UNIVERSITY OF SOUTHAMPTON

Faculty of Medicine

Clinical and Experimental Sciences

Volume 1 of 1

Microevolution of *Neisseria lactamica* during prolonged colonisation of the nasopharynx

by

Anish Kumar Pandey

Thesis for the degree of

PhD

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MEDICINE

Microevolution of *Neisseria lactamica* during prolonged colonisation of the nasopharynx

Authors: Anish Pandey[1],
Supervisors: David W Cleary[1], Jay R Laver[1], Andrew Gorringe[2] & Robert C Read (Primary) [1]

1. Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton
2. Public Health England, Porton Down.

Abstract

Carriage of *Neisseria lactamica* occurs naturally at high frequency in infants and low frequency in young adults. There is an inverse epidemiological relationship between *N. lactamica* carriage and disease caused by *Neisseria meningitidis* (meningococcus). Serogroup B meningococci remain the dominant cause of invasive meningococcal disease in the developed world and have frustrated the production of polysaccharide-conjugate vaccines. While two, recombinant, OMV based vaccines (Richmond *et al.*, 2012; Vernikos and Medini, 2014) have been created and elicit immunological responses, they are less effective on infants (one of the groups most at risk of IMD) and have limited effect on meningococcal carriage and subsequently, on herd immunity. A human experimental challenge study in which healthy, young adult volunteers were inoculated with *N. lactamica* Y92-1009 showed that carriage of *N. lactamica* both displaced and inhibited reacquisition of wild type *N. meningitidis*, and although rare, co-colonization of the two species was also observed in a small number of cases (Deasy *et al.*, 2015). This study provided the opportunity to investigate whether there is a genomic basis for *N. lactamica*'s effect on meningococcal carriage as the mechanism for this interaction remains unknown. Secondly, the use of whole genome sequencing, paired with mutation analysis via the breseq pipeline (Barrick *et al.*, 2014) will comment on the mutability of *N. lactamica,* a potential bacterial medicine, during 6 months of *in vivo,* human challenge. Thirdly, this allows us to track the within-host microevolution of an identically administered commensal *Neisseria spp.* over the course of 6 months of carriage (chapter 5)

Isolates obtained from individuals who were co-colonised by *N. meningitidis* and *N. lactamica* for a prolonged period were examined for evidence of the effect of recombination (r/m) as well as loci affected by it (chapter 6). In addition to the majority of volunteers who solo carried *N.*

lactamica Y92-1009. Recombination was determined for; volunteers in which inoculated *N. lactamica* was the sole *Neisseria spp.* detected, seven, artificially inoculated, *N. lactamica*/meningococcal co-carriers and two extra volunteers who were naturally co-colonised. Using ClonalFrameML (Didelot and Wilson, 2015), we detected minimal homologous recombination events among *N. lactamica* Y92-1009 and no examples of interspecific allele transferred with co-colonising meningococci. In contrast, we found evidence of a dynamic, interspecific relationship and a number of recombination events occurring among co-colonised volunteers with naturally acquired *Neisseria*.

A separate, short term clinical trial utilizing multiple colony sampling (chapter 4) examined the difference in mutational profiles of longitudinal samples *N. lactamica* strain Y92-1009 sourced from *in vitro* conditions versus *in vivo* conditions over one month. Larger numbers of SNPs, nonsense and recurring mutations were observed among the *in vitro* cohort and the quantity/diversity of phase variable mutations was more pronounced among the *in vivo* cohort. Chapters 4 and 5 are supported by a highly-accurate reference genome. The sequencing, assembly, annotation and characterisation of the first complete *N. lactamica* Y92-1009 genome is described in chapter 3 (Pandey *et al.*, 2017). This chapter also revealed the presence of a large but uncharacterised prophage sequence in the strain. The very first example of a species encompassing, pan genomic analysis of *N. lactamica* (chapter 7**)** revealed that *N. lactamica* Y92-1009 possess fewer unique genes/alleles than other members of the species with no virulence factors detected among the results.

In conclusion, the *N. lactamica* Y92-1009 genome is a self-curated system with plastic elements that (like other *Neisseria spp.*) could facilitate rapid changes in expression via its phase variable elements. However, it appears to have remained genetically stable during the 6-month course of carriage in human volunteers. Demonstrating little recombination, no interspecific gene transfer with co-colonising meningococci and an average mutation rate for a *Neisseria* species. While efforts need to be made to improve the acquisition and retention of carriage, *N. lactamica* appears to be a safe, naturally competent, potential bacterial therapeutic, capable of a broad-spectrum reduction of meningococcal carriage.

(637 words)

Computational Microbiology

Thesis for the degree of Doctor of Philosophy_

Microevolution of *Neisseria lactamica* during prolonged colonisation of the nasopharynx

MICROEVOLUTION OF NEISSERIA LACTAMICA DURING PROLONG COLONISATION OF THE NASOPHARYNX

Anish Kumar Pandey

Table of Contents

List of Tables

List of Figures

DECLARATION OF AUTHORSHIP

I, ... [please print name]

declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

[title of thesis] .....

.....

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. [Delete as appropriate] None of this work has been published before submission [or] Parts of this work have been published as: [please list references below]:

Signed: .....

Date: .....

Acknowledgements

I wish to acknowledge a very special group of people who have aided the development of this project from inception to almost written creation. Rest assured, I, the author am permanently indebted to this special group of men and women who have offered me the very best in terms of academic and emotional support.

I wish to thank my teachers: Dr Jay Laver, for his friendship, advice and outstanding supervision that has gone above and beyond the call of anything I could expect of him. Thank you, Dr David Cleary, without you I would not have gained the fundamentals of computational biology or the enthusiasm to continually grow my interest in this field. I also thank the estimable Prof. Read for continuing to entrust me with his positivity and support during one of the most challenging periods of my life. Thank you, Sir.

To the members of my research group (ZP, AV, AD & RuH) and adoptive office, thank you for filling my days with laughter, interesting conversation, helpful advice and engaging science.

Most importantly, I wish to thank Ma and Papu (VP & RKP), my little sisters (AP, TofP), and oldest friend (PGR) for their continual and unconditional support in sharing the trials, tribulations and successes of thesis writing.

Thank you to the innumerable number of software developers and scientists that I have asked from and received such great advice and kindness. I promise to pay it forward whenever I get the chance.

Lastly, to any PhD students reading this in the future (my sympathies!), I offer the following advice: try your very best, don't forget to self-reflect and be proud of yourself for the road taken to get where you are and lastly, live presently. These years are fleeting and transient but there's a lot to appreciate and enjoy if you look around.

Anish

Definitions and Abbreviations

- **CDS:** Coding sequence. DNA sequence coding for proteins comprising a gene
- **Contig:** A contiguous stretch of sequence. When combined with assembly gaps, this is known as a **scaffold.**
- **GCR**: Guanine Cytosine (Bases) Ratio. Can be gene or genome specific
- **Gold standard genome**: A genome widely held to be the most accurate and representative of a given sample.
- **Homolog**: A similar or the same gene due to descent from a common ancestor
- **IMD**: Invasive meningococcal disease
- ***N.lac****: Neisseria lactamica*,
- ***N.men*** *Neisseria meningitidis*
- **NGS**: Next generation sequencing
- **Maximum likelihood** (Phylogenetics): is a statistical method for estimating the unknown parameters of a given probability model
- **MLST:** Multi Locus Sequence Typing
- **PacBio**: Pacific Biosciences
- **Paralogue**: A duplicated copy of a gene
- **SMRT**: Single molecule real time sequencing, offered by companies such as Pacific Biosciences
- **Sequence Type (ST):** Classification scheme based on seven housekeeping loci used to categorise bacteria a step beyond species level.
- **SSR:** simple sequence repeats may be homo or Multimeric. Examples of repeat tracts in this study include the *feta* variable tract include 10 G ("5' GGGGGGGGGG 3'" and *modA* variable tract(CGAT)^{14} CGATCGAT…etc
- **WGS:** Whole genome sequencing
- **ZMW**: Zero mode wave guide, technique used during SMRT sequencing
- **5'** :Genetic direction 5 prime,
- **3'**: genetic direction: 3 prime

Chapter 1: Introduction

1.1 Neisseria meningitidis, Neisseria lactamica and the Nasopharynx

Neisseria meningitidis (also known as the meningococcus) is a diplococcoid, nasopharyngeal-dwelling organism that specifically colonises humans. In rare circumstances, this bacterium can progress from natural carriage to disease state, crossing the blood-brain barrier and causing invasive meningococcal disease (IMD).The disease has a case-fatality rate of 10% (Caugant and Maiden, 2009) and 20% of survivors suffer permanent, severe injures such as hearing loss, neurological damage and limb amputation (NHS, 2018). The meningococcus was first isolated in 1887 and its disease manifestation is responsible for epidemics/pandemics with symptoms including meningitis and septicaemia (alongside less common symptoms like conjunctivitis, septic arthritis, pericarditis and pneumonia) affecting hundreds of thousands of individuals across the globe, annually (Rosenstein *et al.*, 2001). Notable outer membrane, surface-expressed, virulence factors of this organism include the pili (pilin subunits; *pil* loci), polysaccharide capsule loci *cps*, adhesins *opa* and *opc,* the porins *porA* and *porB*, iron sequestration genes, and lipooligosaccharise (LOS) endotoxins (Stephens, 2009).

Four conditions have to be met before *Neisseria meningitidis* can cause invasive disease. Firstly, a human has to encounter an invasive strain of the bacterium. Secondly, the bacterium must successfully colonise the naso/oropharyngeal mucosa. Thirdly, the bacterium must bypass the mucosal epithelium before fourthly, the bacterium survives and proliferates in the bloodstream (Van Deuren, Brandtzaeg and Van Der Meer, 2000). Once progression past these stages is reached IMD can be rapidly fatal and *N. meningitidis* remains either the first or second most common cause of meningitis worldwide(Crum-Cianflone and Sullivan, 2016). The major pre-requisite of invasive disease by nasopharyngeal pathogens is carriage. Meningococcal carriage varies widely between countries and age groups; with the age demographics most at risk of IMD being children aged 0-5 and young adults aged 15-24 (Read, 2014). Carriage rates for this demographic in Europe and Africa vary from 10-35% but recent studies in South America, where rates of meningococcal disease are lower, shows carriage rate dropping to 1.6-6.9%. A number of social and behavioural factors have been correlated with increased meningococcal carriage. These include socioeconomic status (Cleary *et al.*, 2016), smoking, salivary exchange and mass social attendance in pubs & clubs (MacLennan *et al.*, 2006). A Columbian study suggests cultural practices, namely students attending university from home in Columbia, compared to dormitories in the UK, as the most likely reason for lower observed meningococcal carriage rate in this country's demographic compared to Europe (Araya *et al.*, 2015).

1.1.1 Typing

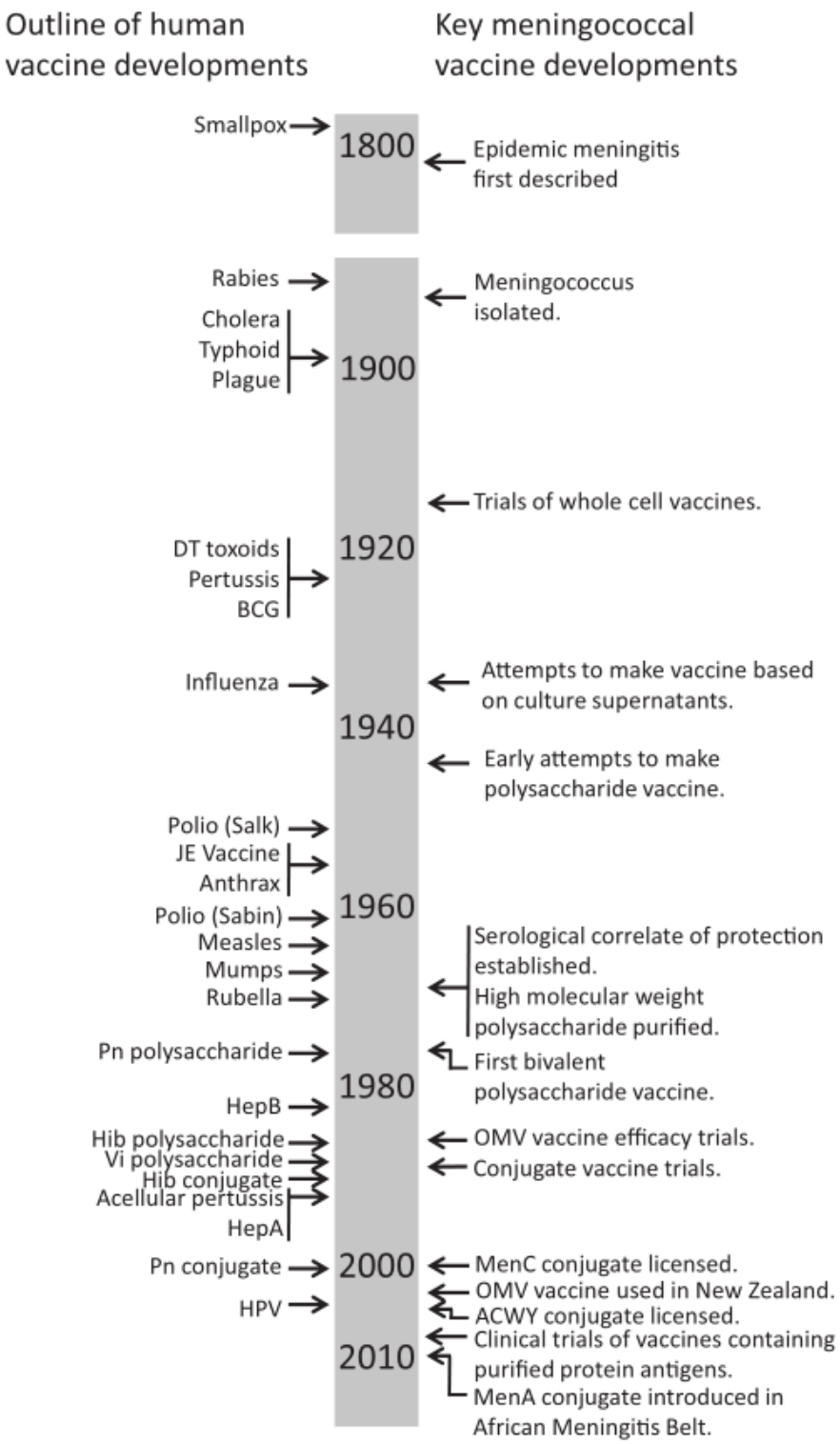
While *N. meningitidis* may be either encapsulated or unencapsulated (designated capsule null, cnl); the capsule enables both the evasion of host immunity and blood survival. The vast majority of IMD burden is caused by six of twelve capsule types/serogroups A, B, C, W135, X and Y (Johswich *et al.*, 2012). IMD is endemic to certain regions with serogroup A IMD implicated in Africa and Asia while serogroups B and C are more prevalent in Europe and the USA/Canada (Chang, Tzeng and Stephens, 2012). *Neisseria meningitidis* is a panmitic organism and the phenotyping approach (which typed via polysaccharide capsule, outer membrane proteins and LOS sub types) has been superseded by genotyping methods (Read, 2014). Molecular typing of *Neisseria* spp. was first performed by multi locus enzyme electrophoresis (MLEE) and was used to describe a hyperinvasive lineage of *Neisseria meningitidis* ET-5 (Caugant *et al.*, 1986). This typing technique was in turn supplanted by multi locus sequence typing (MLST) (Maiden *et al.*, 1998). While both techniques characterize meningococcal variation based on a small subset of housekeeping genes, MLST was able to capitalize based on advances in DNA sequencing and easier data sharing. The ET-5 strain is now known as ST-32, a serogroup B meningococcus, the clones of which have been responsible for epidemics of IMD spanning 40 years and four continents (Harrison *et al.*, 2015) The "gene by gene" approach to MLST was expanded to 53 ribosomal protein subunits (rps loci) and allowed bacterial typing to distinguish at the strain level (Bennett *et al.*, 2012). Currently, MLST, rMLST and an even further typing depth via core and accessory genomic content of *Neisseria* spp. can be performed on the Neisseria PUBMLST database (http://pubmlst.org/neisseria/)(Bratcher *et al.*, 2014).

1.1.2 Meningococcal Vaccines: A brief history

While *Neisseria meningitidis* infections are treatable with a variety of antibiotics (Stephens, 2007), due to the status of IMD as a global source of morbidity and mortality, there have been a number of preventative advances made with regards to its containment. These are presented alongside general advances in human vaccine development in **Figure 1-1**. A polysaccharide vaccine is an inactivated (as opposed to a live-attenuated vaccine) immunity-stimulating solution comprised of fragments of long, sugar chains sourced from bacterial cell-walls. The first meningococcal polysaccharide vaccine targeted serogroup C strains and was successfully trialled by 1970 (Artenstein *et al.*, 1970). Soon after a polysaccharide vaccine was also introduced for serogroup A (Wahdan *et al.*, 1973). These monovalent vaccines have been modified to cross-protect against multiple serogroups of meningococcal disease. Nowadays bivalent (serogroups A and C), trivalent (serogroups A, C, W135) and quadrivalent (serogroups A, C, W135 and Y) polysaccharide vaccines are available (Vipond, Care and Feavers, 2012). Problems with polysaccharide-based vaccines

include that they are T cell-independent and rely on a typically short-lived, humoral immune response. This means they need to be re-administered which has been found to result in a lower, secondary immune response among both children (Gold *et al.*, 1975) and adults (Granoff *et al.*, 1998). This led to the development of conjugate vaccines, where the polysaccharide was combined (conjugated) with a protein molecule (normally a bacterial toxin) to trigger an immune response from T-cells. The advantages of this vaccine type over polysaccharide are: a) improved antibody responses, b) stimulating memory B-cell response which provided longer-lasting protection, c) generation of a mucosal immune response via IgA/IgG, d) generation of a "booster" effect where re-exposure to vaccine antigen can increase immunity to previous, protective levels and e) reduction of bacterial carriage generating herd immunity (Durando, Faust and Torres, 2015).The first conjugate vaccine was developed for *Haemophilus influenzae* type B (Peltola *et al.*, 1992) which stimulated efforts to produce a conjugate vaccine for *N. meningitidis*. In 1999, the UK was the first country to introduce a conjugate, serogroup C, monovalent vaccine to the national immunisation program (Miller, Salisbury and Ramsay, 2001), this led to a decrease in national carriage of serogroup C meningococci (Maiden and Stuart, 2002), the establishment of herd immunity (Borrow *et al.*, 2013) and subsequently, fewer instances of serogroup C IMD. As it currently stands, the UK has a plethora of conjugate and polysaccharide vaccines options available (Table 1, (Crum-Cianflone and Sullivan, 2016)) as part of the national immunisation program or for individuals travelling to areas with an increased risk of IMD.

Figure 1-1 History of meningococcal vaccines alongside key developments in human vaccines



A timeline of meningococcal vaccine development alongside key, global advances in the development of human vaccines. Image taken from (Vipond, Care and Feavers, 2012)

1.1.3 The search for a Serogroup B vaccine

Despite preventative advances in the management of IMD from serogroups A, C, W135 and Y, serogroup B meningococci are currently responsible for 85-90% of IMD in the UK. However, this is still rare, with 600-1400 cases per year (*The Lancet Infectious Diseases*, 2014).

To date, no polysaccharide or glycoconjugate vaccine has been able to be developed for serogroup B meningococci. This is because the serogroup B polysaccharide, an α2-8-linked polysialic acid was found to be too similar to α2-8-linked sialyated human glycoproteins such as foetal, neural cell adhesion molecules (Finne, Leinonen and Mäkelä, 1983). Similarities also exist between lacto-*N*-neotetraose (present on meningococcal LOS) and human glycolipid, paragloboside (Tsai and Civin, 1991). These meningococcal surface carbohydrates are thought to mimic human molecules to guard the bacterium from immune attack (Lo, Tang and Exley, 2009) and raise the issue of malignant autoimmunity in creating a polysaccharide vaccine. Furthermore, meningococci stochastically vary the expression of many antigenic genes through the mechanism of phase variation (Saunders *et al.*, 2000), which in one example, allowed the organism to evade the bactericidal action of a monoclonal antibody (Bayliss *et al.*, 2008). Despite autoimmune and phase-variation concerns, the poor immunogenicity of a candidate, *N*-propionylated, polysaccharide vaccine conjugated with tetanus toxoid stopped further attempts to develop conjugate serogroup B meningococcal vaccines (Bruge *et al.*, 2004).

As opposed to polysaccharide and polysaccharide conjugate vaccines, protein-based vaccines consist of purified outer membrane vesicles (OMVs) shed by serogroup B meningococci, in an effort to induce protective immunity against the pathogen. OMV based vaccines have been used successfully to combat outbreaks of serogroup B IMD in Norway (Bjune *et al.*, 1991) and New Zealand (Wong *et al.*, 2007). In both of these instances the vaccine was found to induce protection against the specific strain the OMVs were collected from and directed immunity against the meningococcal porin protein, PorA. However PorA is a poor, universal therapeutic target due to antigenic variation via the mechanism of phase variation (Tauseef, Ali and Bayliss, 2013). And in the UK, PorA type was seen to vary among a decade of invasive, meningococcal lineages (Gray *et al.*, 2006). Therefore, these OMV vaccines would offer little cross-protection outside of localised epidemics. A double-blind, Chilean study testing 2 OMV vaccines found they elicited an immune response (4-fold rise in SBA titre) in 33% children (2-4 years) and 60% adults (17-30 years) but none in infants (<1 year old) against heterologous meningococcal strains (Tappero *et al.*, 1999). Attempts to broaden the scope of OMV vaccines by incorporating more PorA subtypes into a hexavalent PorA vesicle vaccine found that protection was limited to strains

harbouring the same 6 PorA subtypes (De Kleijn *et al.*, 2001)
(OMV vaccines are reviewed in (Granoff, 2010; Shea, 2013))

1.1.4 Recombinant protein OMV Vaccines

Reverse vaccinology is a technique used to bioprospect (data mine) genomic data with a view to generating novel gene targets for vaccine development (Seib, Zhao and Rappuoli, 2012). This approach was first utilised to identify antigens in a serogroup B meningococcal genome and predicted 350 putative surface exposed-proteins to be used as a list of potential vaccine candidates (Pizza *et al.*, 2000). After eliminating candidate antigens based on: a) unsuccessful cloning and expression in *Escherichia coli*, b) unsuccessful purification for use in murine infection studies, c) unsuccessful elicitation of a serum bactericidal response (SBA: a standardised test of antibody production by participant sera against bacteria in response to vaccine inoculation) d) low antigenic expression by meningococcal strains e) absence of candidate genes in hypervirulent strains and f) antigenic variability due to phase variation (Granoff, 2010). A recombinant, protein-based, serogroup B vaccine called Bexsero (manufacturer: GSK, alias: 4CMenB) was created (Gorringe and Pajon, 2012; Vernikos and Medini, 2014). This vaccine consists of four components: factor H binding protein (*fHbp*), adhesin *(nadA*), heparin-binding protein (*nhba*) and porin (*porA*) which was derived from outer membrane vesicles (OMVs) from a New Zealand epidemic strain (Wong *et al.*, 2007). Evidence suggested that Bexsero was consistently immunogenic over multiple doses as SBA titres were maintained among inoculated adolescents after 2 years (Santolaya *et al.*, 2013). In 2015, Bexsero was added to the UK national immunization program in 2015 and, nine months later, was shown to reduce the number of serogroup B IMD cases by 42% (Parikh *et al.*, 2016). Limitations of 4CMenB are as follows. Bexsero performs poorly on the age demographic arguably most in need of IMD protection, infants. Infant bactericidal antibody response was seen to decline rapidly following 3 dose course of 4CMenB, with boosters suggested to be required at 1 year of age (Snape *et al.*, 2016). Infants are less likely to elicit the cross-protective responses demonstrated in adolescents and adults. Poor cross-reactive antibody response in response to the *fHbp* antigen was also observed (Findlow *et al.*, 2010; Brunelli *et al.*, 2011), with the latter study commenting that even after multiple doses (3 doses plus 1 booster dose) bactericidal antibody responses became increasingly weak to subvariants of *fHbp*, which was consistently expressed by the examined panel of 10 meningococci. In addition to this, the effect of 4CMenB on meningococcal carriage suppression was judged to be modest at best (Read *et al.*, 2014) with a more recent study finding no correlation between vaccine-conferred serum bactericidal antibody titre and suppression of meningococcal carriage (Read *et al.*, 2017). Cost-effectiveness is also a drawback of utilising 4CMenB. While Serogroup B IMD remains dominant in the developed world,

the rate of IMD remains low in these regions, annually affecting ~1-2 people per 100,000. The current cost of 4CMenB dosage is £75 per dose, a 2014 study found this would need to lowered significantly (<£4 per dose) to be cost-effective (Christensen *et al.*, 2014). A study among Dutch infants reported a similar cost-benefit analysis, pricing dosage at €10 to be cost effective (Pouwels *et al.*, 2013).

