	1778	1977	1746	1834	Poor assembly
485	rRNA-5s ribosomal RNA	+	+	+	+	+	+	+	+	+	+	4bpΔ
486	Aminoglycoside phosphotransferase	-	-	1930	-	-	-	-	-	-	1876	99bpΔ
487	Aminoglycoside phosphotransferase	-	-	1930	-	-	-	-	-	-	1876	382bpΔ
487	recA regulator, RecX	1902	1705	1931	19240	1835	1983	1847	2044	1811	1877	382bpΔ
488	Hypothetical	1914	1717	1943	19380	1909	1995	1864	2057	1826	1889	210bpΔ
489	LytTr DNA-binding domain, response regulator	1915	1718	1944	19390	1910	1996	1865	2058	1827	1890	12bpΔ, Poor assembly
490	Autolysin, LytA	1937	1737	1966	19600	1932	2016	1895	2087	1847	1911	14bpΔ
492	Hypothetical	1986	-	2008	20060	1978	2057	1966	2127	1887	1949	323bpΔ
	ABC transporter, ATP-binding, antimicrobial peptide transport	1987	1784	2009	20070	1979	2058	1967	2128	1888	1950	
495	Sensor histidine kinase, hk11	2001	1799	2026	20230	1996	2075	1983	2145	1903	1967	Δs over 557bp
500	IS630-Spn1, transposase orf2	+	+	+	+	+	+	+	+	+	+	Poor assembly
501	IS630-Spn1, transposase orf2	+	+	+	+	+	+	+	+	+	+	Poor assembly
503	Competence protein, cglC, comGC	2051	1861	2089	20740	2057	2139	2046	2205	1966	2018	68bpΔ
504	Competence protein, cglB, comGB	2052	1862	2090	20750	2058	2140	2047	2206	1967	2019	Poor assembly
506	LysM domain protein	2063	1874	2101	20870	2069	2153	2058	2219	1978	2030	204bpΔ
507	rRNA	+	+	+	+	+	+	+	+	+	+	2bpΔ
508	rRNA	+	+	+	+	+	+	+	+	+	+	6bpΔ
510	Transposase	-	+	+	+	+	+	+	+	+	+	2bpΔ
511	ABC transporter, ATP-binding	2075	1902	2130	21000	2097	2182	2086	2263	2014	2042	2bpΔ
512	UDP-glucose-1-phosphate uridylytransferase, *glaU*	2092	1919	2147	21170	2113	2198	2102	2279	2030	2058	Poor assembly
513	Hypothetical	2093	-	2148	-	-	-	-	-	-	-	1,456bpΔ
514	5'formyltetrahydrofolate cyclo-ligase family	2095	1921	2150	21190	2115	2201	2104	2283	2032	2060	4bpΔ
515	5'formyltetrahydrofolate cyclo-ligase family	2095	1921	2150	21190	2115	2201	2104	2283	2032	2060	7bpΔ
516	5'formyltetrahydrofolate cyclo-ligase family	2095	1921	2150	21190	2115	2201	2104	2283	2040	2060	3bpΔ
517	a-glucan phosphorylase, glgP - starch phosphorylase	2106	1932	2160	21310	2127	2211	2114	2295	2043	2070	3bpΔ
518	3-ketoacyl-(acyl-carrier-protein) reductase	+	+	+	+	+	+	+	+	+	+	453bpΔ
519	Hypothetical	2120	-	2175	21480	2143	2227	2129	2310	2057	2086	6bpΔ
521	Transporter, major facilitator family	2122	1951	2178	21510	2145	2247	2133	2313	2059	2089	53bpΔ