In 2014, a bivalent, recombinant vaccine called Trumenba (manufacturer: Pfizer, aliases: rLP2086/ MenB-FHbp) was approved by the US food and drug administration but has yet to be approved in Europe. Trumenba contains 2 variants, one from each of the two broad sub-families of *fHbp* (Richmond *et al.*, 2012), an antigenically diverse virulence factor commonly expressed by meningococci that aids them in avoiding compliment-mediated immune response (Murphy *et al.*, 2009). Trumenba has consistently elicited strong immune responses in adolescents and young adults against a variety of meningococcal strains responsible for invasive IMD in Europe and the USA (reviewed, (Shirley and Taha, 2018)). Limitations of Trumenba include post-inoculation, adverse reactions. The vaccine is not being approved for use in individuals under the age of 10 due to causing fevers in the majority of participants in that demographic (Martinon-Torres *et al.*, 2014) and a "real world study" also reported symptoms including injection site pain, myalgia, fatigue and fever among adolescents (Fiorito *et al.*, 2017). In addition, Trumenba was seen to have no effect on serogroup B meningococcal carriage in either supressing it or preventing its acquisition (McNamara *et al.*, 2017; Soeters *et al.*, 2017).

1.1.5 N. lactamica

Neisseria lactamica is a gram negative, diplococcoid, commensal organism that colonises the human nasopharynx. While *Neisseria lactamica* shares a large degree of genomic homology with *N. meningitidis* (Bennett *et al.*, 2010), it does not express Bexsero antigens; PorA (the highly variable surface protein target that limited the effectiveness of pre 2010 meningococcal OMV vaccines), NadA or FHbp (Lucidarme *et al.*, 2013). Part of a group of more commensally associated *Neisseria spp.* such as *Neisseria mucosa, Neisseria sicca and Neisseria cinerea,* carriage of *N. lactamica* has infrequently led to infection but is reported (Wilson and Overman, 1976). Examples of more common neisserial pathogens include the causative agents of gonorrhoea and IMD; *Neisseria gonorrhoeae* and *N.meningitidis* respectively (Criss and Seifert, 2012). Named due to its innate ability to ferment lactose and produce β-D-galactosidase, *N. lactamica* was previously able to be biochemically and phenotypically differentiated from the rest of the *Neisseria.* But this trait thought to be unique to this *Neisseria* species has since also been identified in *Neisseria oralis* (Bennett, Jolley and Maiden, 2013). *N. lactamica* is still able to be genetically differentiated from the rest of the *Neisseria* via MLST and *N. lactamica*-specific sequence types and clonal complexes have been designated.

A study that examined asymptomatic carriage of *N.meningitidis* and *N. lactamica* in 2969 healthy infants and children discovered a peculiar pattern of epidemiology (Gold *et al.*, 1978). Carriage rates of *N. lactamica* peaked in 18-month-old infants and declined to a much lower rate in teenage children. Conversely, a low level of meningococcal carriage was detected in infants during the first four years of life. This was not however the case in teenagers aged between 14 and 17, as *N.meningtidis* carriage rates increased. These findings have since been confirmed by further studies, most notably, the Stonehouse survey (Cartwright *et al.*, 1987) that found *N. lactamica* carriage was six times higher in children between the ages of 0-5, with *N.meningitidis* carriage observed to be lower in this demographic relative to the rest. Similar results were also observed in a more recent study (Bakir *et al.*, 2001) and from carriage studies in the Faroe Islands, Nigeria, Spain, Turkey, Greece and New Zealand (reviewed in (Gorringe, 2005)). These studies postulate that *N. lactamica* may have a role in protecting the surveyed school children from meningococcal colonisation in the earlier years of childhood. This natural carriage dynamic has been subsequently examined to develop countermeasures (i.e. vaccines and therapeutics) against *N. meningitidis* carriage, a pre-requisite for invasive meningococcal disease.

1.1.6 *N. lactamica* Y92-1009: an answer to meningococcal serogroup B disease burden?

The strain *N. lactamica* Y92-1009 has been used as a model organism of the species for a variety of multi-disciplinary work over the past 15 years. It was originally isolated from a 1992 carriage study of school pupils in Londonderry, Northern Ireland (Oliver *et al.*, 2002), where OMVs cultured from the *N. lactamica* Y92-1009 demonstrated high levels of cross reactivity against a range of *N.meningitidis* antigens. Advances of knowledge in recent years using this strain include the fact that systemic immunisation of mice by live *N. lactamica* Y92-1009 culture has elicited a high serum bactericidal response against a plethora of *N. meningitidis* strains, including model, serogroup B strain, *N.*meningitidis MC58 (Li *et al.*, 2006). The development of an anti-meningococcal, *N. lactamica* OMV vaccine was found to be safe (only a few, minor side effects) and elicited broad-spectrum, opsonophagocytic activity against serogroup B meningococci but showed only modest increases in SBA titres (Gorringe *et al.*, 2009).

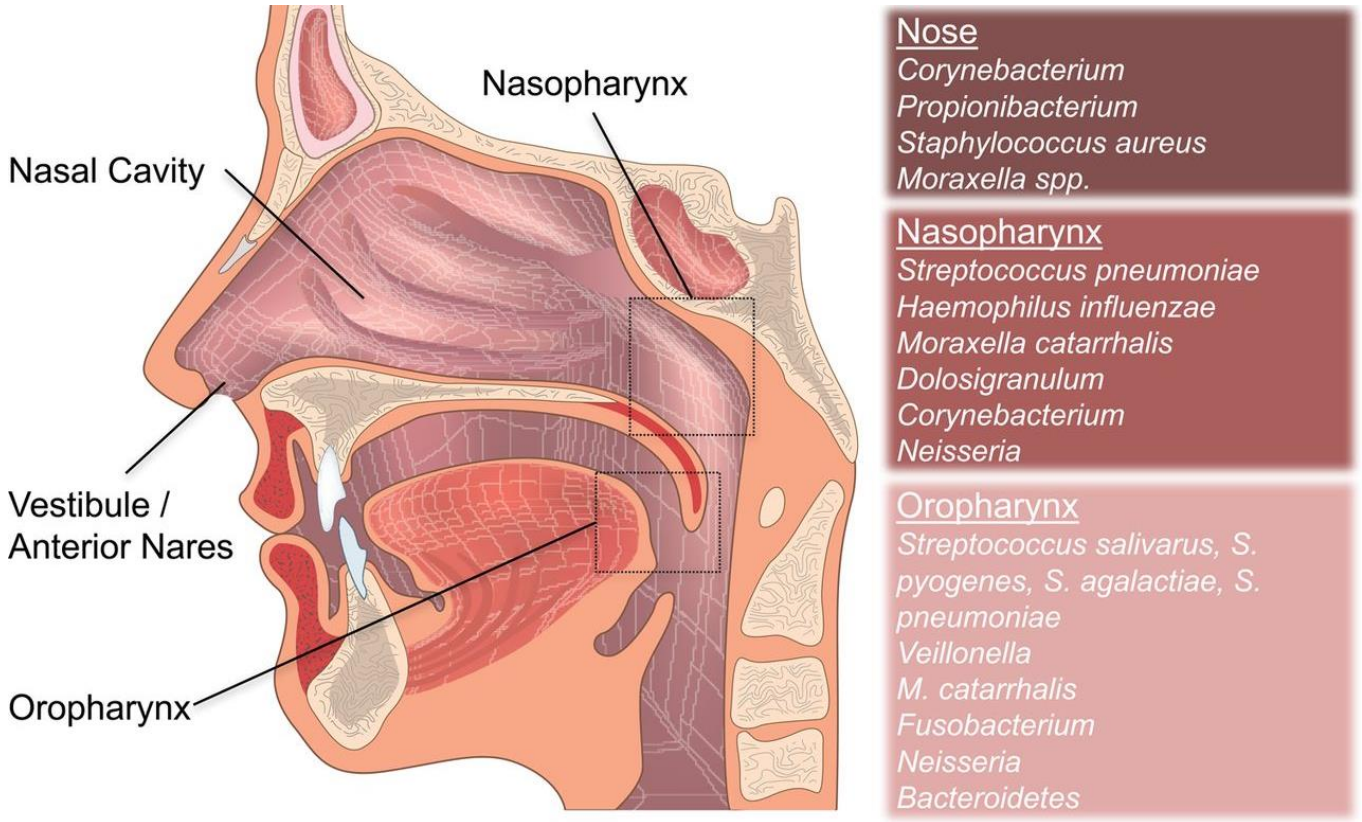
The carriage of commensal *N. lactamica.* has been thought to be confer the induction of cross protective antibodies via OMVs and LOS, against meningococci. These structures show a high degree of homology between both bacteria and produce antigens that are fully cross -reactive with one another. An experimental colonisation study of 61 student-volunteers (Evans *et al.*, 2011) showed that those students that were inoculated with *N. lactamica* Y92-1009 showed very little carriage of *N. meningitidis.* Carriers developed weak opsonophagocytic, anti-meningococcal antibodies but negligible amounts of serum bactericidal response. The study also theorised that some individuals were intrinsically resistant to *Neisseria* carriage. As a follow up, suppression of meningococcal carriage by induced *N. lactamica* carriage was demonstrated (Deasy *et al.*, 2015). This block randomised, human challenge study divided 310 students into two cohorts, inoculated half with 10^4 CFU ml^{-1} *N. lactamica* Y92-1009 and used a sham control for the others. Carriage rates of both *Neisseria meningitidis* and *Neisseria lactamica* were monitored in both cohorts over 26 weeks. A baseline meningococcal carriage rate of 22% increased to 33.6% during the course of the study. Among the 34% of individuals in which *N. lactamica* carriage was established, carriage rates fell from the baseline to 15%. This effect was observed as being broad-spectrum, showing no preference in reducing carriage of a specific meningococcal clonal complex/sequence type. The inhibition of meningococcal carriage was observed in *N. lactamica* carriers due to carriage displacement and once meningococcal carriage was displaced, it was not re-acquired by the volunteers over the course of the study.

1.2 The nasopharynx: the niche and the microbiota

The nasopharynx (NP) is part of the upper respiratory tract (URT). This region contains the anterior nares, nasal, middle ear and oral cavities, sinuses, Eustachian tube, mouth, pharynx (throat) and larynx (voice box). The NP is located the topmost part of the pharynx connecting it with the nasal cavity above the soft palate. Due to its proximity to the cavity, the NP is exposed to a constant flux of air, is lower than body temperature (approximately 34°C, relative to host location and season), and is rich in oxygen and blood (Marks, Reddinger and Anders, 2012). This means that the NP is an area that requires specific adaptions for successful microbial colonisation and therefore possesses different microbial communities from other regions of the URT (Yan *et al.*, 2013; Laver, Hughes and Read, 2015; De Boeck *et al.*, 2017)(**Figure 1-2**). Studies have also suggested that URT regions possesses microenvironments. Therefore, any reports of bacterial interaction must be viewed in the backdrop of this dynamic and complex environment (de Steenhuijsen Piters, Sanders and Bogaert, 2015; Cleary and Clarke, 2017).

The advent of low cost sequencing coupled with metagenomic work done by organisations such as the human microbiome project (Turnbaugh *et al.*, 2007) have allowed researchers to identify and quantify microbiota occupying various niches within the human body. This has been done using techniques such as *16S rRNA* sequencing where clustering approaches such as operational taxonomic units (OTU) allow phyla and genera to be expressed as percentage proportions. The nasopharynx supports a diverse microbial environment which varies from person to person. Work done on the microbial composition of the NP has predominantly focused on children who present with an less abundant and diverse composition of microbiota than that found in adults (Stearns *et al.*, 2015). Factors influencing the composition of the NP microbiota include disease status. The NP microbiota has been found to contain higher numbers of non-*Neisseria* pathobionts such as *Streptococcus pneumoniae*, *H.influenzae, Staphylococcus spp.* and *Moraxella catarrhalis* among study participants presenting with both bacterial and viral infections of the URT and acute otitis media (Garcia-Rodriguez, 2002; Chonmaitree *et al.*, 2017). Another factor influencing NP microbiota is smoking (Greenberg *et al.*, 2006; Charlson *et al.*, 2010)

Figure 1-2 Features of the upper respiratory tract and variable compositions of microbiota



Due to the differences in local environment caused by anatomical structure and epithelial type. Regions of the upper respiratory tract contain variable microbial communities with diversity increasing with distance from cavity exposure. General, pre-dominant, bacterial taxa are indicated for each region above.

Image taken from (Cleary and Clarke, 2017)

1.2.1 Summary of metagenomic research on the nasophayngeal microbiota in healthy, adult humans.

While the majority of metagenomic approaches towards NP microbiota has occurred in children and/or diseased individuals. Our experimental demographic is healthy adults. This section will serve to summarise metagenomic research on this demographic. A Chinese study looking at nasopharyngeal microbiota among 10 undergraduates (Ling *et al.*, 2013) found Actinobacteria (38%), Firmicutes (32%), Proteobacteria (5%) and Bacterioidetes (>1%) as the predominant phyla out of 14 detected. At genus level, the study found *Corynebacterium* (32.4%), *Dolosigranulum* (15.3%), *Staphylococcus* (7.7%), *Lactobacillus* (5.4%) and *Propionibacterium*, *Gardnerella*, *Anaerococcus* and *Prevotella* (1-2%) as the most dominant genera. A British study of 28 volunteers (after filtering out data from an additional 10, natural, pneumococcal carriers) found *Corynebacterium* (37.7%), *Staphylococcus, Dolosigranulum,* and *Streptococcus* (9-12%) as the dominant genera out of 30 identified in total (Table S2, (Cremers *et al.*, 2014)). *Neisseria spp.* were detected in 7/28 participants with a 0.5% OTU proportion. A Canadian study that sampled 19 adults using higher-resolution sequencing found high OTU levels of Firmicutes: *Staphylococcus* (17.7%), , Lachnospiraceae genera (6.9%), *Streptococcus* (4.9%) Carnobacteriaceae genera (4.7%), Actinobacteria: *Corynebacterium* (13.3%), *Bifidobacterium* (2.9%) and Proteobacteria *Pseudomonas* (4.1%) (Table S3, raw text file edited in spreadsheet, (Stearns *et al.*, 2015)). *Neisseria* species were detected in 17/19 participants at 1.2% abundance. A Belgian study of 92 volunteers used high resolution sequencing twinned with a different approach to catalogue microbial diversity (amplicon sequence variants (ASVs) instead of the OTU clustering) than the previous studies (De Boeck *et al.*, 2017). While this allowed them to sub-group genera more finely, their raw data is not available to be interrogated in a comparable way. In contrast to data based on *16S rRNA* sequencing/ OTU clustering, this study found that that just less than half of their participants were pre-dominantly colonised by one specific genus. *Moraxella* (19.6% Participants), *Streptococcus* (13% participants), *Fusobacterium* (8.7% participants), *Neisseria* (2.2% participants) while just over half of participants presented with “intermixed bacterial profiles” with varying levels of *Staphylococcus, Corynebacterium* and *Dolosigranulum* abundance.
Taken together, these four studies on nasopharyngeal microbiota among healthy adults suggest that; a) geographical location is a variable in NP-microbiota composition, b) *Staphylococcus, Corynebacterium* and *Dolosigranulum spp.* are abundant in the NP-niche, c) the genus of the organism of interest of this thesis, *Neisseria* has a low prevalence and abundance compared to other nasopharyngeal colonisers and d) as reviewed in (Hugerth and Andersson, 2017) the

sequencing method and metagenomic metric defining microbiota composition could lead to varying results.

1.3 Sequencing

1.3.1 Overview and definitions

DNA sequencing informs about the genomic content of a sample, what combination and number of bases; adenine, thymine, cytosine & guanine (henceforth referred to as A, T, C & G) generates the identity of coding sequence, a series of bases that are known to form a given protein product. This protein product, a combination of amino acids, is a gene. When all genes (and therefore all the coding sequence) from a single biological organism as well as all the non-coding DNA (intergenic segments) are fully sequenced, this data is known as a whole genome sequence. It is also important to note that the range of sequencing targets can be as narrow as a single gene or collection of genes of biological interest. While not utilised in this thesis, RNA sequencing describes the presence and quantity of RNA, the transcribed form of DNA. This is most commonly used to study what genes are actively being expressed in the sample instead of merely being present. Bacteria are commonly subjected to whole genome sequencing (WGS) and a wealth of sequencing information is currently accessible to microbial researchers around the world in the form of public data repositories. These can be vast, accepting sequencing data from practically any sample (i.e. NCBI: REFSEQ/Biosample repositories) or can be personalized for an organism of interest. (Examples include the 1000 genome project: containing sequences from *Homo sapiens* samples & pubMLST: a bacteria specific database housing sequence data from many genera.) The advent of increasingly easier and cheaper next generation sequencing (NGS) techniques has led to a wealth of DNA and RNA sequencing. Advances in computing have also changed how NGS data is interpreted and analysed. And a largely open source approach adopted in developing software for bioinformatics analyses has allowed unparalleled levels of global collaboration and method refinement.

1.3.2 Read length, Genome assembly, Quality Control and Annotation

The *in silico* DNA *or* RNA data generated by a sequencer corresponding to whatever sample was processed; are known as sequencing reads. Hundreds of these overlapping strings of data can be examined to resolve errors, generate a consensus sequence to be assembled into contigs or mapped to reference assembly to detect variants. (N.B. These concepts are excellently reviewed in (Ekblom and Wolf, 2014)).

Read length and per-base accuracy are important concepts to consider when utilising either technology. Shorter read lengths correlate with more contiguous assemblies and higher assembly error rates as there is greater potential for error towards the beginnings and end of reads and

subsequently the beginnings and ends of contigs assembled from these reads. Long read sequencers (e.g. SMRT sequencing) avoid this issue by sequencing reads of a higher size (up to mega bases long) relative to short read sequencers (Illumina machines, typically 75bp-300bp.) However, this often comes at a cost of lower per base accuracy and enhanced difficulty with repeated sequences. Shorter sequenced reads typically have a much higher per base quality. The Illumina short paired end technology used for isolate resequencing in this study has been found have one of the highest phred quality scores, the units used to measure per base accuracy rates (Loman *et al.*, 2012).

An assembly is a consensus sequence put together by its compositional sequencing reads. The size of an assembly can be the size of a gene as well as whole genome encompassing assemblies. In the case of the latter (whole genome assembly), assembling algorithms position sequencing data present in reads to resolve contiguous stretches of sequence. Once a genome sequence has been assembled into contigs, the next essential step is linking them together, this is known as scaffolding and is guided by mate-pair data which help to resolve repeats or identify sequencing errors (such as missing data) between contigs (Wetzel, Kingsford and Pop, 2011). The accuracy of the assembled whole genomes can be gauged partly by metrics such as contig number, N50 value and coverage. Given a set of assembly contigs ordered from smallest to largest, an N50 value is defined as the size of the contig found to contain 50% of the genome sequence. For example, a typical neisserial genome is 2.2mb. As the contigs describing this genome are calculated, the contig containing the 1,100,000th base would be defined as the N50 contig. In higher quality, low contig genomes this number would be higher as even the smallest contigs should contain larger amounts of sequence since there are less of them overall. Coverage is defined as the average number of reads clarifying the presence of a given nucleotide. Although there is as of yet no gold standard test to judge genome assemblies, attempts have been made to develop software that takes assemblies as input and outputs statistics of assembly accuracy. The first of these, "Assemblathon" (Earl *et al.*, 2011) evaluated 41 short read *de novo* genome assembly algorithms on a variety of sample DNA under the assumption that the best assemblers can reconstruct a genome assembly with high coverage and good accuracy irrespective to sample type. GAGE (Salzberg *et al.*, 2012)is a similar but updated evaluation of genome assemblers.
Annotation is the process by which coding regions are differentiated from non-coding (intergenic) regions and genetic information is ascribed to a whole genome assembly's sequence data. This can be done by either submitting an assembly to an annotation portal (I.E. RAST (Aziz *et al.*, 2008) or generating customised annotations via open source pipelines such as PROKKA (Seemann, 2014) which uses sub-programs for gene and RNA finding as well database matching.

1.3.3 Historical background and first-generation sequencing

In 1977, Fred Sanger and colleagues developed the dideoxynucleoide (ddNTP)/chain termination method of sequencing (Sanger, Nicklen and Coulson, 1977). Much like the polymerase chain reaction (PCR), this technique used a primer adjacent to the region to be sequenced, one strand of DNA as a template and added deoxyribonucleotide triphosphates (dNTPs) which were incorporated onto the elongating strand using DNA polymerase. Where this method differed from PCR is the next step. The reaction took place in four tubes with each tube containing a specific dideoxy modified base (ddGTP, ddATP, ddCTP and ddTTP), unspecific dNTPs and a radioactively labelled primer. When the ddNTP was incorporated onto the synthesising DNA strand, polymerisation was terminated. Because the polymerase had the option of attaching either a standard dNTP or modified ddNTP onto the DNA template the net result was a production of DNA fragments of differing sizes from each tube. These fragments were pooled, separated via gel electrophoresis and when viewed side by side allowed the interpretation of the complementary strand which could be used to determine the original DNA template sequence (explanation adapted from (Moran, 1994)). This technology was used a year later to sequence the whole genome (5,375 nucleotides) of bacteriophage ϕX174 (Sanger *et al.*, 1978) and was followed by the first use of what is now called bioinformatics, a computer program used to order the sequence gel readings (Staden, 1979). By 1986, Applied Biosystems built the first commercial, automated sequencing machine, the ABI Prism 370A. In 1996, this was followed by the release of Prism 310 machine which downscaled the gel-electrophoresis reaction to occur in long and thin capillary tubing (Springer, 2006). By this point, sequencing technology had sufficiently evolved to target bacterial whole genomes and the first whole genome sequence of a free-living organism, *Haemophilus influenzae* was performed by Craig Venter and colleagues (Fleischmann *et al.*, 1995)

1.3.4 Second generation (Next generation) Sequencing

While the eventual goal of the 1st generation of sequencers was an accurate whole genome sequence. The goals of the second generation of sequencers were a) the parallel synthesis of millions of sequencing reads, b) reducing the cost and therefore increasing the availability of sequencing c) the removal of electrophoresis in detecting sequencing output and d) increasing the speed of sequencing compared to the first generation (Kchouk, Gibrat and Elloumi, 2017).

A new way of sequencing was developed that tracked the proportion of light emitted due to pyrophosphate synthesis via the dual enzymatic action of ATP-sulfurylase and luciferase. Sequence data could be inferred as each observable nucleotide from the DNA template was washed through the reaction (Nyrén, 1987). This technique known as pyrosequencing allowed the

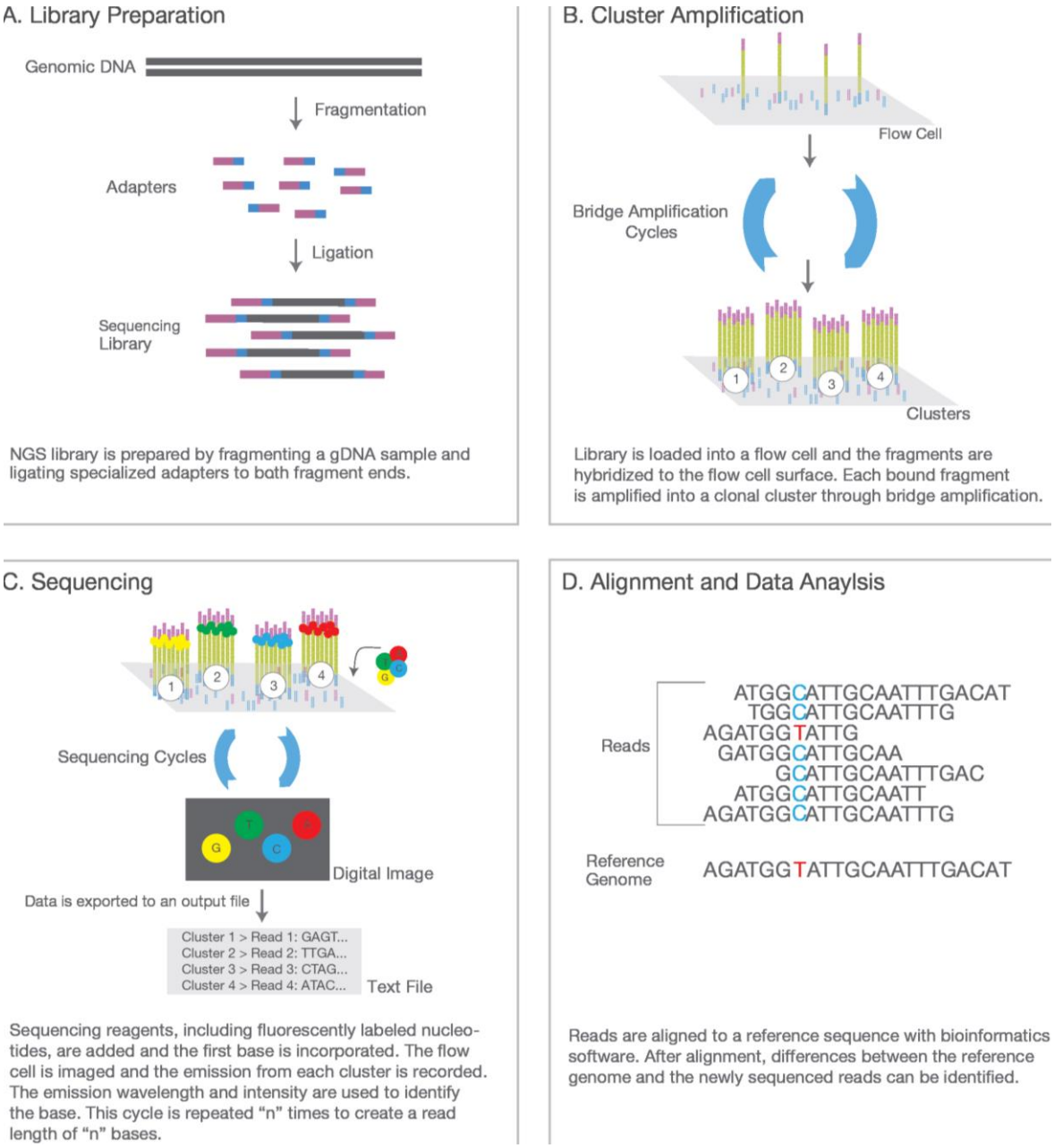
observation of the real time DNA polymerase synthesis (Ronaghi *et al.*, 1996). The technique was licensed to 454 technologies who coupled it with the use of clonally-populated, DNA-coated beads (created with emulsion PCR) interacting with enzymes and dNTPs over a picolitre, reaction plate to monitor pyrophosphate release (Margulies *et al.*, 2005). This resulted in 400-500bp read lengths per well (out of millions of wells sequenced in parallel) and produced orders of magnitude greater output, with high accuracy (outside of homopolymeric sequence) for lower cost compared to first generation sequencing. The release of the 454 (later Roche) GS20 sequencer utilising this technology was the advent of next-generation, high-throughput sequencing (Wheeler *et al.*, 2008).

The 2004-2010 period saw the introduction of a number of companies and sequencing protocols such as Applied Biosystems' SOLiD and Life Technologies' Ion Torrent. All the technologies mentioned so far in this chapter use the sequencing by synthesis method as they monitor the effect of the DNA polymerase synthesis enzyme to detect output. Sequencing by oligonucleotide ligation and detection (or SOLiD) utilised the mismatch sensitivity of DNA ligase to identify the nucleotide present at a given point in a sequence (McKernan *et al.*, 2009). This method had high per base accuracy even in homopolymeric stretches of sequence but struggled with palindromic sequences (Huang *et al.*, 2012). Ion torrent used a similar methodology to that outlined for pyrosequencing with the caveat that nucleotide integration was detected due to a difference in pH caused by the release of hydrogen ions during polymerisation. This detection was performed using a complementary metal-oxide-semiconductor (Rothberg *et al.*, 2011). This method is fast compared to most other sequencers with a low cost but suffers from inaccuracy in homopolymeric sequence.

In retrospect, the most popular NGS technology was developed by Solexa and purchased by Illumina. This "bridge-amplification" technology was a variation of sequencing by synthesis which used adapter bracketed DNA passed over an amassment of complementary oligonucleotides which via solid phase PCR produces neighbouring, clonally-populated DNA clusters. This is repeated for both flow-cell-binding DNA strands which are forced to arch to prime the next round of polymerisation (Bentley *et al.*, 2008). The Illumina sequencing workflow is summarised in **Figure 1-3.** A high level of base call accuracy is maintained throughout this method due to the fact that base calls are made based on flow cell emission signatures unique to adenosine, cytosine, guanine, and threonine. A further improvement using this method of sequencing was the generation of paired end reads. Since the distance between each paired end is known, there is an extra layer of positional data available over unpaired reads. These can be used to more accurately map reads to a sequence assembly and are helpful in resolving homopolymeric regions of DNA.

N.B. The following review articles were helpful in writing this section and should be consulted if the reader wishes for more information on next generation sequencing (Mardis, 2008; Voelkerding, Dames and Durtschi, 2009; Heather and Chain, 2016; Kulski, 2016; Kchouk, Gibrat and Elloumi, 2017)

Figure 1-3 Overview of Illumina NGS sequencing workflow



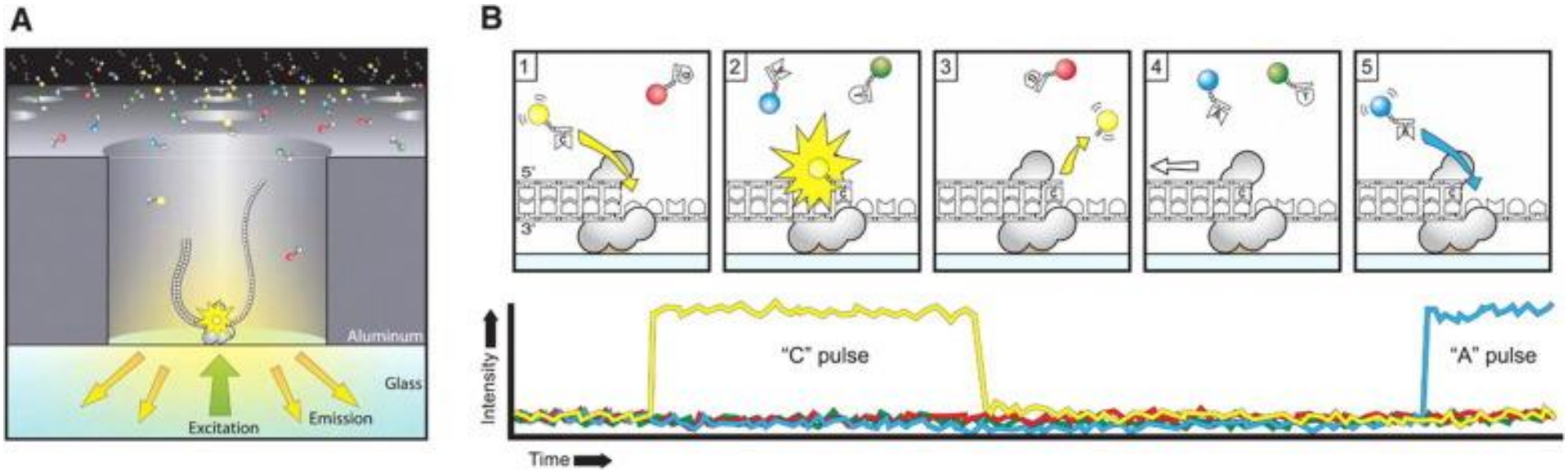
(Image adapted from (Illumina, 2017)). Extracted genomic DNA (A) is fragmented and modified with adapter sequences on both ends. This is repeated until the threshold required for the generation of a library of adapter annealed sequences is reached (B). The library is hybridised onto flow cells and amplified up into a cluster. (C) Reagents are added to the cluster, the flow cell image and emission wavelength/intensity are used to identify bases, these are recorded. This step is repeated multiple times until (D). The newly generated raw sequencing reads can be accurately mapped onto a reference assembly or in the absence of a suitable reference, assembled *de novo.*

1.3.5 Third generation of sequencing

Having reduced costs and running times while increasing informational output during the second generation. The goal of the third generation of sequencing technologies was to increase read length via the direct sequencing of DNA molecules in order to produce higher quality genome assemblies with fewer contig breaks (Bleidorn, 2015). The two major technologies involved in this generation are Pacific Biosciences' single molecule real time sequencing (SMRT) and Oxford Nanopore Technologies' nanopore sequencing. With the latter company having the additional goal of producing portable sequencing machines (Quick *et al.*, 2016).

The primary difference between SMRT sequencing and other types of genome sequencing is the use of the zero-mode wave guide (ZMW). This is a ~70nm diameter, ~100nm depth, structure that guides electromagnetic waves in a two or three-dimensional plane, which allows the isolation of a single instance of nucleotide binding to DNA polymerase (Levene *et al.*, 2003). A fluorescent signal is generated by each individual nucleotide, which decays following the DNA polymerase binding event, causing a registered signal decay (reaction displayed in **Figure 1-4**). SMRT cells contain tens of thousands of ZMWs, this allows for a massively parallelised sequencing approach that simultaneously detects modifications like methylation. As a result, SMRT has been credited with rejuvenating the methylation pattern recognition field (Fang *et al.*, 2012; Sater *et al.*, 2015). The long reads generated by SMRT sequencing have allowed bacteria to be assembled to a "finished" (i.e. one contiguous, circular assembly) standard (Koren *et al.*, 2013). The disadvantages of using SMRT sequencing is high cost, lower throughput and a higher base error rate. The latter concern has prompted the use of short read sequences from the same sample to correct long read finished assemblies (Walker *et al.*, 2014) as well as hybrid assembly (Au *et al.*, 2012)

Figure 1-4 SMRT cell reaction summary



The target double stranded DNA is ligated with hairpin adapters to form a closed, single-stranded template called a SMRTbell. [A] When the SMRTbell is loaded onto a SMRT cell chip it diffuses into a sequencing unit with a single, immobilised, DNA polymerase at the bottom called a zero-mode waveguide (ZMW). The polymerase binds to either hairpin adapter of the SMRTbell to start DNA replication.

[B] The four nucleotides G (red), C (yellow), T (green) and A (Blue) are labelled in distinct fluorescent dyes which release pulses of light with varying emission spectra as they're held by a polymerase. Here both cytosine and adenine bases are held [B2, B5] which excites their fluorescent signal intensities from the baseline. Following this output, the dye-linker pyrophosphate product is cleaved from the nucleotide and diffuses out of the ZMW [B3] while the polymerase moves to the next position for another fluorescently-labelled nucleotide to associate with the enzyme's active site [B4]. This figure is adapted from (Rhoads and Au, 2015).

1.3.6 Challenges for and limitations of bacterial WGS and bioinformatics

The breakthroughs of various strategies with high DNA input, parallelised sequencing technology has been dubbed the "genomics revolution". Between 2004-2010, the capabilities of DNA sequencers grew at a rate outpacing even Moore's law (the number of transistors in a dense integrated circuit doubles approximately every two years) by doubling every five months (Stein, 2010). For example, Solexa's 2005 genome analyser produced one gigabase of data per run. By 2014 the data output by Illumina HiSeqX was 1.8 terrabases, a thousand fold increase over 9 years (Illumina, 2017). Therefore, a major rate-limiting step for genome sequencing is the technology powering both the processing and storage capacity of increasingly vast quantities of sequencing data (Muir *et al.*, 2016).

Universal access to the internet has changed the way scientists both share and think about generated data. A number of essential tools for bioinformatics and methods of learning the skills required to process NGS data utilise open-source design (Quackenbush, 2003). Organisations such as the Ensembl project (Yates *et al.*, 2016)and NCBI (Tatusova *et al.*, 2016) have made bacterial genome browsing, submission and annotation open to all. While this has been successful in diversifying the number of research groups undertaking sequencing and subsequent informatic analysis on bacterial species, it has posed a challenge of creating specialist databases so species-specific research can be usefully interpreted, curated and iterated upon. The majority of work presented in this thesis focusing on a commensal *Neisseria* species would be impossible without a *Neisseria*-specific resources and databases such as PubMLST Neisseria (Jolley and Maiden, 2010) or NeisseriaBase (Zheng *et al.*, 2016). But as the number of novel, sequenced genera and species grows, the number of well-curated databases must grow in tandem and presents an ongoing challenge to the bacterial genomics community.

1.3.7 Genome-wide association studies

A genome-wide association study (GWAS) is defined as the observational study of a genome encompassing set of genetic variants. Since multiple chapters of this thesis deal with microevolution that occurred during a longitudinal study, it would be pertinent to class those chapters as an example of GWAS. Important limitations to consider in the experimental design of such a study include a requirement for stringent and fine-tuned statistics in calling variants (Ribeiro *et al.*, 2015), genotyping errors (Pompanon *et al.*, 2005) and a lack of functional information on hypothetical proteins of interest (Pearson and Manolio, 2008). These limitations and measures taken to address them will be discussed in greater detail in chapter 8.

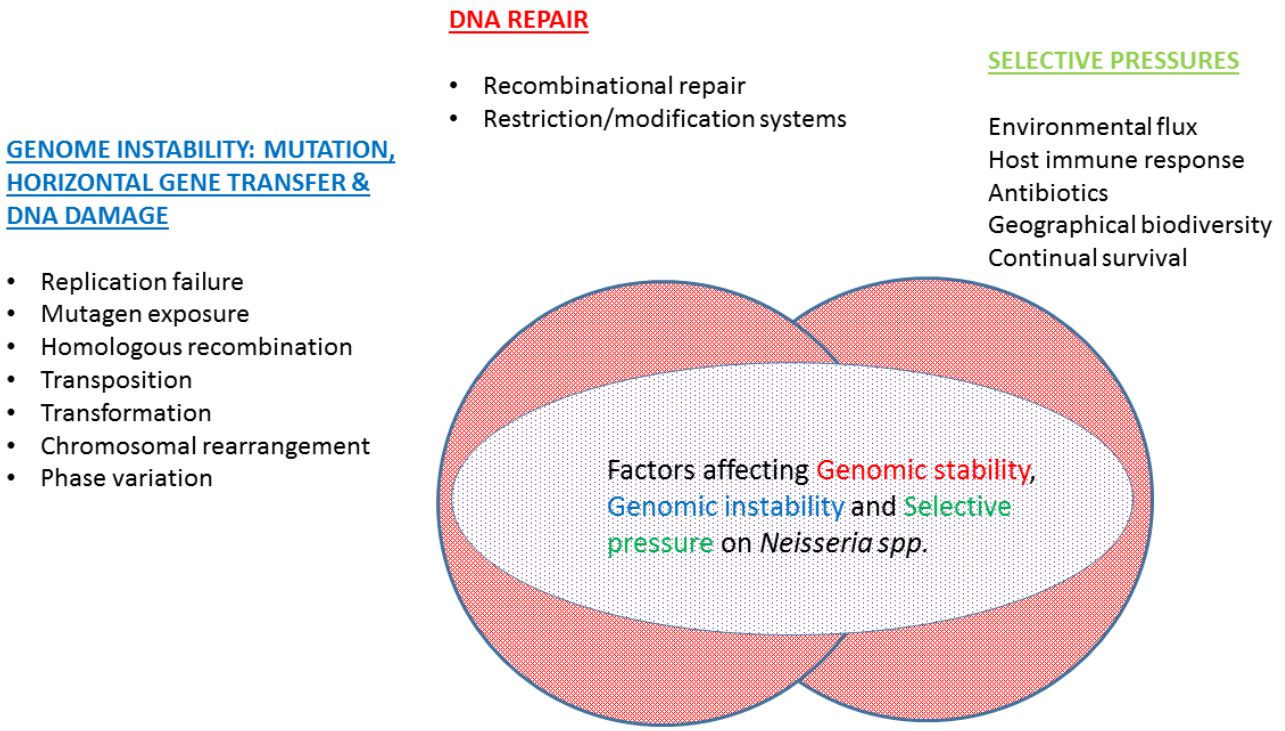
1.4 Mutations and mutational analysis

1.4.1 Point mutations and single nucleotide polymorphisms

A bacterial organism's genetic structure is built on the simple foundation of 4 bases, A, T, C and G. Three of these bases in a sequence (codon) produces an amino acid. These amino acids combine together, like constructive blocks, to form a protein. Mutations are a common form of genetic variation made more likely by enhanced selective pressure and genome instability as well as a failure of genome stabilising factors (summarised in **Figure 1-5**.). It has been thought that a genome more pliable to mutations alongside vast numerical advantages provides a bacterial species with the resources necessary to survive and thrive within a hostile and complex environment (Aguilera and García-Muse, 2013; Darmon and Leach, 2014).

If a mutation changes the protein coding sequence/ putative transcribed amino acid sequence it is described as non-synonymous. If a mutation does affect the CDS but not the resultant protein being encoded, it is described as being synonymous. Point mutations are single base modifications. They can insert or delete (insertions and deletion events are called indels) a base in a DNA coding sequence (dCDS) which may in turn result in frame shift, a mass divergence in the original protein coding sequence (pCDS). Single nucleotide polymorphisms or SNPs are isolated point mutation substitution events in a given dCDS. SNPs have the potential to activate or inactivate an entire gene. This is because multiple codons of DNA bases can code for the same amino acid residue. When a mutation changes the codon but not the resultant amino acid, it is same sense or silent. Instead of an amino acid, three of the codon combinations encode stop codons, sequences that terminate transcription and prematurely stop protein synthesis. A SNP leading to a premature stop codon is called a nonsense mutation. Amino acids can be further discriminated by whether they possess polar or non-polar charges. Therefore, it is of greater consequence to the gene as a whole if a mutation leading to a change in amino acid (missense mutation) is conservative or non-conservative. A conservative mutation will generate an amino acid with the same charge as the original amino acid. The product of the non-conservative mutation has a different polar charge and could therefore negatively affect the tertiary structure of the encoded protein. Examples of the effect mutations have on coding sequence are displayed in **Figure 1-6**. These mutations could be detrimental for a species; leaving its progeny with no chance to survive in the environment it finds itself in. Or they may confer a selective advantage, such as modifying a gene to enable the bacterium to circumvent the effects of an antibiotic or radiation (Bryant, Chewapreecha and Bentley, 2012).

Figure 1-5 Overview of sources of genetic stability, instability and selective pressure on the genomes of *Neisseria* spp

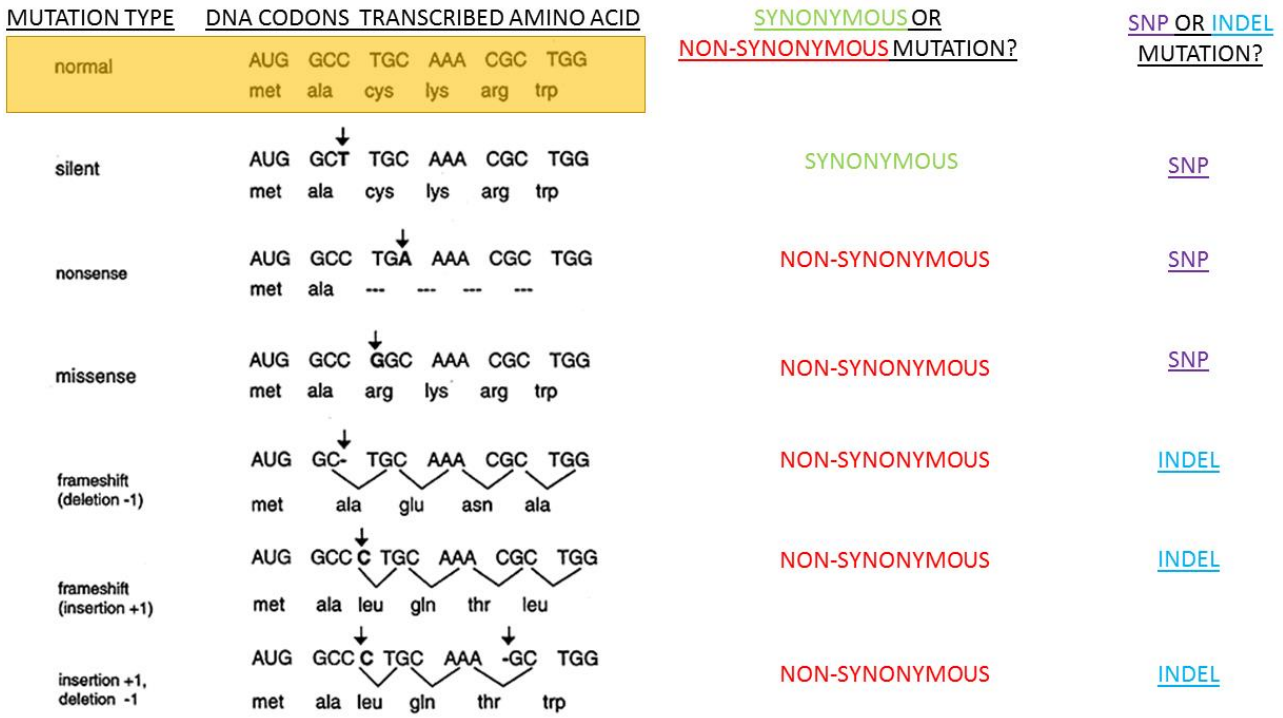


.

(Image adapted from page 123, Handbook of Meningococcal Disease (Frosch and Maiden, 2006))

N.B. Recombination can act as both a source of DNA repair and genome instability.