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
525	N-acetyl-β-D-glucosaminidase	2141	1969	2192	21740	2167	2269	2154	2334	2081	2111	17bpΔ
526	ROK family protein	2142	1970	2192	21740	2167	2269	2154	2335	2082	2112	54bpΔ
527	ROK family protein	2142	1970	2192	21740	2167	2269	2154	2335	2082	2112	4bpΔ
528	ROK family protein	2142	1970	2192	21740	2167	2269	2154	2335	2082	2112	22bpΔ
529	Sugar hydrolase, α-mannosidase	2143	1971	2194	21760	2169	2271	2156	2336	2083	2113	9bpΔ
530	Twin-arginine translocation pathway signal	2144	1972	2195	21770	2170	2272	2157	2337	2084	2114	31bpΔ
531	α-1,2-mannosidase, cell wall antigen	2145	1973	2196	21780	2171	2273	2158	2338	2085	2116	Poor assembly
532	α-1,2-mannosidase, cell wall antigen	2145	1973	2196	21780	2171	2273	2158	2338	2085	2116	Poor assembly
533	FucA, α-L-fucosidase	2146	1974	2197	21790	2172	2274	2159	2339	2086	2117	45bpΔ
534	FucA, α-L-fucosidase	2146	1974	2197	21790	2172	2274	2159	2339	2086	2117	92bpΔ
535	FucA, α-L-fucosidase	2146	1974	2197	21790	2172	2274	2159	2339	2086	2117	6bpΔ
536	Alcohol dehydrogenase, iron-containing	2157	1985	2208	21890	2183	2283	2170	2351	2096	2125	37bpΔ
537	L-fucose isomerase	2158	1986	2209	21900	2184	2284	2171	2352	2098	2126	76bpΔ
539	Permease, major facilitator superfamily	-	2007	2230	22110	2206	2306	2193	2374	2120	2146	157bpΔ
540	CbpA	2190	2017	2242	22240	2217	2316	2208	2388	2135	2158	53bpΔ
541	CbpA	2190	2017	2242	22240	2217	2316	2208	2388	2135	2158	Poor assembly
542	CbpA	2190	2017	2242	22240	2217	2316	2208	2388	2135	2158	23bpΔ
543	CbpA	2190	2017	2242	22240	2217	2316	2208	2388	2135	2158	423bpΔ
544	CbpD	2201	2028	2254	22340	2227	2328	2219	2399	2147	2168	Poor assembly
545	M16 family peptidase	2224	2052	2277	22580	2251	2352	2243	2420	2170	2190	13bpΔ

Table A.4 BLAST results of strain 1861 discrepancies

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
1	Hypothetical	+	+	+	+	+	+	+	+	+	+	Poor overlap
2	Transposase	+	+	+	+	+	+	+	+	+	+	4bpΔ
3	Transposase	+	+	+	+	+	+	+	+	+	+	3bpΔ
4	Transposase	+	+	+	+	+	+	+	+	+	+	4bpΔ, 5bpΔ
5	Transposase	+	+	+	+	+	+	+	+	+	+	32bpΔ
6	Transposase	+	+	+	+	+	+	+	+	+	+	7bpΔ
7	Transposase	+	+	+	+	+	+	+	+	+	+	3bpΔ
8	Pneumococcal surface protein, excalibur domain family	0667	0579	-	06030	0618	0726	0691	0762	0608	0623	180bpΔ
9	Transposase	+	+	+	+	+	+	+	+	+	+	44bpΔ
10	Transposase	+	+	+	+	+	+	+	+	+	+	48bpΔ
11	Hypothetical	+	+	+	+	+	+	+	+	+	+	3bpΔ
12	Degenerate transposase	+	+	+	+	+	+	+	+	+	+	5bpΔ
13	Degenerate transposase	+	+	+	+	+	+	+	+	+	+	18bpΔ
14	Degenerate transposase	+	+	+	+	+	+	+	+	+	+	5bpΔ
16	Tn*5253* , conserved hypothetical	-	1185	1143	13170	-	-	-	1231	1293	1341	5bpΔ
16	Replication initiation protein, repA	1350	-	1144	13160	-	1070	-	1232	1292	1340	5bpΔ
17	Replication initiation protein, repA	1350	-	1144	13160	-	1070	-	1232	1292	1340	19bpΔ
18	Replication initiation protein, repA	1350	-	1144	13160	-	1070	-	1232	1292	1340	4bpΔ, 1bpΔ
19	Replication initiation protein, repA	1350	-	1144	13160	-	1070	-	1232	1292	1340	9bpΔ
20	Replication initiation protein, repA	1336	-	1144	13160	-	1070	-	1233	1291	1339	33bpΔ, 3bpΔ
21	Replication initiation protein, repA	1336	-	1144	13160	-	1070	-	1233	1291	1339	20bpΔ
22	Replication initiation protein, repA	1336	-	1144	13160	-	1070	-	1233	1291	1339	84bpΔ, poor overlap
22	Tn*5253* , conserved hypothetical	+	1183	+	+	1251	+	-	1234	+	1338	84bpΔ
23	Tn*5253* , conserved hypothetical	+	1183	+	+	1251	+	-	1234	+	1338	18bpΔ
24	Tn*5253* , conserved hypothetical	+	1183	+	+	1251	+	-	1234	+	1338	2bpΔ
25	Tn*5253* , conserved hypothetical	+	1183	+	+	1251	+	-	1234	+	1338	1bpΔ
26	Tn*5253* CAAX amino terminal protease family	+	1180	+	+	1248	+	-	1482	+	+	1bpΔ
27	Tn*5253* CAAX amino terminal protease family	+	1180	+	+	1248	+	-	1482	+	+	115bpΔ
28	Tn*5253* CAAX amino terminal protease family	+	1180	+	+	1248	+	-	1482	+	+	3bpΔ
29	Tn*916* , Hypothetical	-	-	-	13061	-	-	-	1407	-	-	24bpΔ
30	Tn*916* , TetM protein	-	-	1155	13050	-	-	1919	1049	1231	0171	2bpΔ