Figure 1-6 Mutations: nomenclature examples and transcriptional consequence



This figure describes the mutations occurring to the original DNA codons/amino acid sequence (highlighted in gold, "AUG-GCC-TGC-AAA-CGC-TGG", "met-ala-cys-lys-arg-trp") The bases undergoing mutation are indicated with arrows. A hyphen ("-") indicates a deleted base and three hyphens together ("---") indicate a deleted amino acid. (Figure adapted from Strum, 2015)

1.4.2 Phase variation

Organisms exist in a rapidly changing host environment in terms of nutritional availability and innate/adaptive immune defence system. This can force bacteria to counter this selective pressure with a diverse array of metabolic pathways and rapid but stochastic phenotype variation by hypermutable mechanisms and a short generation time. The mechanisms responsible for this effect can be as a result of site-specific and homologous recombination (Didelot and Maiden, 2010) and changes in simple sequence repeats causing frameshift (Bidmos and Bayliss, 2014). The effect of these mechanisms (**Figure 1-7**) that introduce reversible and stochastic mutation to affect the downstream expression of genes are described as phase variation. This phenomenon has been described in many bacterial genera associated with the infection of multiple sites in the human host (Van Der Woude and Bäumler, 2004). Loss or gain of repeat components in repetitive sequence structure in the coding sequence of a gene may cause frameshift, leading to alteration of the transcription/translation of said gene and a subsequent loss or gain of genetic function. The specific adaptations that enable the organism to survive best are maintained and permeated in the population colonizing the host tissue. The average mutation rate of a bacterial gene has been shown to increase a million-fold in certain hypermutable genes dubbed “contingency loci.” (Moxon, Bayliss and Hood, 2006; Bayliss *et al.*, 2008). Many of these loci encode outer membrane associated proteins which functionally act as mediators in bacteria-host or bacteria-phage interactions.

Classically, homopolymeric (I.E. “A-A-A-A-A”) and multimeric sequences (I.E. “G-C-A-G-C-A-G-C-A”) have been difficult to analyse with short read sequencers. Draft, *de novo* genome sequences assembled using earlier iterations of sequencers demonstrated contig breaks (and subsequent loss of confidence and coverage) in many of these repetitive sequences. However, improvements in per-base-accuracy in short read sequencing as well as a using longer read sequencing to resolve areas around repetitive sequences, have helped to alleviate this problem (Goldberg *et al.*, 2014).

Figure 1-7 Mechanisms underpinning Phase variation

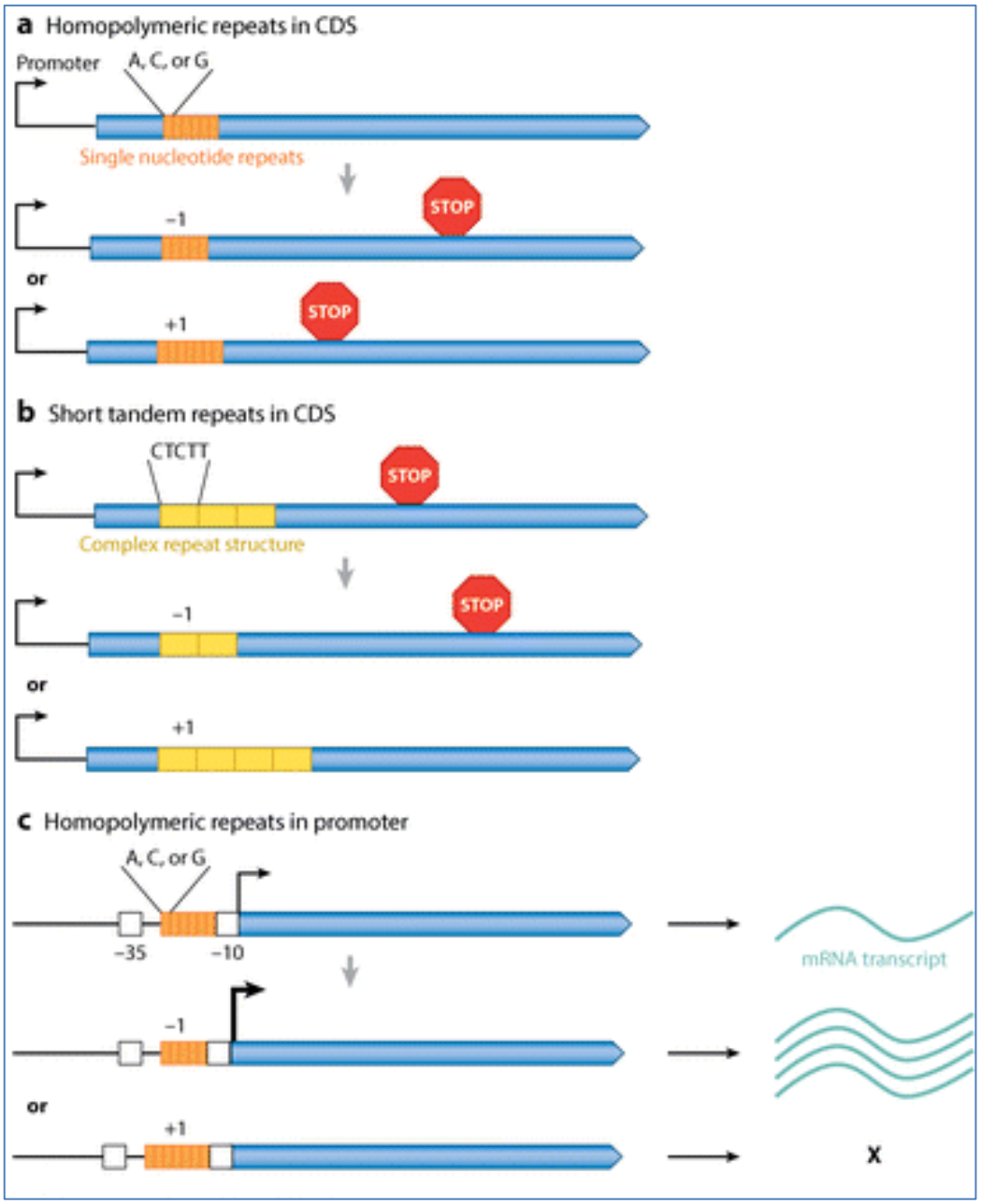


Figure 1-7 legend

A: An insertion or deletion in the homopolymeric tract located in the coding sequence of a gene can completely halt expression as shown (**stop symbols)**. These types of phase variants have copy number expression patterns in triplicates. In the first example above, a 6-base tract (**orange section)** phase varying to 9 or 12 bases may still allow expression. But a 5 or 7 base tract (as shown in the example) does not.
B: Multimeric repeat tracts work in the same way as homopolymeric tracts but have more complicated copy number expression patterns. In the second example, the tract does not allow expression until the insertion of an addition "C-T-C-T-T" residue (**yellow segment**). This brings the number of "CTCTT" residues from 3 up to 4 (15 bases to 20 bases) allowing expression. Deleting a CTCTT residue (15 bases to 10 bases) also affects expression negatively.
C: Intergenic promotor associated phase variants are typically located between the -10 and -35 binding sites downstream of the phase variable locus. These variants affect the spacing of the transcriptional binding site which has knock on effects for efficiency of transformation. If the phase variable examples (A and B) could be thought of as "On/Off switches", this type is a "dimmer switch", modulating protein expression. In the example shown above, a deletion of one base to the 6 base homopolymeric tract (**orange section)** led to an increase in mRNA transcription from baseline levels. When one base was added instead, expression was halted. (Figure taken from (Bidmos and Bayliss, 2014)).

1.4.3 Phase variable, sub-unit modifiers in *Neisseria*: *pgl* and *lgt* loci

Post translational modifications are strongly suggested to be important to the continued lifestyle of *Neisseria spp.* (Snyder and Saunders, 2006). A large proportion of all mutations detected were in genes involved in transferring chemical groups to alter the structure of pilins and lipo-oligosacharide (LOS) structures. The core *pgl* genes are *pglB, pglC, pglD* and *pglF* which are involved in the synthesis of undecaprenyl diphosphate or in the case of the latter, pilin translocation into the periplasm (Børud *et al.*, 2010). *pgl*A acts as a galactosyltransferase and modifies the monosaccharide molecule generated by the core *pgl* genes into di and tri-saccharide forms. These can be modified even further by *pglL* which is responsible for O-acetylation. *pglI* shares functional homologs with pathogens, *Pseudmomonas aeruginosa* (*wbpC*) and *Salmonella typhimurium* (*oafA*) and is also responsible for the addition of acetyl molecules to these structures (Faridmoayer *et al.*, 2008).

The LOS structure of *Neisseria* spp. consists of an outer membrane anchor and an inner oligosaccharide core. *lgt* genes A-E are glycosyltransferases responsible for the addition of different sugars to the alpha chain of this core. This is responsible for the generation of such an immunologically reactive LOS endotoxin (Rotman and Seifert, 2014). Additionally, lgtG when phased ON has been shown to aid immune evasion of the bacterium from monoclonal antibody b5 (Bayliss *et al.*, 2008). The net result of these post-translational modification genes are contributions to an ever-expanding repertoire of surface antigens via LOS and type IV pili (Rotman & Seifert., 2014.).

1.4.4 Sources of variation: Restriction-Modification systems

The neisserial genome is subject to many sources of stress and change through DNA damage or mutation. It is thought the restriction modifications systems of the *Neisseria* police some of these mutations and regulate them. The main reason for this is that natural, unbeneficial modifications to these restriction modifications systems have been shown to be responsible for enhanced rates of virulence dissemination and genetic transfer into the target genome (Rotman and Seifert, 2014). Recent advances in long read sequencing technology have identified the type III restriction, *mod* genes (Tan *et al.*, 2016); mediating epigenetic regulation by altering expression through methylation pattern change.

1.4.5 Sources of variation: Recombination

Recombination is a driver for the rearrangement of genetic material, breaking and re-joining DNA strands. It can occur as a result of DNA repair or during meiosis. But among naturally competent organisms like *Neisseria,* a penchant for transformation and homologous recombination has driven the genetic diversity of the species, acting as another source of horizontal gene transfer between similar strains (Kong *et al.*, 2013). Understanding and analysing recombination is essential to understanding how genetic diversity is created and then maintained within this genus (Didelot and Maiden, 2010). Unlike mutation analysis, identifying the specific direction of a recombination event/ breakpoint is possible by contrasting gene sequences between parent and progeny genomes. These can be quantified and statistically analysed over the course of an experiment. The potential of homologous recombination to influence genome sequence identity and subsequently derived phylogenetic inference has led to a rise in maximum-likelihood based approaches in determining ancestry among recombinogenic bacteria (Croucher *et al.*, 2014; Didelot and Wilson, 2015)

1.5 Experimental aims

The importance of this research will be contextualised below based off topics covered in greater detail in the introduction.

There are no polysaccharide conjugate vaccines available for serogroup B meningococcal disease. Therefore, there has not been an reductive effect on the carriage of serogroup B meningococci eliciting herd immunity, as has been reported for serogroups A, C, W135 and Y (Borrow *et al.*, 2013). Carriage is a pre-requisite for invasive meningococcal disease and serogroup B IMD remains the most prevalent in the developed world (Read, 2014). While recent efforts have generated two recombinant protein vaccines shown to have an immunogenic effect, they have displayed limited evidence of meningococcal carriage reduction (Read *et al.*, 2017; Soeters *et al.*, 2017), particularly to the extent and duration of carriage reduction by *N. lactamica* as observed in a long-term, experimental, human challenge study (Deasy *et al.*, 2015). Artificial inoculation of *N. lactamica* Y92-1009 was shown to reduce meningococcal carriage to a greater extent than even glycoconjugate vaccines successfully developed and implemented for meningococcal serogroups A, C, W135 and Y. The effect of natural *N. lactamica* carriage on meningococcal carriage reduction has been an observable phenomena among infants for 30+ years (Gold *et al.*, 1978; Cartwright *et al.*, 1987) but remains poorly understood. Therefore, the potential exists for this organism to be used either as a standalone bacterial therapeutic, or to be paired with vaccine therapy. However, inoculating humans with *N. lactamica* presents risks of its own and the following questions must be answered.

1) Is the carriage of artificially inoculated *N. lactamica* harmless in a wide population?
2) Does *N. lactamica* change genetically while residing in the human nasopharynx?

Chapter 5 investigates the in vivo microevolution of artificially inoculated *N. lactamica* Y92-1009 sampled (~100 isolates) longitudinally over 26 weeks and ~40 volunteers from **the long-term, experimental, human challenge study** (Deasy *et al.*, 2015). *N. lactamica* inoculation was previously shown to have a broad-spectrum reduction on meningococcal carriage. This experiment will allow the genomic interrogation of *N. lactamica* Y92-1009 isolates from; (i) volunteers in which meningococcal carriage was seen to be reduced, (ii) volunteers in which meningococcal carriage was seen to persist and (iii) volunteers in which *N. lactamica* was the sole *Neisseria* coloniser observed during 26 weeks. This allowed the determination of the stability of the *N. lactamica* genome during 6 months of in vivo carriage. Secondarily, this experiment observed for evidence of adaptive mutation; namely, consistent, identical mutations to the genome across multiple volunteers. Considering all volunteers were inoculated from the same bacterial stock, a gross mutational change in all isolates may represent the adaptive transition of

the bacterium from cryostorage to in vivo carriage. The unique resource of isolates from the Deasy study will allow us to comment on the within host evolution of this organism with all volunteers having been inoculated the same way, at the same time, using the same bacterial stock. This is a level of standardisation not normally seen in within host evolution studies which typically monitor disease causing isolates with no background history of strain acquisition (Didelot, Walker, *et al.*, 2016). Monitoring microevolution necessitates the use of a highly accurate assembly. In **chapter 3,** the complete genome sequence of *N. lactamica* Y92-1009 (Pandey *et al.*, 2017) was sequenced, assembled and annotated. As well as describing notable features of interest, repeat sequence motifs were searched and compared with other members of the *Neisseriaceae* in order to better characterise the genome. The assembly was compared to the only other complete *N. lactamica* genome (strain 020-06, (Bennett *et al.*, 2010)) as the discovery of novel segments in *N. lactamica* Y92-1009 may shed light on what caused this strain to tread a different evolutionary path from a common ancestor. The accuracy of this assembly is tantamount to downstream work in chapters 4, 5 and 7. Therefore, it was also evaluated against a previously sequenced *N. lactamica* Y92-1009, draft assembly (Vaughan *et al.*, 2006).

Despite isolates from **chapter 5** being sourced from volunteers carrying *N. lactamica* Y92-1009 and minimally passaged. The cryostorage and *in vitro* recovery of isolates may introduce mutational bias. In **chapter 4**, differences in mutations were observed when comparing the microevolution of *N. lactamica* Y92-1009 isolates grown in *in vitro* versus *in vivo* conditions. This was done by using mutational analysis, a highly accurate reference genome to compare against and **a second, short-term clinical study** in which multiple colony sampling was utilised over 1 month. Since mutations are known to occur stochastically as well as adaptively, this experiment was essential in determining whether mutations observed *in vivo* and thought to be influenced by those conditions also appeared among the *in vitro* cohort. As far as can be searched in the literature, this is the first experiment to contrast mutation differences between *in vitro* and *in vivo* sourced isolates among *Neisseria lactamica.* In addition, looking at the multiple samples per time point study will indicate whether a realistic picture of mutational diversity is being observed as the longer-term study utilises one sample per time point.

Genetic recombination and nucleotide substitution are drivers of diversity that enable natural selection. In *Neisseria spp.*, homologous recombination shapes the diversity of species within the genus and transfers genetic material interspecifically. To further answer questions relating to the genomic stability/mutability of bacterial inoculant *N. lactamica* Y92-1009, **chapter 6** quantified the effect of recombination and identified genes affected by homologous recombination-mediated, genetic importation. The analyses were performed on artificially inoculated *N. lactamica* Y92-1009 sole-colonised and co-colonised with the meningococcus sourced from the

long-term, experimental, human challenge study (Deasy *et al.*, 2015). Furthermore, recombination data from the two cohorts were contrasted with isolates from an additional two volunteers among whom wild type *N. lactamica* and wild type *N. meningitidis* co-colonisation was detected. The detection of co-colonisation among the two volunteers who were not experimentally inoculated with *N. lactamica* Y92-1009 makes them ideal candidates for contrasting the level of recombination observed between a recently established co-colonised relationship (between *N. lactamica* Y92-1009 and resident meningococci) and a more mature, co-colonised relationship (wild type *N. lactamica* and resident meningococci.)

Chapter 7 quantified the extent of genetic diversity found among the representative genomes of every available sequence type of *Neisseria lactamica* by constructing a pan-genome of the species. These metrics were used to compare and contrast core genome size of the pan-genomic analysis versus those calculated from other bacterial species. An experiment of this type hasn't been performed since a genus encompassing *Neisseria* pan genome was calculated (Bennett *et al.*, 2010). Since then, many more *N. lactamica* whole genome sequences have been made publicly available for interrogation. Specifically, a species encompassing, *N. lactamica* pan genome has never been determined up until now. In calculating the pan-genome, the genomic content that distinguishes *Neisseria lactamica* Y92-1009 as a strain within this species was quantified. This data is crucial for investigators using this strain as an experimental inoculant in future carriage studies because it would inform of any potentially "harmful genes (i.e. virulence factors, antibiotic resistance markers, and restriction modification systems) unique to this strain when compared to other *N. lactamica*.

.

Chapter 2: Methods

This section describes the general methods used in experiments throughout this thesis. Most of bioinformatics pipelines and analytical techniques are unique to certain chapters and therefore will be described in the methods sections of chapters 3-7.

2.1 Culture and extraction of bacterial samples

2.1.1 Culturing

N. lactamica stock was plated onto blood plates and grown overnight at 37°C. Colonies were tested with XGAL + phosphate and two blue colonies were sub cultured into 10 X 5ml bijoux tubes, each containing 2 ml Trypticase soy broth + 0.2% yeast extract (Both, Sigma Aldrich). The cultures were grown overnight at 37°C in a gently shaking platform.

2.1.2 DNA extraction

DNA was extracted using the Wizard Genomic Purification Kit. The instructions were followed to extract the DNA from samples using the steps outlined for Gram negative bacteria by the manufacturers.

The following options/modifications to the protocol were as follows
a) In the RNase solution step, the solution was incubated for an hour at 37°C before cooling to room temperature.
b) The DNA was rehydrated overnight by rotating the DNA extract containing Eppendorf in a gentle mixer overnight.

All DNA was extracted using this method. The exception to this was the DNA required for PacBio sequencing elsewhere. DNA extracts were stored at -20°C.

2.1.3 Assessing DNA extract purity

Purity of DNA samples was assessed by examining both the trace and absorbance levels of the 260/280 and 260/230 absorbance ratios in a nanodrop 1000 spectrophotometer. The ratio of absorbance measured between 260nm and 280nm wavelengths is used to assess DNA and RNA purity. A ratio value of ~1.8 is considered pure for a DNA extract, while a pure RNA extract has a value of ~2.0. If the ratio drops below this threshold, it may indicate contaminants such as proteins in the extract. These absorb more to the 280nm wavelength over the 260nm wavelength

due to containing protein components such as tryrosine instead of nucleobases like adenine.

The 260/230 ratio is used to assess nucleic acid purity, a pure sample demonstrates a ratio of ~2.0. As in the case of the 260/280 ratio, contaminants in the sample will absorb more strongly at A230 over A260 and subsequently decrease the ratio.

2.1.4 Assessing DNA extract concentration

DNA extracts were quantified using the Qubit 2.0 fluorometer and BR dsDNA kit (Invitrogen). The instrument was calibrated, samples concentration read and volume of the original sample were calculated. In addition, this experiment was run in batches of twenty four samples; to help reduce the possibility of DNA sample degradation. The pending DNA samples to be tested were stored at -4 °C. While recently read samples were returned to -20 °C cryostorage. This process was carried out as described in pages 14-24 of the instrument manual available in the following URL [https://tools.thermofisher.com/content/sfs/manuals/mp32866.pdf]URL supp. data section 2.1.5

2.2 Sequencing

2.2.1 Long read SMRT cell sequencing PacBio systems model RS II

The Gentra Puregene yeast/bacteria kit and protocol (Qiagen) was used in order to produce high molecular weight (>40kb) DNA, improving sample quality for long red sequencing.

Pure DNA samples were collected until a threshold of 30µg DNA was reached. The sample (260/280: 1.83, 260/230: 1.86) was collated and sent to the Earlham Institute, Norwich for *de novo* long read sequencing using the Pacific Biosciences RSII instrument (Pandey *et al.*, 2017). Sample purity was reassessed long read sequencing was performed using the PacBio RSII. Four SMRT cells, each sequencing 50,000 8500bp length reads, were used; taking approximately three hours. A sample of the same stock was sent for Illumina paired end sequencing (Wellcome Trust Centre for Human Genetics, Oxford, Hiseq 2000, 151 bp). The paired end sequencing reads generated by the oxford genome sequencing centre was successfully used to polish the raw, closed genome sequence of any potential errors in homopolymeric regions. This is discussed in full in chapter 3.

NB At the time of writing, there was no gold standard genome available for this specific strain of *Neisseria* lactamica although it should be noted that the species *N.l*actamica does have a gold

standard genome available, *N.lactamica* strain 020-06 (Bennett *et al.*, 2010). Information regarding this existing genome assembly was be used by our third party sequencers (The Earlham Institute, Norwich) as a basis to aid successful completion of the assembly process.

2.2.2 Short read sequencing

Culture samples were defrosted and streaked onto a Columbia blood agar (Oxoid). Culturing was performed as described above. Illumina short read sequencing was carried out using short read, 151 bp performed by Hiseq 2000. Sequencing methodology was performed similar to that outlined in (Deasy *et al.*, 2015). Wild Type and longitudinally isolated *N. lactamica* samples from the long term, experimental human challenge study were extracted using the wizard gDNA kit and protocols Pure, 10µg DNA samples (260/280: 1.8+; 260/230:1.7+, good peak on nanodrop trace) of genomic DNA were sent for Hiseq 2000, 151 Bp, paired end sequencing (Wellcome Trust Centre for Human Genetics, Oxford). Samples were quantified using Qubit 2.0 and the BR assay kit (Invitrogen), libraries were constructed using the EBNext DNA sample Prep Master Mix Set 1 Kit (New England Biolabs) and illumina nextera adapters were excised.

2.2.3 Ethanol precipitation

Ethanol precipitation of DNA was performed in steps outlined in the lamitina lab protocol. this step was performed to recover DNA from high yield, low purity samples as identified by the spectrophotometer.

The volume of the DNA sample was measured. 1/10 volume of 3M sodium acetate, pH 5.2, (final concentration of 0.3 M) was added and mixed well. 2-2.5 times the volume of cold 100% ethanol (calculated after salt addition) were added and mixed well. The mixture was placed on ice or at -20 degrees C for >20 minutes before being spun at a maximum speed in a microfuge 10-15 min. The supernatant was carefully decanted. and 1 ml 70% ethanol was added. This was mixed and spun briefly in a centrifuge before the supernatant was carefully decanted. The pellet was air dried and then resuspend in the appropriate volume of distilled water.

This protocol is available from the following URL
http://docs.wixstatic.com/ugd/803ab9_1cd1cb09279649b388391953899ae1f9.pdf
Visualisation of Data

2.2.4 MView (embl-ebi)

MView (Brown, Leroy and Sander, 1998) was used to visualise and check the accuracy of .fasta format alignments. The MView program is hosted on the following URL (
The alignments for this analysis were usually generated gene by gene using MAFFT aligner (Katoh and Standley, 2013) available via the genome comparator tool available on pubMLST Neisseria (Jolley and Maiden, 2010).

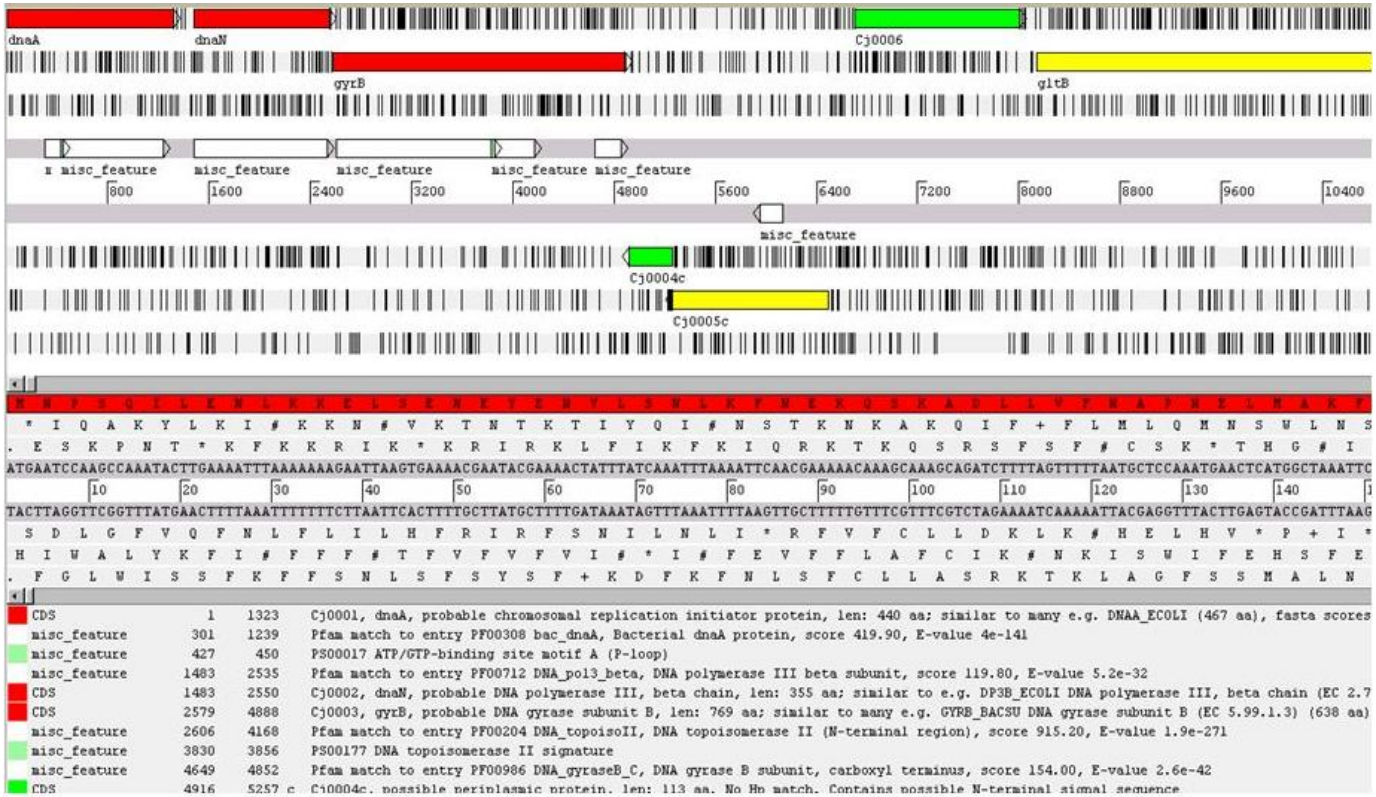
2.2.5 Figtree and other phylogenetic visualisers

Phylogenetic trees were generated in both the nexus and newick formats in this thesis.
Figtree (Rambaut, 2009) was run as a java executable available from the following download link. This software was used because it accepted tree files in the newick format and possessed a wide variety of intuitive tools for labelling phylogenetic tree images.
SplitsTree4 (Huson and Bryant, 2006) was used to visualise the neighbour net phylogenetic files (.nexus) outputted by default during analyses using the genome comparator tool on [pubmlst.org/Neisseria].

2.2.6 Artemis

Artemis ((Rutherford *et al.*, 2000), **Figure 2-1**) by the Sanger institute was used as a general all-purpose genome browser and editor. Tablet (Milne *et al.*, 2009) was used to examine read coverage in .FASTQ files. While Artemis can also be used for this purpose, Tablet allows for faster browsing. Both genome visualisers accepted files in formats such as draft genome assemblies (.fasta), raw sequencing files (.fastq), bam and sam. These visualisers can also be used to look at average coverage across sections of the genome.

Figure 2-1 Artemis Genome Browser screenshot.



Genes may be studied at both a DNA (middle, green-yellow column) and peptide (red, green, yellow) residue level

2.3 SNP calling; Breseq pipeline and other SNP callers

The process of SNP calling uses an accurate, reference sequence as a template for positionally specific, mapping of trimmed next generation sequencing reads by aligners such as BWA (Li, 2013) or bowtie2 (Langmead and Salzberg, 2012). Once created, this alignment file is binarised into BAM format and sorted, a process that is essential to downstream use in many programs/scripts. SNP callers "pile up" positionally aligned reads to reference assembly sequence and generate a BCF file containing; all genomic positions, what reads overlap the position and any indels (insertions or deletions) detected at that position in the BAM file. This rich BCF file can be parsed into the variant call format (VCF), used by many downstream applications.

2.3.1 Trimming adapter sequence from raw reads (Trimmomatic)

Trimming of raw sample read files (.fastqs) was carried using the trimmomatic program (Bolger *et al*., 2014), a comma separated value (.csv) file containing nextera adapters used in the sequencing process (2.2.2) was used to excise these contaminants from raw read sequencing data in.fastq format. The script used is available in the **Appendix section A.1.2**

2.3.2 Mutation detection and statistical analysis of mutation and phase variation

Experimental procedures for mutational analyses were performed using the Breseq pipeline (Deatherage and Barrick, 2014) under standard parameters. All mutations observed and reported were manually checked for false positives arising because of low read coverage and subsequent statistical bias. Chapters 4 and 5 made use of these techniques. Breseq was run on the University of Southampton's high computing cluster (Iridis4) and utilised a PBS script displayed in **Appendix Figure A-1.**

The CL-TABULATE command using the breseq pipeline was used to generate a table containing all homopolymeric tract lengths between 7-16 base pairs in length, their genomic location and the number of reads supporting each tract. This was repeated for every isolate screened. Genes that displayed read divergence between tract lengths were amalgamated into gene-specific tables. A phase variable tract change was identified where following

i) breseq's standard mutation Kolmogorov-Smirnov test (to test for base quality bias supporting a novel variant).

II) A statistically significant (fisher's twin-tailed exact test <0.05) number of sample reads indicated a divergence from that detected in the reference sequence.

III) High quality reads encapsulating the variable regions of contingency loci were filtered to remove any reads not spanning the entire length of the variable region.

iv) The presence of a high-quality base, before or after the variable region but matching the reference sequence was also pre-requisite.

Chapter 3: The *N. lactamica* Y92-1009 genome sequencing, assembly, annotation and insights

3.1 Introduction

While pathogenic *Neisseria* have several, complete, reference whole genomes available for analysis; *Neisseria lactamica* currently has one, *N. lactamica* 020-06. Since the *N. lactamica* Y92-1009 assembly demonstrated sufficient coverage and quality to be closed; the sequence should be compared to its closest analogue for a homology analysis of genomic regions and specific genes. *Neisseria* spp. are known for sharing highly-specific gene content (Bennett *et al.*, 2010). As such, the average homology of both the DNA and subsequently the protein content, of both genomes will be highly similar in this analysis. But the potential discovery of novel segments in the genome of *N. lactamica* Y92-1009 may shed light on what caused this strain to evolve on a different evolutionary path from a common ancestor and will inform on what genomic regions have driven the evolution of both of these nasopharyngeal dwelling, commensal organisms that compete for the same niche.

The importance of an accurate and reliable reference genome of the inoculum strain (*N. lactamica* Y92-1009*)* is tantamount to the accuracy of all subsequent mutational analyses. The only genome publicly available for this strain was shotgun sequenced and assembled into 44 contigs by Public Health England over a decade ago (Vaughan *et al.*, 2006). Each of these contigs represents a potential region of low coverage and can also occur in genomic regions containing repetitive sequences. These areas are known to be frequent targets of neisserial microevolution (Marri *et* al., 2010). Therefore, it is also essential to be able to accurately detect any changes occurring to them. Since the time this genome (PHE 2006 assembly) was sequenced, more sophisticated technologies (SMRT) have emerged and these will be used to construct a more accurate and complete reference genome. The difference between both genomes will be quantified.

Repetitive sequences play important roles in *Neisseria* genome modification and gene expression. The ten base DNA uptake sequence (DUS) has been shown to be pre-requisite for transformation in *Neisseria* spp. (Frye *et al.*, 2013). DUS-containing sequences have permeated the *Neisseria* genus core genome, indicating these sequences can survive genome diversification via recombination (Treangen *et al.*, 2008). Another repeat type is named dRS3. This is an abundantly recurring 20bp repeat sequence known to flank larger repeat sequences and act as a site for

phage integration (Rotman and Seifert, 2014). Lastly, the transposon-like Correia repeat enclosed elements (CREEs) may combine with native sequence to form gene promotors as well as affect post transcriptional gene expression (Lin, Ryan and Davies, 2011). As a result of this, CREEs have often been observed as hot spots of DNA rearrangement and recombination (Siddique, Buisine and Chalmers, 2011). Repeat sequence content can reflect on the evolutionary history of an organism. *N. meningitidis* possesses many more CREEs than any other member of the genus. This is thought to have arisen after the species diversified away from a common ancestor (Rotman and Seifert, 2014). These repeat motifs will be discovered for *N. lactamica* Y92-1009 and contrasted with results from previous *Neisseria* studies.

3.2 Methods

3.2.1 Long read sequencing and assembly

This is described fully in section 2.2.1 and (Pandey *et al.*, 2017)

3.2.2 Annotation and database searching

The Prokka pipeline (Seemann, 2014) was used to putatively assign genetic function and identify RNA and pseudogenes. As part of this annotation pipeline, prodigal (Hyatt *et al*., 2010) was initially used to identify all co-ordinates of CDSs from the assembly but did not assign a putative gene product. Once all CDSs were detected, gene prediction is normally inferred by comparing an unknown protein to a database containing known protein sequences. To ensure maximum possible accuracy, this sequence-database homology comparison is staggered hierarchically in the following way by Prokka.

i) All putative CDSs are matched with a trusted list of proteins from the only manually curated *N. lactamica* reference genome (Bennett *et al.*, 2010).

ii)All unannotated proteins are then compared to the uniprot bacterial database

iii) All unannotated proteins are then compared to a *Neisseria* specific RefSeq database (enabled with the –genus and –usegenus flags)

The genome assembly was loaded into BLAST2GO (Conesa *et al.*, 2005). This was used to search against the MMMPFAM (Finn *et al.*, 2014), SignalPHMM (Petersen *et al.*, 2011) and THHMM (Krogh *et al.*, 2001) databases to identify genes with PFAM domains, signal peptides and transmembrane helices respectively.

Any CRISPRs annotated were found using CRISPRfinder (Grissa, Vergnaud and Pourcel, 2007). The bacteriophage identification tool PHAST (Zhou *et al.*, 2011) was used to identify, annotate and display the detected prophage sequence. Insertion sequences within Prophage regions were identified using IS finder (Siguier *et al.*, 2012).

3.2.3 Protein clustering

Paralogous genes were identified using CD-HIT (Fu *et al.*, 2012) and a file of *in silico* translated proteins from the *N. lactamica* Y92-1009 genome. A sequence identity of 0.9 was used alongside a

word size of 5 for the search parameters. This was performed according to the developer's instructions available at http://www.bioinformatics.org/cd-hit/cd-hit-user-guide.pdf

3.2.4 Repeat Motif searching

Motifs for repetitive sequences were acquired from a *Neisseria* genus wide study that reported DUS and dRS3 motifs among ten species (Marri *et al.*, 2010), correia repeat type patterns (Roberts *et al.*, 2016) and a study on the overrepresentation of DUS motifs (dialects)among the *Neisseriaceae* (Frye *et al.*, 2013). These motifs were searched in the genome using fuzznuc as part of the EMBOSS package (P Rice, Longden and Bleasby, 2000) available at http://www.bioinformatics.nl/cgi-bin/emboss/fuzznuc

3.2.5 Cgview comparison Tool

The Cgview comparison tool python script (Grant, Arantes and Stothard, 2012) was used to assign COGs to *N. lactamica* Y92-1009, in addition the program generated DNA and CDS comparison maps. A blast atlas calculation and whole genome-maps were constructed as per the developers' instructions available at the URL http://stothard.afns.ualberta.ca/downloads/CCT/tutorials.html#tutorial-2.

A genbank format (.gbk) *N. lactamica* Y92-1009 annotation was used as a reference to compare both DNA and coding sequence content with *N. lactamica* 020-06 .gbk . For the DNA vs DNA comparison. Blastn (word size 11) was used to compare the total genomic region of *N. lactamica* 020-06 against $1\text{^}10^6$ base pair segments (this is due to a threshold limit of blastn query size used by the software) of reference, *N. lactamica* Y92-1009 sequence.
For the CDS vs CDS comparison BlastP was used to compare putative protein content encoded for by both genomes. Any proteins identified as being present in *N. lactamica* Y92-1009 and absent in *N. lactamica* 020-06 were protein blasted against the nr database as well as the pubMLST Neisseria locus database.

3.3 Results

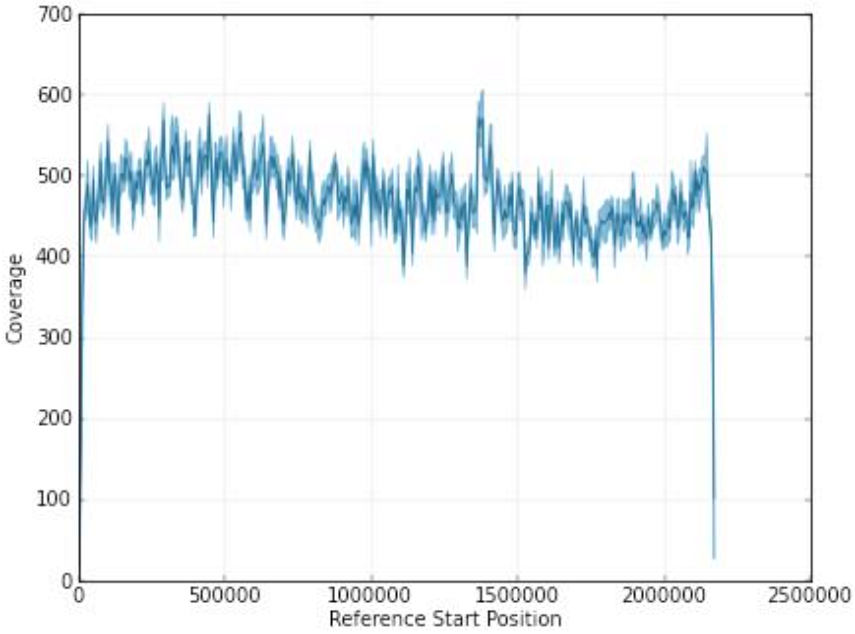
In these sections I will first describe the quality metrics and error correction of the Pac Bio RS1 sequenced genome. I will then describe the genome's annotation, COG analysis and repeat sequence make up. I will then go on to contrast the new assembly against the genomic content of *N. lactamica* 020-06, the closest related closed genome. Lastly, I will describe how much more accurate this assembly is compared to the existing *N. lactamica* Y92-1009 assembly (Vaughan *et al*., 2006).

3.3.1 The long read sequenced, day zero *N. lactamica* Y92-1009: genome assembly and quality control

The Pac Bio sequenced genome assembly consisted of 2,176,116 bases assembled into one contig. As shown in **Figure 3-1**, coverage was consistently high (~500X) over the span of the bases existing as one contiguous run of sequence. The assembly created by long read sequencing technology showed drops of coverage at both the start and finish of the genome sequence. The assembly was analysed for any low coverage and erroneous bases. This was done using the program Pilon and 151 bp sequenced (Illumina) reads from the same sample used to provide the DNA extraction for long read sequencing (**Figure 3-2**). Most of these corrections were intergenic and of no consequence to positions designated as undergoing mutations by subsequant SNP analyses. Position 256,700 located 4bp into the homopolymeric tract in the *pglA* gene was the exception. This gene is phase variable which means the addition or deletion of a base in the tract could cause a frameshift.

The 151 bp illumina reads were also mapped to the genome assembly using the Breseq pipeline V.0.26a in order to identify and trim low coverage areas present at the beginning and end of the circular genome sequence. This low coverage area was shown to span approximately 30,000 bases containing 34 putative genes (**Table 3-2).** Its removal reduced the assembly size down from 2,176,116 bp to 2,146,723 bp.

Figure 3-1 Coverage plot for SMRT cell sequenced raw genome assembly.



This shows an average value of coverage (coverage is defined the number of times evidence is shown that a given base is represented by sequencing read data) across the N. lactamica Y92-1009 genome sequence. There are two drops of coverage visible at opposite ends of the assembly

Figure 3-2 A list of erroneous positions calculated by Pilon from the Pac Bio assembly.

Position	Change	Position biasing SNP call?
87879	Del A	No, Intergenic
256700	Ins G	Yes, *pglA*
289417	Del A	No, Intergenic
389182	Del T	No, Intergenic
429284	Del T	No, Intergenic
564569	Ins C	No, Intergenic
692038	Del A	No, Intergenic
1294754	Del T	No, Intergenic
1336629	Del A	No, Intergenic
1385873	Del A	No, Intergenic
1648253	Del T	No, Intergenic
2135627	Del A	No, Intergenic

"Ins" refers to the insertion or addition of the corresponding base to that position.

"Del" refers to a deletion or subtraction.

Figure 3-3 Unassigned missing coverage evidence detected by Breseq when using the Pac Bio genome assembly as a reference and short Illumina reads from the same sample

Unassigned missing coverage evidence

			seq id	start	end	size	←reads	reads→	gene	description
*	*	÷	PROKKA_contig000001	1	16017–3774	3774–16017	NA [0]	[102] 109	*tbpB_1–[PROKKA_00018]*	***18 genes*** *tbpB_1, tbpB_2, tbpB_3, tbpB_4, PROKKA_00005, PROKKA_00006, PROKKA_00007, PROKKA_00008, PROKKA_00009, PROKKA_00010, PROKKA_00011, PROKKA_00012, PROKKA_00013, PROKKA_00014, PROKKA_00015, PROKKA_00016, tyrA_1, [PROKKA_00018]*
*	*	÷	PROKKA_contig000001	2162748–2176091	2176116	26–13369	107 [105]	[0] NA	*PROKKA_02073–nsrR_2*	***16 genes*** *PROKKA_02073, PROKKA_02074, PROKKA_02075, PROKKA_02076, PROKKA_02077, PROKKA_02078, PROKKA_02079, PROKKA_02080, PROKKA_02081, PROKKA_02082, PROKKA_02083, PROKKA_02084, tyrA_2, PROKKA_02086, PROKKA_02087, nsrR_2*

This table extrapolates upon and quantifies the two areas of low coverage shown in **Fig 3-1.** The first row shows the genes detected in the area of low coverage at the beginning of the assembly (positions 1-16,017) while the second row shows the genes detected in the area of low coverage at the end of the assembly (position 2,162,748-2,176116) The "**gene**" column does not refer to what genes have been detected in the low coverage regions. It describes that the list of genes found in that region run from the first gene (tbpB_1, in the first row; description column) up to last gene (PROKKA_00018, in the first row; description column). The full list of putative genes detected in both of these regions is in the "**Description**" column to the far right of the table. The **start** and **end** columns show what position number (out of a total of 2176091 bases) and therefore what section of the genome assembly these low read coverage regions encompass.

3.3.2 The genome annotation, COG analysis and repeat sequence content of *N. lactamica* Y92-1009

The *N. lactamica* Y92-1009 genome assembly contained 2,146,723 bp with approximately 460 fold mean coverage depth and 52.3% GC ratio. The assembly was predicted to contain 2053 putative ORFs, 1980 of which coded for proteins. There were 72 genes predicted to encode RNA genes and three CRISPR repeats were detected (**Figure 3-4**.)

Furthermore, 74.3% of total putative ORFs matched with the COG database; these results are presented in **Figure 3-5** and displayed in a circular genome diagram in **Figure 3-6**. The three most common COG categories were S (n= 163, function unknown), J (n= 148, Translation, ribosomal structure and biogenesis) & L (n=137, Replication, recombination and repair.) It should be noted that approximately a quarter of the number of coding sequences in the *N. lactamica* Y92-1009 genome were not assigned a COG (n=526).

The frequency and pattern of dRS3, CRE and DUS repeat sequence motifs observed in *N. lactamica* Y92-1009 is shown in **Figure 3-7**.

Figure 3-4 Genome annotation statistics

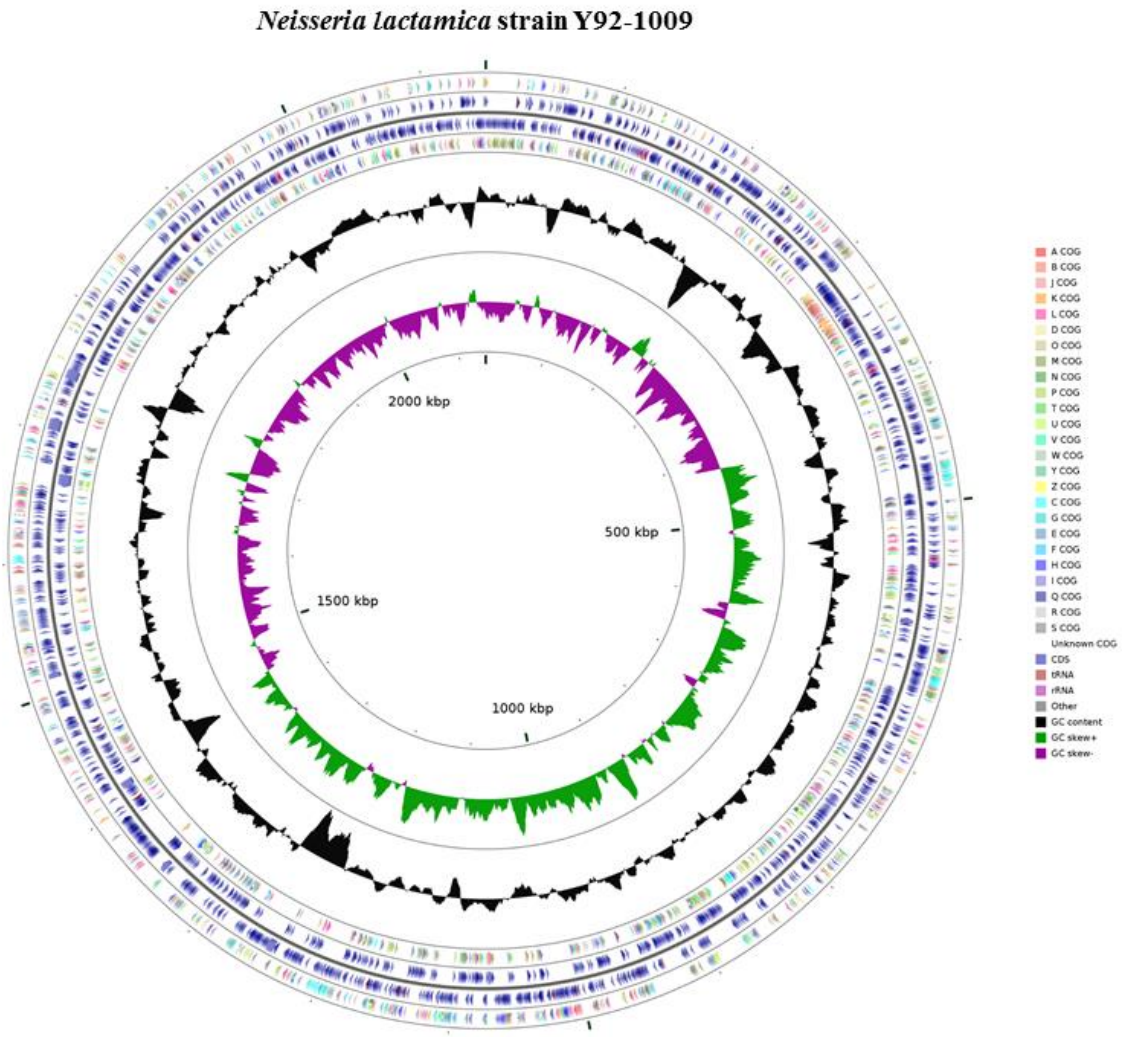
Attribute	Value	% of Total
Genome size (bp)	2146723	100
DNA coding (bp)	1831541	85.3
DNA G+C (bp)	1123594	52.3
DNA scaffolds	1	100
Total genes	2053	100
Protein coding genes	1980	96.4
RNA genes	72	3.5
Pseudo genes	16	0.8
Paralogues	16	0.8
Genes with function prediction	1918	93.4
Genes assigned to COGs	1527	74.3
Genes with Pfam domains	5	0.2
Genes with signal peptides	0	0
Genes with transmembrane helices	0	0
CRISPR repeats	3	0.1

To clarify, pseudogenes were defined as potentially defunct but contain high sequence homology with the genes they are seen to mimic. Paralogous genes were either a duplicate or near duplicate copy of a gene pre-existing in the genome. The methods for identifying pseudogenes, paralogous genes & genes containing Pfam domains, signal peptides, CRISPRs or transmembrane helices were described in the methods of this chapter.