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
31	Tn*916* , TetM protein	-	-	1155	13050	-	-	1919	1049	1231	0171	7bpΔ
32	Tn*916* , TetM protein	-	-	1155	13050	-	-	1919	1049	1231	0171	8bpΔ
33	Tn*916* , TetM protein	-	-	1155	13050	-	-	1919	1049	1231	0171	26bpΔ
34	Tn*916* , NLP/P60 family	-	-	1158	13030	-	-	1916	1412	1233	0168	18bpΔ
35	Tn*916* , NLP/P60 family	-	-	1158	13030	-	-	1916	1412	1233	0168	19bpΔ
36	Tn*5251* membrane protein	-	-	1159	13020	-	-	1915	1413	1234	0167	70bpΔ
37	Tn*5251* membrane protein	-	-	1159	13020	-	-	1915	1413	1234	0167	60bpΔ
38	Tn*5251* membrane protein	-	-	1159	13020	-	-	1915	1413	1234	0167	29bpΔ, 1bpΔ
39	Tn*5251* membrane protein	-	-	1159	13020	-	-	1915	1413	1234	0167	24bpΔ, 25bpΔ
40	Tn*5251* membrane protein	-	-	1159	13020	-	-	1915	1413	1234	0167	3bpΔ, 15bpΔ
41	Tn*5251* membrane protein	-	-	1159	13020	-	-	1915	1413	1234	0167	16bpΔ
42	Conjugative transposon protein	-	-	1160	13010	-	-	1914	1414	1235	0166	9bpΔ
44	Putative bacteriocine, Tn*5253*	-	1174	1169	12910	-	-	-	1240	1282	1315	41bpΔ
45	Tn*5253* , conserved hypothetical	-	-	1175	12850	-	-	-	1246	-	1309	1bpΔ, 10bpΔ
46	M23 peptidase domain	-	-	1176	12840	-	-	-	1247	1275	1308	48bpΔ
47	M23 peptidase domain	-	-	1176	12840	-	-	-	1247	1275	1308	15bpΔ
48	M23 peptidase domain	-	-	1176	12840	-	-	-	1247	1275	1308	25bpΔ
49	Hypothetical	-	-	1177	12830	-	-	-	1248	1274	-	625bpΔ
50	Tn*5253* , SNF2-related helicase	-	-	1179	-	-	-	-	-	1272	1305	4bpΔ
51	Tn*5253* , SNF2-related helicase	-	-	1179	-	-	-	-	-	1272	1305	41bpΔ
52	Tn*5253* , SNF2-related helicase	-	-	1179	-	-	-	-	-	1272	1305	22bpΔ
53	Tn*5253* , SNF2-related helicase	-	-	1179	-	-	-	-	-	1272	1305	20bpΔ
60	ATPase with chaperone activity, ATP-binding subunit	-	-	1187	12640	-	-	-	1261	1263	1297	38bpΔ
61	Helix turn helix motif	1050	0930	1188	+	0987	1128	-	-	-	+	28bpΔ
62	Signal recognition particle GTPase	1051	0931	1189	+	0988	1129	-	-	-	+	28bpΔ
63	Signal recognition particle GTPase	1051	0931	1189	+	0988	1129		-	-	+	70bpΔ
64	Signal recognition particle GTPase	1051	0931	1189	+	0988	1129		-	-	+	28bpΔ
65	Phage transcriptional repressor	-	-	1190	-	-	-	-	-	-	-	2,078bpΔ
	UmuC MucB	-	-	1191	-	-	-	-	-	-	-	
66	Tn*5252* , relaxase	-	-	1195	12440	-	-	-	-	1251	1277	27bpΔ
68	Pneumococcal histidine triad protein A/B, PhtA/B	1174	1037	1217	10770	1093	1226	1049	1104	1073	1122	13bpΔ
69	Signal recognition particle-docking protein, FTsY	1244	1101	1281	11390	1157	1306	0985	1359	1134	1059	18bpΔ
70	Chromosome segregation protein, Smc	1247	1104	1284	11420	1160	1309	0982	1362	1137	1056	20bpΔ