Figure 3-5 Number of genes in the *N. lactamica* Y92-1009 genome associated with general COG functional categories.

Code	Value	%age	Description
J	148	7.21	Translation, ribosomal structure and biogenesis
A	1	0.05	RNA processing and modification
K	56	2.73	Transcription
L	137	6.67	Replication, recombination and repair
B	1	0.05	Chromatin structure and dynamics
D	24	1.17	Cell cycle control, Cell division, chromosome partitioning
V	23	1.12	Defense mechanisms
T	25	1.22	Signal transduction mechanisms
M	130	6.33	Cell wall/membrane biogenesis
N	20	0.97	Cell motility
U	42	2.05	Intracellular trafficking and secretion
O	75	3.65	Posttranslational modification, protein turnover, chaperones
C	109	5.31	Energy production and conversion
G	48	2.34	Carbohydrate transport and metabolism
E	129	6.28	Amino acid transport and metabolism
F	45	2.19	Nucleotide transport and metabolism
H	76	3.70	Coenzyme transport and metabolism
I	51	2.48	Lipid transport and metabolism
P	77	3.75	Inorganic ion transport and metabolism
Q	10	0.49	Secondary metabolites biosynthesis, transport and catabolism
R	137	6.67	General function prediction only
S	163	7.94	Function unknown
-	526	25.62	Not in COGs

Figure 3-6 Clusters of Orthologous Groups represented on the *N. lactamica* Y92-1009 genome



The circular genome map was generated with Cgview Comparison Tool. The first (counted from outermost to innermost) ring contains genes identified and assigned to Clusters of Orthologous Groups (COGs). The twenty COG categories correspond to a spectrum of colour from Red to Grey. These are annotated in the right-side figure legend and ran from 5' to 3' (+). The second and third rings display regions containing coding sequence (Blue), tRNA (Orange), and other RNAs (Grey), with the second ring running 5' to 3' (+) and the third ring runs 3' to 5 (-)'. The fourth ring contains open reading frames assigned as COGs and encoded on the negative strand. The fifth, black, graph ring displays GC content while the last ring (purple and green)displays positive (Green) and negative (Purple) GC skew. The alphabetical COG categories can be referenced in **Table 3-4**.

Figure 3-7 Frequency of repeat sequences in N. lactamica Y92-1009 genome and comparison of results with members of the *Neisseriaceae*.

Strain	AT-DUS	AG-DUS	AG-mucDUS	DRS3	Correia Type 1	Correia Type 2	Correia Type 3	Correia Type 4
N. lactamica Y92-1009	1718	262	45	454	50	1	17	17
N. lactamica 020-06	838	10	32	253	29	0	11	10
N. meningitidis MC58	1935		N/A	689	524	N/A	N/A	N/A
N. gonorrhoeae FA1090	1965		109	208	254	N/A	N/A	N/A
N. cinerea 14685	943		113	5	28	N/A	N/A	N/A
N. polysaccharea 43768	2183		142	153	159	N/A	N/A	N/A
N. mucosa 25996	179		1543	63	210	N/A	N/A	N/A
N. sicca 29256	300		3729	35	570	N/A	N/A	N/A

For the genome of interest *N. lactamica* Y92-1009 and the related genome *N. lactamica* 020-06 results were obtained by searching the following repeat motifs using fuzznuc: AT-DNA-Uptake-Sequence [DUS](“AT-GCCGTCTGAA”), AG-DUS (“AG-GCCGTCTGAA”), AG mucDUS (“AG-GTCGTCTGAA”), dRS3 (“ATTCCCNNNNNNNNGGGAAT”), Correia Repeat element types 1, 2,3 and 4 respectively (“ATAG[CT]GGATTAACAAAAATCAGGAC”, “TATAG[CT]GGATTAAATTTAAACCGGTAC”, “TATAG[CT]GGATTAACAAAAACCGGTAC”, “TATAG[CT]GGATTAAATTTAAATCAGGAC”). Data on

repeat values for other bacteria was collated from [Figure 2 (Marri *et al.*, 2010)] and [Supplemental table 1 (Frye *et al.*, 2013)]

3.3.3 DNA and translated coding protein comparison between *N. lactamica* Y92-1009 and *N. lactamica* 020-06 reveals 10 regions unique to *N. lactamica* Y92-1009

The circular diagrams generated by Cgview comparison tool were used in conjunction with underpinning blastn and blastp matches to create quantitative data on what genomic regions and specific genes were present in *N. lactamica* Y92-1009 and absent in *N. lactamica* 020-06. The data present in the diagrams are the result of whole genome or protein sequence alignments between the reference and comparator strain. The success of these DNA or protein alignments was measured in **BLASTn** or **BLASTp identity**. This was defined as the extent to which two (nucleotide or amino acid) sequences had the same residues at the same positions in an alignment often expressed as a percentage (Fassler & Cooper, 2011).

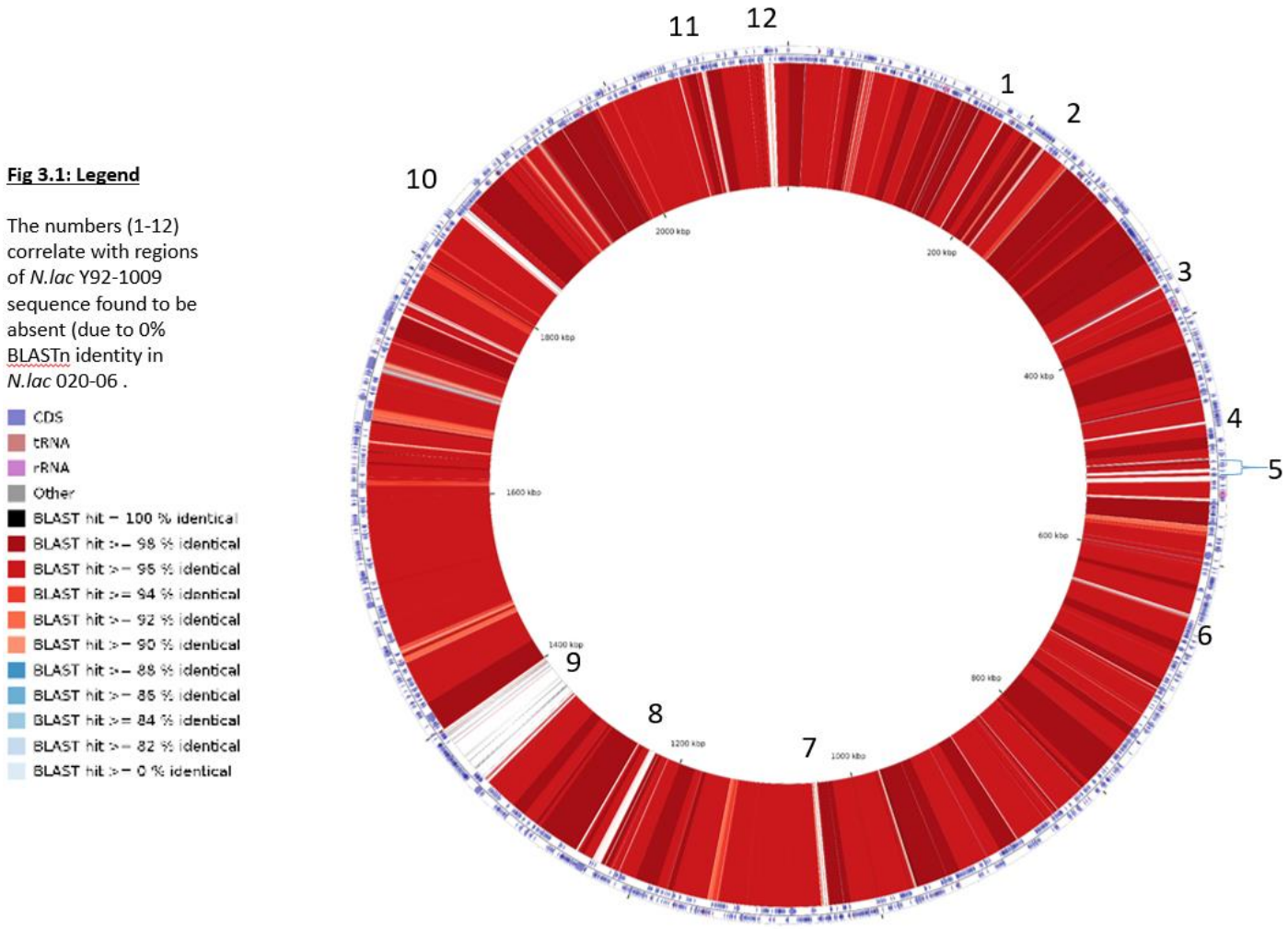
The DNA vs DNA diagram (**Figure 3-8**) revealed an abundance of red or crimson areas bisected with thin white regions. The red and crimson areas contain high BLASTn identity (>94%) and revealed that much of the genome of the newly sequenced *N. lactamica* Y92-1009 could also be found in that of *N. lactamica* 020-06. There was very little sequence data that was matched at a lower homology than 92%. A total of 13 genomic regions were found present in *N. lactamica* Y92-1009 that displayed zero homology to any DNA sequence found in the *N. lactamica* 020-06 assembly. The largest of these regions occurred at the 1,300,534bp -1,399,095 bases, was ~50kbp in size. In total, 2,220,606 bp from the *N. lactamica* 020-06 assembly were compared against the 2,146,723 bp constituting the *N. lactamica* Y92-1009 assembly.

The coding content of both genomes was compared in a separate analysis using translated protein sequences. Like the DNA vs DNA analysis, the CDS vs CDS comparison revealed that the majority of proteins shared high (>90%) but not exact blast identity with those found in *N. lactamica* 020-06

There was also evidence of regions of zero identity correlating with proteins belonging to non-homologous genomic regions identified in the DNA vs DNA analysis. Across 9 regions, 84 proteins were detected in *N. lactamica* Y92-1009, which shared no homology with any found in *N. lactamica* 020-06. For the most part the predicted proteins were annotated as conserved hypothetical proteins detected across *Neisseria spp.*

Region C contained a putative transcriptional accessory protein (alias: NMB0075). This is a highly conserved protein among *N. meningitidis* and it has been included in the pubMLST Neisseria meningococcal core genome scheme. Region D contained a restriction endonuclease protein with a 98% identity match to gonococcal restriction endonuclease (accession: HMSC056A03). In region E a LOS subunit modifier protein *lgtC* was found. Region G held the rest of the functionally annotated proteins and in discussed in greater detail in 3.3.4.

Figure 3-8: DNA vs DNA whole genome-BLASTn comparison of *N. lactamica* Y92-1009 and *N. lactamica* 020-06

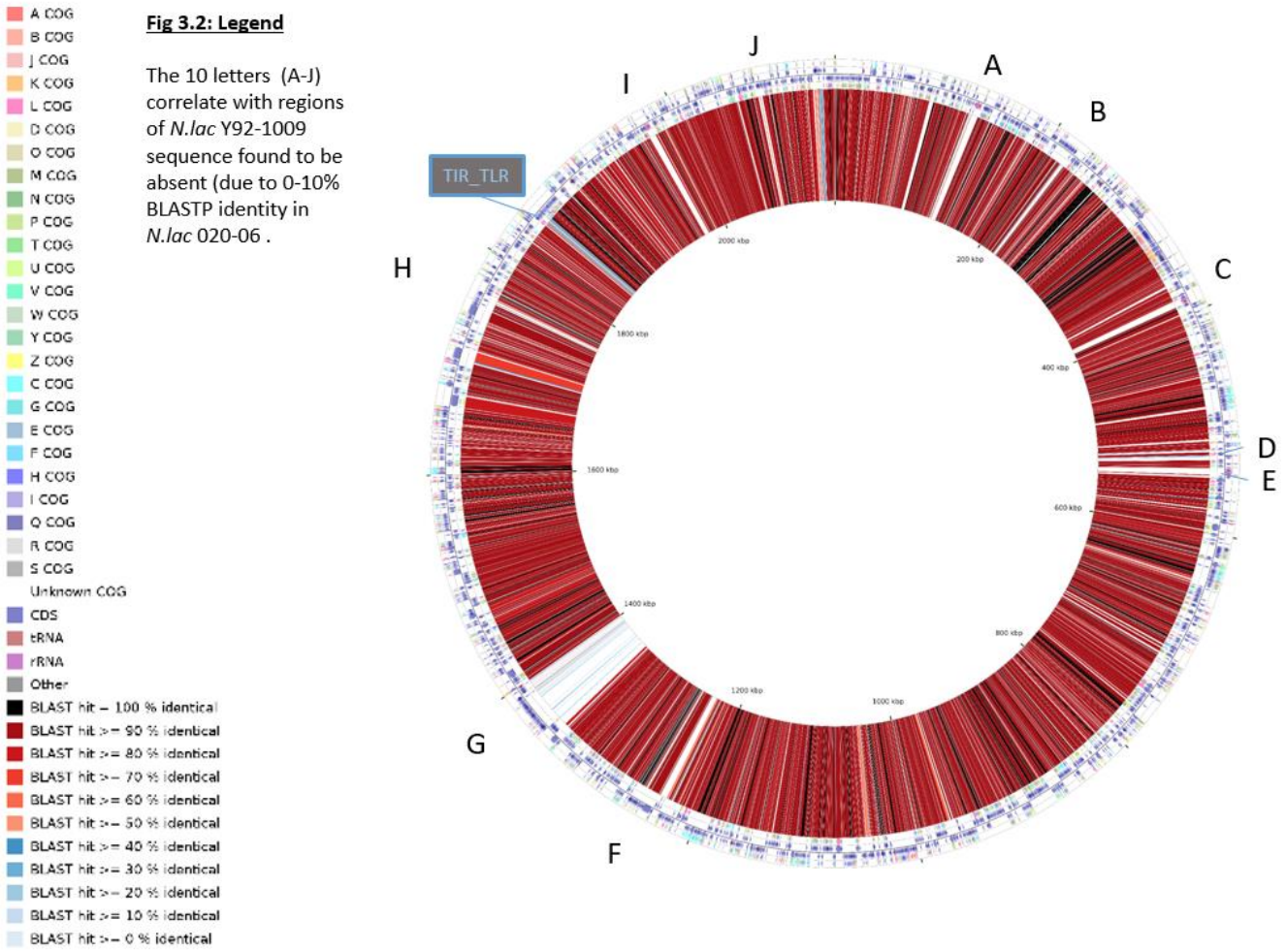


Caption Text for Fig 3.3

The red and crimson regions are areas of high BLASTn identity (>94%) and are found in between the white, low-homology numbered regions. Using the legend on the left, these red regions reveal that most of the *N. lactamica* Y92-1009 genome can be found to an extremely high (94%, red) but not exact (100%, black) BLASTn identity in the *N. lactamica* 020-06 genome. There was very little sequence data that was matched between the 1%- 92% BLASTn identity range (blue sequences).

13 genomic regions were found present in *N. lactamica* Y92-1009 that displayed zero identity to any DNA sequence found in the *N. lactamica* 020-06 genome. The largest of these regions (Region number 9, correlating with **Region G in Figure 3.4**) occurred just before the 1,400 kilo base pair mark on the figure.

Figure 3-9 CDS vs CDS whole genome-BLASTp identity comparison of *N. lactamica* Y92-1009 (reference)and *N. lactamica* 020-06 (comparator)genomes



Caption Text for Fig 3.4

The red, crimson and black regions were areas of high BLASTp identity (>80%) and were found in between ten regions lettered A- J. Using the legend on the left, these regions revealed that most of the genome of the newly sequenced *N. lactamica* Y92-1009 genome could also be found at a high but not perfect identity as that of the *N. lactamica* 020-06 genome. 10 regions were found present in *N. lactamica* Y92-1009 genome that displayed zero homology to any protein content found in the *N. lactamica* 020-06 assembly. The largest of these regions occurred just before the 1,400 kilo base pair mark (G, correlating with **region 9 Fig. 3.3**.) This region contained 72 proteins and includes a prophage sequence described in section 3.3.4. Excepting a TIR domain toll like receptor ((labelled TIR_TLR: alias pfam13676) between letters H and I) There was very little coding/protein content that was matched between 1%- 92% (I.e. blue sequences).

3.3.4 An intact prophage is detected in *N. lactamica* Y92-1009.

The largest region found in both DNA vs DNA and CDS vs CDS cgview analyses respectively is found approximately 1,350,000 bp into the genome. This 48.6 kbp stretch of DNA sequence includes 81 putative ORFs. These ORFs include a RecT like protein, two endonucleases Rus and HNH and bacteriophage subunit genes (head, tail, and terminase). Overall, the majority of proteins found in this entire region are either *Neisseria*-conserved hypothetical proteins or hypothetical proteins with a high identity but low query coverage to phage associated proteins.

Since a number of phage related genes and atypical GC content were seen in a region unique to the *N. lactamica* Y92-1009 genome, the presence of a prophage was investigated by PHAST (Phage Search Tool; http://phast.wishartlab.com/index.html). The circular genome image generated by PHAST revealed the presence of an intact prophage in the same location as region 9/G (**Figure 3-10**). The prophage is 49.8Kb in length, contains 53.98% GC content and possesses 81 proteins, 46 of which are phage associated, while 26 are hypothetical. The phage related proteins include 2 attachment sites, 2 coat proteins, 2 tail-fiber proteins, an integrase, 2 plate, a portal, a tail shaft and terminase subunits (**Figure 3-6)**. The prophage sequence scored a completeness score of 120, where 150 is the maximum and a minimum score of 90 indicates intactness of prophage. A putative attachment site was detected for the prophage sequence.

The 48.6 KBp phage sequence was subject to a nucleotide BLAST against the nr database. The paucity of results with regard to query coverage (<22%) and identity (<92%) indicated that this sequence was not conserved among other members of the genus *Neisseria.* The sequence was then subject to nucleotide BLAST against all *N. lactamica* isolates hosted on PubMLST Neisseria. This revealed that the prophage sequence was only present in other *N. lactamica* Y92-1009 isolates (including the 2006 PHE *N. lactamica* Y92-1009 shotgun sequenced assembly) suggesting it's presence was strain-specific.

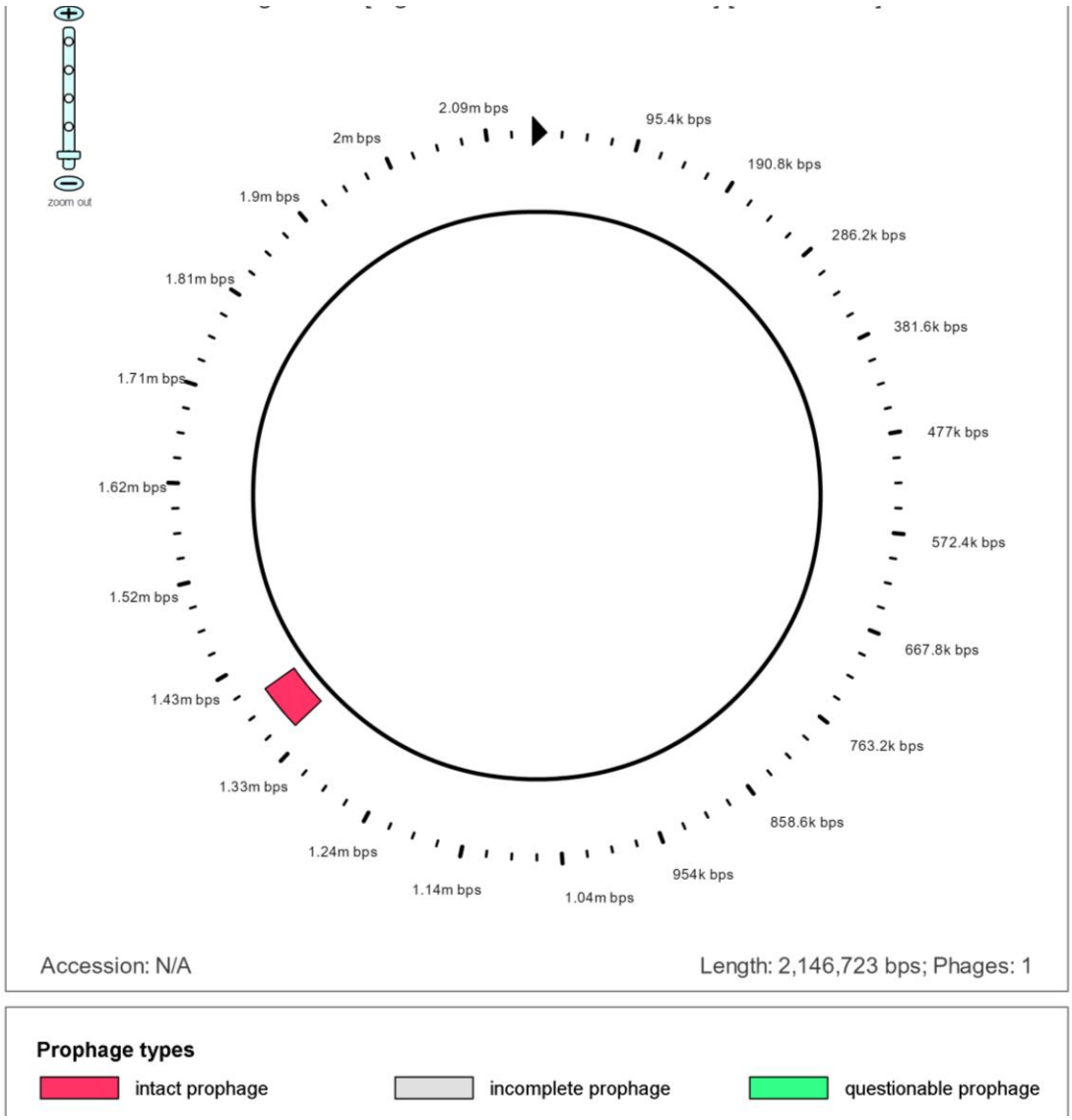
The proteins from the putative, *N. lactamica* Y92-1009 prophage sequence were compared against the PHAST-prophage and nr databases. This revealed that at least one protein in this prophage sequence was detected in over 23 different species of bacteriophage. Three of the four bacteriophages with the greatest number of shared proteins with the *N. lactamica* Y92-1009 prophage were found in *Acinetobacter* phages. However, these prophages shared 18, 17 and 16 proteins respectively out of a potential 81.

The phage sequence was analyzed for the presence of insertion sequences using ISfinder. A total of three insertion sequences were detected using the stringent criteria of no gapped matches,

100% sequence identity and <1.0 E-value. These included the ISStin4 sequence (IS30 family), ISAli17 sequence (IS5 family) and ISMha4 sequence (IS1595 family)

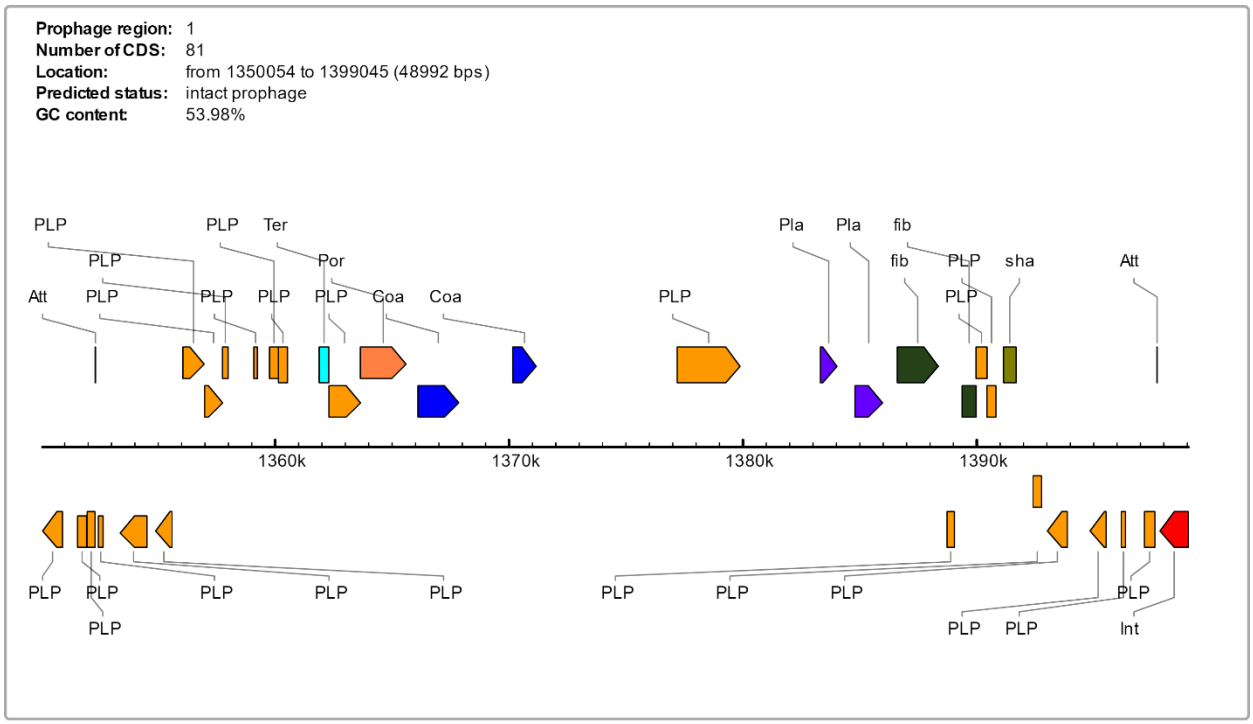
The flanking regions 2000bp upstream and downstream of the phage sequence were analyzed for protein content and function by protein BLAST. The 2000bp of sequence upstream of the identified phage sequence was found to contain four proteins. These included an AlpA family (DNA-binding transcriptional activators) phage regulatory protein (proteome number: 01331, accession: ARB04752) , found disseminated across only other *N.* meningitis and *N. gonorrhoeae* strains, two hypothetical proteins (proteome numbers 01332 & 01334) found in multiple *Neisseria* species and a flagellar formation protein found only in *N. lactamica* Y92-1009 (proteome number: 01333, accession: ARB04754) and a The 2000 bp of sequence downstream of the phage sequence was found to contain three tRNA-serine coding proteins (proteome numbers 01408-01410) and ribonuclease R (proteome number: 01411; accession: ARB04813.)

Figure 3-10 Circular image of the location of the putative prophage sequence



An intact prophage sequence was detected (pink region) by PHAST in the same location as the large region identified as unique to *N. lactamica* Y92-1009 compared to *N. lactamica* 020-06

Figure 3-6 Phage related proteins identified in the intact prophage



Identified CDS types:

1 Lysis	2 Terminase	3 Portal
4 Protease	5 Coat	6 Tail shaft
7 Attachment site	8 Intergrase	9 Other phage-like protein
10 Hypothetical protein	11 Other	12 Transposase
13 Tail fiber	14 Plate	15 tRNA

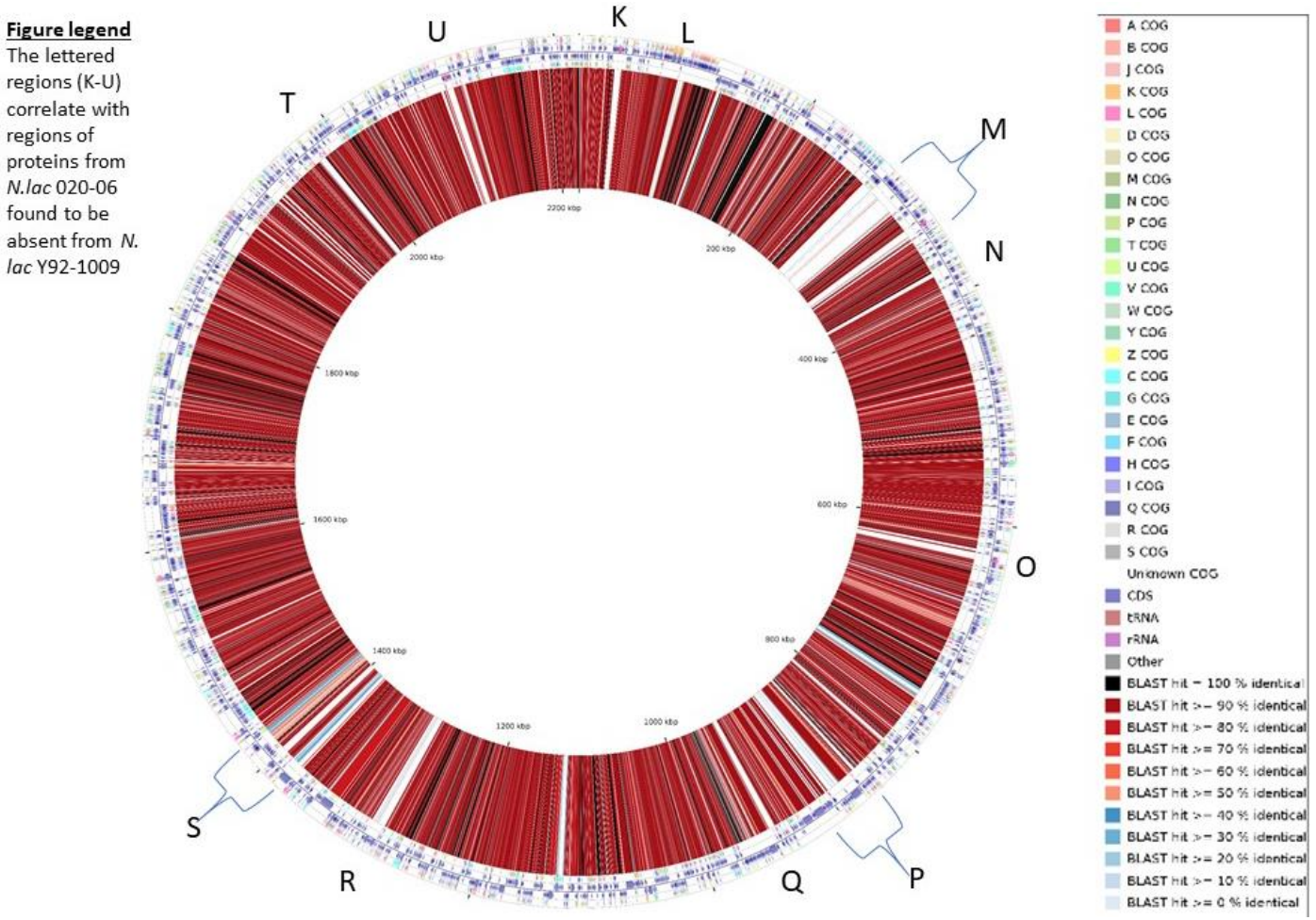
Att (phage attachment site), **Coa** (Phage coat protein), **fib** (Phage Tail Fiber), **Int** (Phage integrase) **Pla** (Phage plate protein), **PLP** (Phage like protein), **Por** (Portal protein) **sha** (Phage tail shaft protein) **Ter** (Terminase)

The diagram above was annotated by and imported from the PHAST-prophage database. 26 Hypothetical proteins were removed from the schematic to increase the clarity of the phage related proteins. The proteins above the base pair line are encoded 5’ to 3’ while the proteins under it are encoded 3’ to 5’.

3.3.5 Translated coding protein comparison between *N. lactamica* 020-06 and *N. lactamica* Y92-1009 and reveals regions absent from *N. lactamica* Y92-1009

Using the same method outlined earlier in the chapter, the role of reference and comparison genome were switched in cgview comparison tool to allow the analysis of proteins/protein regions present in *N. lactamica* 020-06 but absent from the *N. lactamica* Y92-1009 genome. Region K contained three hypothetical proteins (hps) locustagged in the *N. lactamica* 020-06 genbank file as 020_06_00083, 00084 & 00093. Region L contained a multispecies enterochelin ABC transporter (020_06_00130), a kinase protein (00133), a glycosyl transferase containing a Von Willebrand factor domain (00134) and a hp (00132). Region M was the largest region present in *N. lactamica* 020-06 which was absent in *N. lactamica* Y92-1009 and contained an intact prophage sequence which is detailed further in **Figure 3-12** . Region N contained a hp (020_06_00559) and a hp conserved among the *neisseriaceae* (00558). Region O contained two hps (020_06_637 & 00646). Region P was the second largest region found in *N. lactamica* 020-06 but absent from *N. lactamica* Y92-1009 and contained an incomplete prophage detailed further in **Figure 3-12**. Region Q contained 4 hps (020_06_00911-00914) as well as an ATPase protein (00915). Region R contained 5 hps (020_06_01223, 01224, 01225, 01227 & 01229) and the RelE translation repressor protein (01231). Region S contained a putative outer membrane peptidase (020_06_01302), a membrane protein (01319), a hp conserved among the *neisseriaceae* (01300) and five other hps (01301, 01303, 01315, 01321 & 01325). Region T contained a DNA-binding protein (020_06_01870), two hps conserved among the *neisseriaceae* (01868 & 01869) and two other hps (01867 & 01871). Lastly, region U contained 4 hps (020_06_1989-2002)

Figure 3-11 CDS vs CDS whole genome-BLASTp identity comparison of *N. lactamica* 020-06 against *N. lactamica* Y92-1009 genome



Caption for figure 3-6

The red, crimson and black regions were areas of high BLASTp identity (>80%) and were found in between eleven regions of zero to low (<20%) BLASTp identity lettered K- U. Using the legend on the left, these regions revealed that most of the genome of the *N. lactamica* 020-06 genome could also be found at a high but not perfect identity as that of the *N. lactamica* Y92-1009 genome.

.

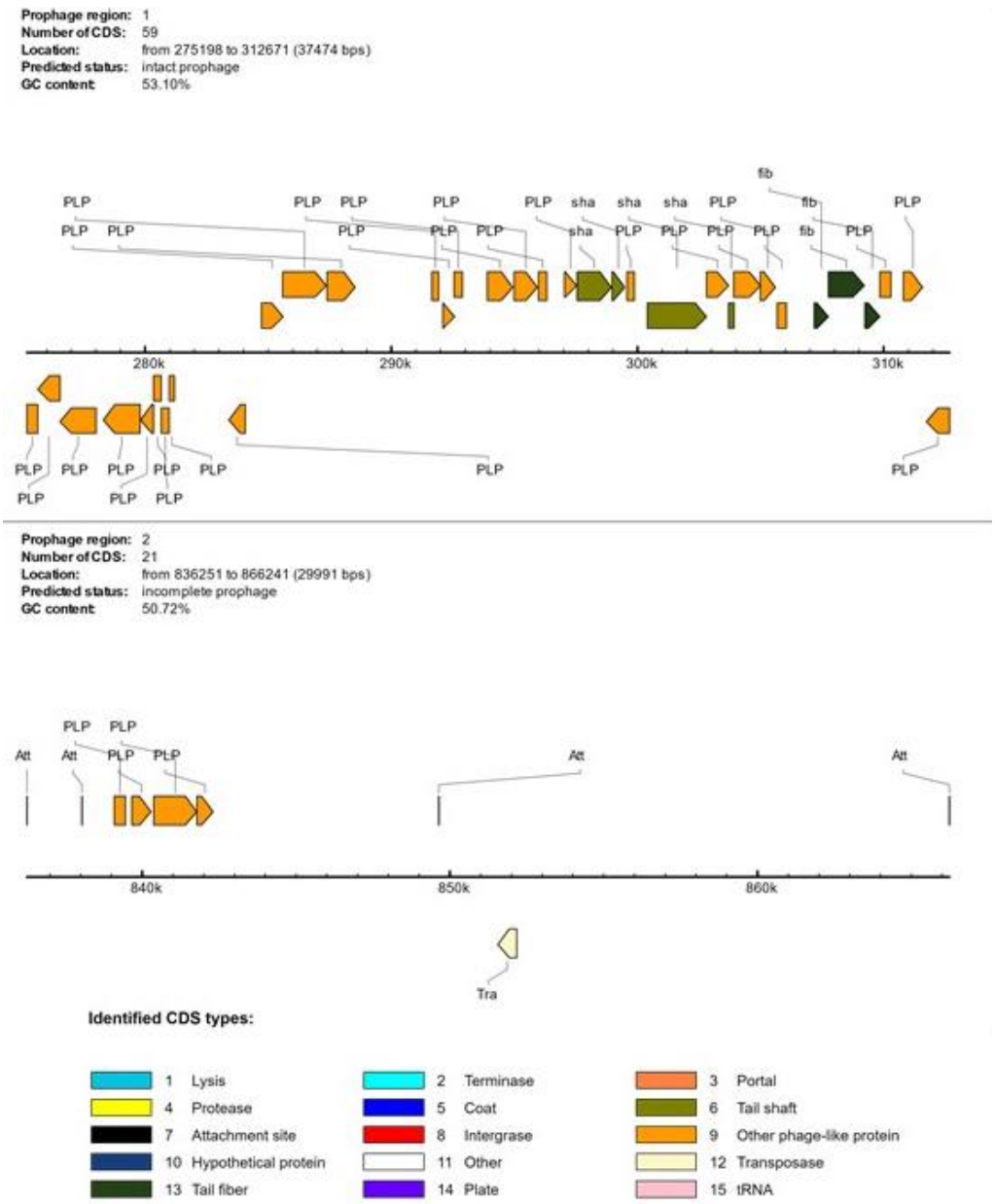
3.3.6 Two prophages detected in the *N. lactamica* 020-06 genome

The two largest regions (regions M and P, **Figure 3-11**) seen in *N. lactamica* 020-06 but absent from *N. lactamica* Y92-1009 were found to contain phage related proteins and were therefore also entered into PHAST to identify prophage regions.

The complete prophage discovered in region M was approximately 37.4Kbp long and spanned positions 275,198-312,671bp in the genome (**Figure 3-12**). The region possessed a GC content of 53.10% similar to the 52.27% calculated for the *N. lactamica* 020-06 whole genome. A total of 59 proteins were identified in this intact prophage, 39 of which were phage associated, 15 of which were hypothetical and 5 of which were bacterial proteins. The phage related proteins included three tail shaft and three tail fiber proteins but no RNA coding genes or attachment sites. This phage sequence was compared against the PHAST database which found the highest matches based on protein constituency to occur among phage phiE255 (accession: NC_009237, score= 24/59 proteins) and BcepMu (NC_005882, score = 23/59 proteins) isolated from *Burkholderia spp.* The 2000bp upstream flanking region of the phage sequence contained a phosphoenolpyruvate protein phosphotransferase (proteome number: 020_06_00293, accession: CBN86522, alias: NMB2044) which is also found in *N. meningitidis* and *N. gonorrhoeae.* The 2000bp downstream flanking region of the phage sequence contained three proteins. The first of these was a hypothetical protein containing a nudix (nucleoside diphosphate hydrolysis) domain (proteome number: 00353, accession: CBN86582, alias: NMB2041) found conserved among the *Neisseriaceae*. The second was a tRNA-met protein (proteome number: 00354) and lastly there was a thiamin biosynthesis protein ThiC (proteome number: 00355, accession: CBN86583, alias: NMB2040) which is also conserved among the *Neisseriaceae*.

Region P contained an incomplete prophage approximately 29.9 Kbp in length and spanned positions 836, 251 - 866, 230bp in the genome (**Figure 3-12**). This region possessed a GC content of 50.72% and contained a total of 17 proteins. These included 6 phage associated proteins including integrase and transposase proteins, 7 hypothetical proteins and 4 bacterial proteins. While no RNA coding genes were identified in this region but four attachment sites (2 X *attL* & 2 X *attR*) were identified. The incomplete phage sequence was compared against the PHAST database but yielded no significant matches.

Figure 3-12 Phage related proteins identified region C (prophage region 1) and region F (prophage region 2) in the *N. lactamica* 020-06 genome.

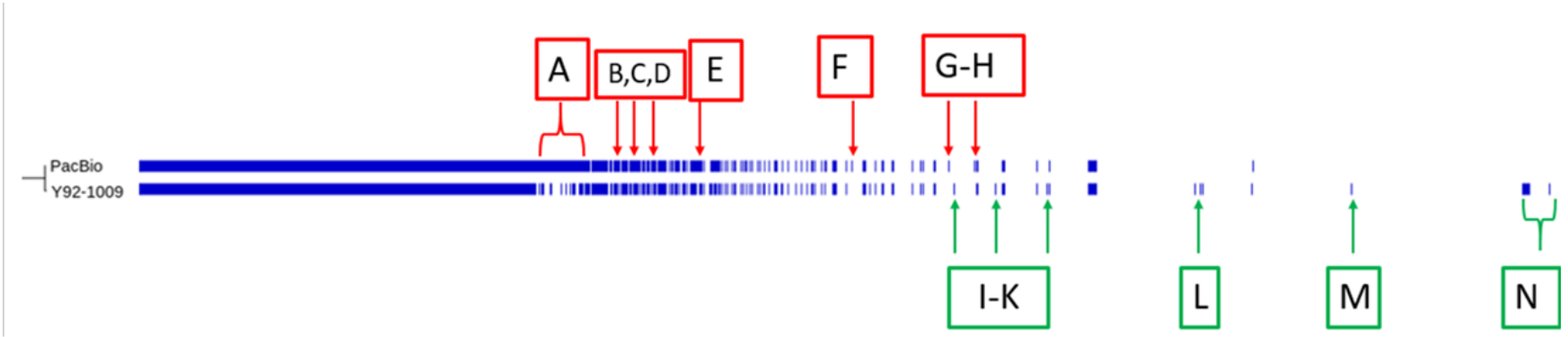


Att (phage attachment site), **fib** (Phage Tail Fiber), **Int** (Phage integrase) **Pla** (Phage plate protein), **PLP** (Phage like protein, **sha** (Phage tail shaft protein) **Tra** (Transposase)The diagram above was annotated by and imported from the PHAST-prophage database. Hypothetical proteins were removed from the schematic to increase the clarity of the phage related proteins. The proteins above the base pair lines are encoded 5’ to 3’ while the proteins under it are encoded 3’ to 5’.

3.3.7 Pan –genomic comparison of gene presence or absence between the PHE 2006 *N. lactamica* Y92-1009 assembly against the Pac Bio assembly revealed a host of quality control issues associated with the prior.

In order to help illustrate the difference in quality between them. Roary, the pan-genomic pipeline (**see chapter 7**) was used to calculate the extent of genetic diversity across the species, *N. lactamica*. Once calculated, the pan genome was used to compare the gene presence or absence of the shotgun sequenced, earlier *N. lactamica* Y92-1009 assembly (Vaughan *et al*., 2006) against the newly long read sequenced Pac Bio assembly. The earlier assembly consists of 44 contigs, the ends of beginnings of each being a potential source of sequence error. The PacBio sequenced genome assembly is closed and has furthermore been error corrected with a short read sequence of an identical isolate. The pan genomic comparison revealed that the earlier PHE assembly had 64 novel genes/gene variants within 5 regions in its assembly but 15 of these were flagged up by kraken to contain quality control issues **(See Appendix Tables 0-1 and 0-2**.) The Pac Bio corrected assembly had 154 novel genes/ gene variants in comparison. These were seen across 8 distinct regions (**Figure 3-13**) showcased by the phandango visualizer (Hadfield 2016.) Despite these differences, there is a high degree of homology between these assemblies. The gene presence and absence analysis also revealed that there were a total 1816 identical genes shared between them.

Figure 3-13 A *N. lactamica* Y92-1009 specific blow up of a section of the pan genome overview diagram generated by phandango (Hadfield 2016.)



Presence and of absence of genome segments contrasting the PacBio rsII long read reference and Y92-1009 shotgun sequence (PHE, Vaughan *et* al., 2006) as determined by pan genomic analysis using the pan-lactamica dataset and Roary v3.4.1. Red regions (A-H) represent areas where genes are present in the Pac Bio assembly and missing in the earlier *N. lactamica* Y92-1009 assembly. Conversely, green regions (I-N) represent areas where genes are present in the earlier *N. lactamica* assembly but not in the Pac Bio assembly. The raw data (specific gene presence and absence) in this analysis is presented in **Appendix Tables 0-1, 0-2, 0-3 and 0-4**..

3.4 Discussion

A variety of phages have been described in *Neisseria* species to date. In addition to a *E.coli* Mu-like phage coding for surface expressed antigens (Masignani *et al.*, 2001), four types of filamentous prophages were discovered among *N. meningitidis* and *N. gonorrhoeae* isolates (Kawai, Uchiyama and Kobayashi, 2005). This was shortly followed by reports of the presence of a 8Kb filamentous prophage (MDAΦ) which correlated among meningococci belonging to hyperinvasive lineages (Bille *et al.*, 2005).However the exact mechanism of this enhanced bacterial invasiveness remains unknown. Five dsDNA prophage sequences were discovered in *N. gonnorhoeae.* The repressor proteins from these regions were found to regulate the expression of other neisserial genes (Piekarowicz *et al.*, 2007). In comparing the genome of *N. lactamica* Y92-1009 against *N. lactamica* 020-06 and vice versa, the largest areas of zero homology were due to the presence of integrated prophages. *N. lactamica* 020-06 harbored two prophage regions, one of these was incomplete and the complete prophage presented with a smaller number and diversity of phage-related genes, lower PHAST completeness score, smaller difference in GC content and no attachment sites compared to the *N. lactamica* Y92-1009 prophage. This may indicate a more recent transpositional event in *N. lactamica* Y92-1009 as a phage sequence adapts to a more integrated prophage state over time (Canchaya *et al.*, 2003). Both intact prophages were flanked by host tRNA coding genes as RNA sequences are more slowly evolving and offer a more “stabile habitat” for phage integration (Bobay, Rocha and Touchon, 2013).