#	Gene name/description	ORF ID from annotated *S. pneumoniae* strains										Discrepancy
		TIGR4 (SP)	D39 (SPD)	P1031 (SPP)	ATCC700669 (SPN23F)	JJA (SPJ)	70585 (SP70585)	Taiwan 19F (SPT)	Hungary 19A (SPH)	G54 (SPG)	CGSP14 (SPCG)	
71	Hypothetical	1250	1108	1288	11450	1164	1313	0978	1366	1141	1053	21bpΔ
71	Endonuclease	1251	1109	1289	11460	1165	1314	0977	1367	1142	1052	21bpΔ
72	3-isopropylmalate dehydrogatase small subunit	1255	1113	1293	11500	1169	1318	0973	-	1147	1048	17bpΔ
73	CTP phosphocholine cytidyltransferase, LicC	1267	1123	1305	11610	1181	1330	0961	1383	1161	1231	7bpΔ
74	CAAX amino protease family	1346	1180	1364	12350	1248	1386	-	1237	1287	1336	78bpΔ
75	CAAX amino protease family	1346	1180	1364	12350	1248	1386	-	1237	1287	1336	4bpΔ
76	CAAX amino protease family	1346	1180	1364	12350	1248	1386	-	1237	1287	1336	4bpΔ, 72bpΔ
77	Tn*5253* conserved hypothetical	1333	1181	1365	13120	-	-	-	1475	1288	1337	5bpΔ, 10bpΔ
78	Tn*5253* conserved hypothetical	1348	1182	1366	13130	1250	1388	-	1235	1289	-	126bpΔ
79	Hypothetical	1349	1183	1367	13140	1251	1389	-	1234	1290	1338	24bpΔ
80	Hypothetical	1349	1183	1367	13140	1251	1389	-	1234	1290	1338	8bpΔ
81	Peptidase, U32 family	+	+	+	+	+	+	+	+	+	+	2bpΔ
82	Hypothetical	+	+	+	+	+	+	+	+	+	+	Poor overlap
83	IS*1380* -Spn1, Transposase	+	-	+	+	+	+	+	+	+	+	Poor overlap
84	NanA	-	1504	1709	16920	1586	1731	1630	1797	1600	1665	415bpΔ
85	NanA	-	1504	1709	16920	1586	1731	1630	1797	1600	1665	525bpΔ
86	Hypothetical	-	+	+	+	+	+	+	+	-	+	89bpΔ, poor overlap
87	Hypothetical	-	+	+	+	+	+	+	+	-	+	Poor overlap
88	Hypothetical	2093	-	2148	-	-	-	-	-	-	-	440bpΔ

A.2 Conference presentations

Harvey, R.M., Stroeher, U.H., Ogunniyi, A.D., Leach, A.J. and Paton, J.C. (14th – 17th April, 2007). "*Identification of potential virulence enhancing genes in Streptococcus pneumoniae type 1 isolates*". 8th European Meeting on the Molecular Biology of the Pneumococcus. Oeiras, Portugal.

Harvey, R.M., Stroeher, U.H., Ogunniyi, A.D., Leach, A.J. and Paton, J.C. (9th – 13th July, 2007). "*Identification of potential virulence enhancing genes in Streptococcus pneumoniae type 1 isolates*" Australian Society for Microbiology, Proffered Paper presentation. Adelaide, South Australia.

Harvey, R.M., Stroeher, U.H., Ogunniyi, A.D., Leach, A.J. and Paton, J.C. (6th – 10th July, 2008). "*Identification of potential virulence enhancing genes in Streptococcus pneumoniae type 1 isolates*" Australian Society for Microbiology, S.A. recipient of BD award presentation. Melbourne, Victoria.

Harvey, R.M., Stroeher, U.H., Ogunniyi, A.D., Leach, A.J. and Paton, J.C. (15th May, 2009). "*Genetic characterisation of Streptococcus pneumoniae type 1 isolates in relation to invasiveness"*. Next Generation Sequencing Workshop. Geneworks, Adelaide, South Australia.

Harvey, R.M., Stroeher, U.H., Ogunniyi, A.D., Leach, A.J. and Paton, J.C. (20th – 23rd Spetember, 2009). "*Potential virulence enhancing genes identified in Streptococcus pneumoniae serotype 1 isolates"*. BacPath 10. Barossa Valley, South Australia.

A.3 Poster presentations

Harvey, R.M., Stroeher, U.H., Ogunniyi, A.D., Leach, A.J. and Paton, J.C. (30th March – 2nd April, 2009). "*Identification of potential virulence enhancing genes in Streptococcus pneumoniae type 1 isolates*". Society for General Microbiology, Burnett-Hayes postgraduate award winner presentation. Harrogate, U.K.