Prophages have been known to confer a variety of genes to their host organism in other bacteria although the small size of the *N. lactamica* Y92-1009 accessory genome (see **chapter 7**) suggests this is not the case here. This prophage is unique to *N. lactamica* Y92-1009 as firstly the sequence was only found to be strain specific when queried against other *N. lactamica* isolates in PubMLST Neisseria via BLASTn and secondly, the RecT like protein encoded by the phage was unique to *N. lactamica* Y92-1009 following pan-genomic comparison against all other *N. lactamica* sequence types. A paucity of matches against other prophages on the PHAST database indicate that this sequence is currently not well characterized. The sequence also contains four transcriptional regulators, a repressor and KilA anti-repressor system suggesting a tight level of regulation of phage mediated genes.

The presence of a prophage in a nasopharyngeal coloniser like *N. lactamica* Y92-1009 may confer several advantages *in situ*. These could include protection from superinfection from lytic phages, a pro-colonization effect on gene regulation as evidenced in *N. gonnorhoeae* (Piekarowicz *et al.*, 2007) but the most compelling argument comes from a recent study that shows that phage

particles enhance meningococcal colonization onto epithelial cells (Bille *et al.*, 2017). MDA phage particles were found to form networks of large bundles linking non piliated meningococci to piliated meningococci, increasing biomass. This is a strategy that would also work for an acapsulate *Neisseria lactamica.*

Motifs for repetitive sequences were acquired from a *Neisseria* genus wide study that reported DUS, CRE and dRS3 motifs among ten species (Marri *et al.*, 2010) and a study on the overrepresentation of DUS motifs (called dialects) among the *Neisseriaceae* (Frye *et al.*, 2013). In comparison to isolates examined in the first study, *N. lactamica* Y92-1009 has a higher number (n=454) of duplicated repeat sequence 3 (dRS3) than all other neisserial species bar *N. meningitidis* MC58 (n=689). This value also exceeds those detected in other *N. lactamica* strains ATCC 23970 (n= 197) 020-06 (n=253) and N. gonnorhoeae FA 1090 (n=208). The dRS3 belong to a family of Neisserial intergenic mosaic elements (NIME) (Parkhill *et al.*, 2000) and are the widely studied sites of phage integration into the neisserial genomes (Power and Moxon, 2006) most notably the self-transposing neisserial filamentous phage (Kawai, Uchiyama and Kobayashi, 2005) and meningococcal disease island (Bille *et al.*, 2005). In addition to this, the presence of dRS3 elements flanking genes are hypothesized to contribute to genome maintenance and regulation in the transferrin binding operon (*tbpAB*) of *N. gonorrhoeae* FA1090 (Vélez Acevedo *et al.*, 2014) and the deletion of the *nalP* autotransporter in *N. meningitidis* MC58 was shown to be a result of a lack of them (Oldfield *et al.,* 2013). The presence of additional dRS3 in *N. lactamica* Y92-1009 may therefore serve a regulatory function for the genetic repertoire it possesses, in addition to serving as markers for phage integration.

N. lactamica Y92-1009 has lower numbers of Correia repeat enclosed elements (CREE) repeats (n= 86) than all other *Neisseria* except for *N. lactamica* 020-06 and *Neisseria cinerea*. CREEs consist of two inverted repeats flanking a variably-large, core sequence. The composition of these repeats (CR) can act as either stronger or weaker transcriptional promoters depending on whether they possess a i) -10 box and utilize native, upstream sequence to form a -35 box ((as found originally in *dcw* transcription in *N. lactamica* (Snyder, Shafer and Saunders, 2003)) or already possess the II) -35 box as found in *uvrB* transcription in *N. gonnorhoeae* (Black, Fyfe and Davies, 1995). CREE can also regulate gene expression post-transcriptionally (Rouquette-Loughlin *et al.*, 2004). The numbers of CREE are variable among specific strains of pathogenic *Neisseria* (Liu *et al.*, 2002) but are typically more common among this group than the commensal *Neisseria* (Marri *et al.*, 2010) which include *N. lactamica.*

In comparison to isolates examined in the Neisseria-wide DNA uptake sequence study, *N. lactamica* Y92-1009 possesses an overrepresentation of AT-DUS sequence (n=1718) which is

typical of other *N. lactamica* strains although the number seen is more than double that of *N. lactamica* 020-06 and more similar to the levels found in more pathogenically associated *Neisseria*. It also possesses small levels of AG-DUS and AG-mucDUS dialects which are found more prominently in species such *as Neisseria polysaccherea, N. cinerea* and *Neisseria. mucosa.* Due to the interspecific barrier to transformation that exists between *Neisseria spp.* with uncomplimentary DUS dialects (Treangen *et al.*, 2008), it unlikely that *N. lactamica* would engage in transformation with *Neisseria* spp. possessing a differing DUS dialect (Frye *et al.*, 2013).

There is still a lot of genetic information that is not known about *N. lactamica* and is yet to be gleaned. The COG analysis revealed that approximately a third (33.5%) of all open reading frames (ORFs) in this genome have either unknown function or do not possess sufficient homology to any sequence in the COG database to be sorted into a COG category. Additionally, there were only five matches when comparing *N. lactamica* Y92-1009 translated ORFs against the PFAM database as well as a total lack of matches with SignalP and THHMM databases. The use of signal peptide (Hiller *et al.*, 2004) and transmembrane helix (Cuthbertson, Doyle and Sansom, 2005) annotators instead of database matching may rectify this.

Preliminary SNP analysis (data not shown due to inaccuracy) was initially undertaken using an Illumina short read, day zero, sequenced reference assembly containing multiple contigs. This was initially done due to the absence of a reference genome for *N. lactamica* Y92-1009 used by our study. While the average coverage depth of the reference assembly was between 25-30X, this value was seen to decrease around some predicted mutations, decreasing confidence in false positive results. The SMRT cell sequenced, closed reference had a much greater sequencing depth both overall and at specific sites. The use of this PacBio RS II generated reference genome has allowed unprecedented depth into understanding the microevolution of this commensal *Neisseria* spp. over time. The post assembly correction tool Pilon found 12 false insertion mutations and rectified them. Of these twelve, only one was present in the coding sequence and therefore caused a recalibration of the data in this thesis presented for this gene (*pglA.)* If this error had not been spotted it would have falsely identify false *pglA* phase variation in both the pilot study and lactamica 2 datasets.

The *N. Lactamica* Y92-1009 genome sequence bears all the hallmarks of a standard Neisseria genome. It is small (2.2Mb) and efficiently arranged with a high amount of coding content (85.3%) The DNA and CDS content in the comparative analysis between *N. lactamica* Y92-1009 and *N. lactamica* 020-06 revealed that most of the *N. lactamica* 020-06 genome is highly homologous with *N. lactamica* Y92-1009. Despite this there is an abundance of CDS found at 92% BLASTp identity. These suggest minor microevolutionary diversification and the possession of the same

genes but with varying alleles. This is consistent with the fact that both genomes belong to the same species and mirrors the results of earlier studies which compared *N. meningitidis* strains. Gross genome statistics found high similarities in GC content, RNA/DNA coding sequence content and genome size between all tested *N. meningitidis* isolates regardless of their differing origins (carriage vs disease), serogroup or clonal complex (Schoen *et al.*, 2008). However the flexibility of the meningococcal genome with regards to phase variation (Power and Moxon, 2006), intragenomic recombination (Schoen *et al.*, 2009) and repeat sequence possession (discussed later) aiding horizontal gene transfer has been shown to diversify meningococci within phylogenetic clades(Budroni *et al.*, 2011) and clonal complexes (Hao *et al.*, 2011).

.

Chapter 4: A short-term, multiple-colony sampled study: Comparing the microevolution of longitudinally sequenced isolates of *N. lactamica* in an *In Vivo* vs *In Vitro* genome cohort.

4.1 Introduction

This chapter aimed to identify the genetic differences undergone by bacteria that have been subjected to selective pressure in the nasopharynx of humans (*in vivo)* and contrast these with bacteria grown on rich media with minimal selective pressure (*in vitro*.) The experiment took place over the course of a month and utilised multiple colony sampling. Since this short-term study will inform future, longer-term mutational analyses (**see chapter 5**,) it is important to ascertain whether the type or frequency of mutations differentially affect the *in vivo* or *in vitro* cohorts. Because isolates from both cohorts were sampled longitudinally, it will also offer an opportunity to determine whether any mutations are stochastic or adaptive. A stochastic mutation is defined as a mutation that appears in a volunteer or control at just one-time point; never to be seen again. An adaptive mutation would be indicated by a mutation arising in one sample from one source of origin (i.e. either a volunteer's nasopharynx or an *in vitro* culture) and persisting until the next time point. To do this the *N. lactamica* Y92-1009 stock that was mass produced, purified (Cell-Bank) and artificially inoculated into human volunteers (or streaked onto blood plates) was sequenced, annotated and used as a reference genome to monitor base changes in the genomes of longitudinally acquired isolate genomes. The ongoing *Escherichia coli* long term evolution experiment has shown that bacteria cultured *in vitro* can develop traits not normally observed *in vivo.* These traits included citrate utilization (Blount, Borland and Lenski, 2008) (due to gltA gene mutation (Quandt *et al.*, 2015)) and hypermutation. Studies contrasting *in vivo* vs *in vitro* mutations in *Mycobacterium tuberculosis* found enhanced antibiotic resistance spanning different antibiotic classes among *in vitro* isolates when compared to *in vivo* (Bergval *et al.*, 2009; de Steenwinkel *et al.*, 2012). Among the *Neisseriaceae* no experiment to date has contrasted mutations to the genome observed *in vitro* passage versus *in vivo* carriage although a study noted distinct variation in colony morphology of gonococcal cells following repeated *in vitro* passage (Kellogg *et al.*, 1968)

SNP calling aligns informationally-rich, raw read data generated by sequencing to a reference genome. Variants detected because of this "mapping" are identified, statistically tested and

reported. There are multitude of tools that can do this. Whilst we have tested a variety of SNP callers including Harvest tools, kSNP and the bam/samtools pipeline, the SNP calling software which has worked best for our dataset so far was the breseq pipeline. Breseq was developed for resequencing projects such as this study (Deatherage and Barrick, 2014) and has unique data collation features that will express complex mutational data (SNPs, Indels, substitutions and repetitive tract changes) in a format that allows easy quantification and detection of longitudinally persistent mutants. This smaller dataset of *In vitro* and *in vivo* cohorts featuring a month of multiple colony sampling allowed us to test this technology before upscaling it to a larger dataset involving a greater number of volunteers and sampling over 6 months (Chapter 5.)

4.2 Methods

4.2.1 Short-term study dataset: Culture and isolation

A pilot study was undertaken by Paul Morris and Robert Read. Ten volunteers were intranasally inoculated with cell bank purified 10^4 CFU ml^{-1} *N. lactamica* Y92-1009 stock. Colonisation was established in five of these volunteers represented by the codes; AN, HE, SW, TN and EB (**Figure 4-1.**) Nasopharyngeal swabs were used to obtain samples at days 1, 2, 3, 14 and 28 post successful colonisation. If possible, DNA from up to a maximum of five colonies from each time point was extracted and sent for sequencing. The *in vitro* isolates of this study (113 & 115) were collected after serially passaging a sweep of 1-3 colonies of *N. lactamica* stock on Columbia blood Agar + 5% horse blood (ThermoFisher, UK) every weekday for up to a month. Four day-zero controls from the *in vitro* samples 113 and 115 and *in vivo* samples HE and SW were also sequenced.

4.2.2 Sequencing

The isolates were sequenced as described in (Evans *et al.*, 2011). Isolates from the study were sent to The Sanger Institute for 101bp Illumina paired end sequencing. Raw read sequencing files (fastq) were assembled into NEIS alleles to be used for allelic variation analysis. Both sets of files were uploaded onto the pubMLST Neisseria repository by Dr Julia Bennet (University of Oxford). The sequences and accession numbers are tabulated for the *in vitro* (**Figure 4-2**) and *in vivo* (**Figure 4-3**, **Figure 4-4**) cohorts.

Figure 4-1 : Number of isolates sequenced per time point per volunteer/ control during the study.

Cohort		Day 1	Day 2	Day 3	Week 2	Week 4	Total
In vitro **controls**	**113**	5	2	5	5	2	19
	115	0	0	0	3	5	8
In vivo isolates	**AN**	0	0	0	4	5	9
	HE	0	3	3	3		9
	SW	0	0	0	5	5	10
	TN	0	0	0	0	5	5
	EB	0	0	0	1	0	1

White cells indicate the numbers of isolates sampled from serially passaged blood plates (113 & 115) at a given time point. Blue cells indicate the numbers of isolates and time points where the isolates were sampled from 5 volunteers (AN, HE, SW, TN and EB). All isolates were sequenced. The total number of isolates from each cohort were n= 27 (*in vitro* cohort) & n= 34 (*in vivo* cohort).

Figure 4-2 *In vitro Cohort:* Passaged control strain PubMLST Id, contig number in assembly & SRA accessions

In vitro Control	Day	Colony	PubMLST id	Contigs	SRA accession
113	1	3	26747	122	ERR026921
113	1	2	26748	135	ERR026922
113	1	4	26749	137	ERR026923
113	1	5	26750	180	ERR026924
113	1	1	26753	130	ERR026927
113	2	1	26783	188	ERR028550
113	2	5	26794	199	ERR028562
113	3	3	26784	215	ERR028551
113	3	5	26785	217	ERR028552
113	3	1	26787	209	ERR028555
113	3	4	26795	246	ERR028563
113	3	2	26796	237	ERR028564
113	14	3	26786	235	ERR028554
113	14	1	26788	202	ERR028556
113	14	4	26789	216	ERR028557
113	14	5	26792	231	ERR028560
113	14	2	26797	244	ERR028566
113	28	5	26735	108	ERR026908
113	28	4	26798	249	ERR028567
113	28	3	26804	255	ERR028573
113	28	2	26805	243	ERR028574
115	14	2	26799	261	ERR028568
115	14	5	26803	262	ERR028572
115	14	1	26806	216	ERR028575
115	28	3	26736	100	ERR026909
115	28	5	26743	133	ERR026916
115	28	4	26744	140	ERR026917
115	28	2	26745	139	ERR026919
115	28	1	26752	145	ERR026926

Caption for Table 4-2

All isolates possessed the following allele numbers, MLST scores & strain designations; **abcZ**: 80, **adk:** 45, **aroE**: 98, **fumC**: 100, **gdh**: 94, **pdhC**: 158, **pgm:** 56, **Sequence type**: 3493, **clonal complex:** ST-613, **species**: *Neisseria lactamica*

Figure 4-3 *In vivo* cohort: volunteer AN, EB, HE isolates, PubMLST id, assembly contig number and SRA accession

Volunteer	Day	Colony	PubMLST id	Contigs	SRA accession
AN	14	1	26779	237	ERR028546
AN	28	1	26763	153	ERR028526
AN	14	2	26770	183	ERR028534
AN	28	2	26762	157	ERR028525
AN	28	3	26771	148	ERR028537
AN	14	4	26778	251	ERR028545
AN	28	4	26773	200	ERR028539
AN	14	5	26775	241	ERR028542
AN	28	5	26772	164	ERR028538
EB	28	1	26754	153	ERR26928
HE	3	1	26739	98	ERR026912
HE	14	1	26756	185	ERR028519
HE	2	2	26738	95	ERR026911
HE	14	2	26781	224	ERR028548
HE	3	3	26802	251	ERR028571
HE	2	5	26800	275	ERR028569
HE	3	5	26801	255	ERR028570
HE	14	5	26780	237	ERR028547

Caption for Table 4-3

All isolates possessed the following allele numbers, MLST scores & strain designations; **abcZ**: 80, **adk:** 45, **aroE**: 98, **fumC**: 100, **gdh**: 94, **pdhC**: 158, **pgm:** 56, **Sequence type**: 3493, **clonal complex:** ST-613, **species**: *Neisseria lactamica*

Figure 4-4 (CONTINUED) *In vivo cohort*: volunteer SW & TN isolates, PubMLST id, assembly contig number and SRA accession

Volunteer	Day	Colony	PubMLST id	Contigs	SRA accession
SW	14	1	26768	154	ERR028532
SW	14	2	26767	153	ERR028531
SW	14	3	26774	152	ERR028540
SW	14	4	26765	173	ERR028528
SW	14	5	26764	170	ERR028527
SW	28	1	26766	135	ERR028530
SW	28	2	26755	159	ERR028518
SW	28	3	26808	159	ERR028536
SW	28	4	26758	130	ERR028521
SW	28	5	26807	166	ERR028535
TN	28	1	26759	170	ERR028522
TN	28	2	26760	194	ERR028523
TN	28	3	26761	169	ERR028524
TN	28	4	26769	172	ERR028533
TN	28	5	26782	263	ERR028549

Caption for Table 4-4 (continued)

All isolates possessed the following allele numbers, MLST scores & strain designations; **abcZ**: 80, **adk:** 45, **aroE**: 98, **fumC**: 100, **gdh**: 94, **pdhC**: 158, **pgm:** 56, **Sequence type**: 3493, **clonal complex:** ST-613, **species**: *Neisseria lactamica*

4.2.3 Long read sequencing preparation

This method was performed as described in section 1.2.2

4.2.4 SNP calling

Paired, raw read files for each isolate were error checked and trimmed for nextera adapters using FastQC(Andrews, 2010) and trimmomatic (Bolger, Lohse and Usadel, 2014). These files were used to detect mutations for each isolate using the Breseq pipeline (Deatherage and Barrick, 2014)against a long read reference.

Breseq tabulates the SNP and annotation into an easily viewable and editable format and also creates a .gd file which contains data on mutations. Breseq is bundled with a series of tools that can be used to manipulate these bespoke mutation files called gdtools which can be used to;
a) Collate SNP data for samples, volunteer specifically into a browsable .csv file. This means that if a volunteer was sampled 3 times. Then the data on what genes were affected by mutation from day zero up to the third sample can be visualised against each other, to easily spot any emerging mutation patterns.
b) Visualise read coverage for a given genomic position and output a .png graphic of said position for
c) Perform a maximum-parsimony phylogenetic analysis.

Breseq can also be used to aid users in the manual curation of their reference genome. The pipeline automatically detects and reports areas of little or no read coverage (these are called “unassigned or missing coverage regions”) that are identified to the program user as they contain a much greater potential to bias mutation data. This feature is used later in the chapter (Figure 3-3) to trim the low-coverage zones present at the very beginning and end of an otherwise high-coverage and accurate assembly sequence.

Using the gdtools [subtract] command, any resultant .gd files were filtered for mutations that occurred because of comparing the long-read sequence of the Day zero inoculum strain of *N. lactamica* Y92-1009 to its corresponding short read sequence. The gdtools [annotate] command was used to generate & collate SNP tables and convert mutations into PHYLIP format for maximum parsimony phylogenetic analysis. An independent workflow using BWA(Li and Durbin, 2009), BAMtools (Barnett *et al.*, 2011) , Samtools (Li *et al.*, 2009)and

Tablet (Milne *et al.*, 2009) was implemented to predict mutations and reconfirm the results generated by the breseq pipeline.

4.2.5 Mutational analysis

To discriminate between the *in vivo* and *in vitro* cohorts, analysis focused on the mutation types and patterns of variation that were unique to the *in vivo* cohort while contrasted with the *in vitro* cohort.

4.2.6 PubMLST allelic analysis of genes detected as undergoing variation

The genome comparator tool on PubMLST Neisseria (Jolley and Maiden, 2010) was used to determine whether mutations identified among certain isolates correlated with alleles found to be variable. The genes in which coding sequence mutations were identified with Breseq were matched with NEIS loci on PubMLST Neisseria via BLAST (parameters: 100% identity, e-value <0.00, no gaps). These NEIS loci were then used as input into genome comparator along with their PubMLST isolate IDs

4.3 Results

In these sections I will compare and contrast the results of mutational analyses (sub divided by mutation type) of sequenced *N. lactamica* Y92-1009 isolates from an *in vivo* cohort of isolate genomes successfully recovered from colonised human volunteers) versus an *in vitro* cohort of isolate genomes recovered from serially passaged cultures.

4.3.1 Short-term study: Coding sequence SNP analysis

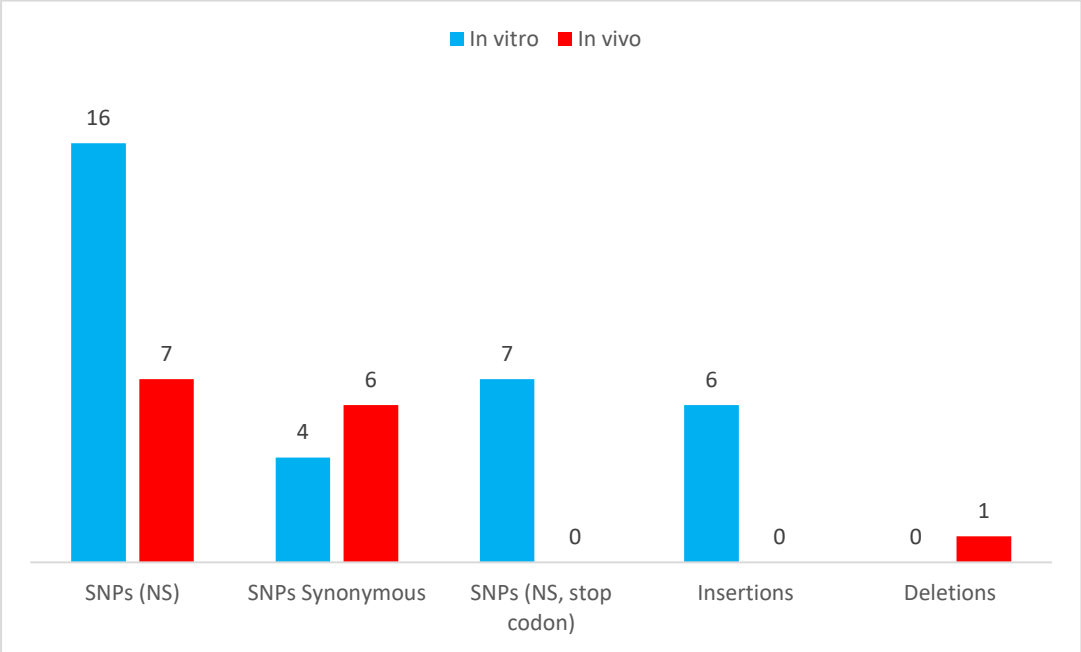
This section focuses on describing the results of the SNP analysis. This dataset is quantified in **Figure 4-1**. To summarise, we sequenced 29 *N. lactamica* Y92-1009 isolates which were restreaked on two sets (113 & 115) of Columbia blood agar and 34 isolates from 5 individuals artificially inoculated with a cell bank purified, day zero stock of strain Y92-1009. Therefore, the re-streaked isolates and those derived from humans were split into *in vitro* and *in vivo* cohorts to compare the mutational analyses between. This allows a comparison of the numbers and types of CDS and intergenic mutations observed in total. The nature of these mutations and whether they recur over time is explored for both cohorts in the rest of this results section.

The *in vitro* cohort displayed a greater number of non-synonymous & stop codon single nucleotide polymorphism mutations to that observed *in vivo* (**Figure 4-5**.) There were also more insertions observed for this cohort. The *in vivo* cohort showed marginally more synonymous SNPs but no insertions, only one deletion and no examples of SNPs introducing premature stop codons.

Non-synonymous SNPs detected in the *gltA* gene of blood plate control 113 occurred in all the isolates sampled at week 2 and week 4 **(Figure 4-6.)** The insertion and stop codon SNPs observed in the *mqo* and *pilP* genes of 115 also recur. This contrasted with the *in vivo* cohort in which there was limited evidence of longitudinal mutation recurrence **(Figure 4-7.)** Except the non-synonymous *nuoL* mutation present in all the isolates recovered from volunteer AN week 2, all mutational events occurred only once per time point despite multiple colony sampling. Mutations in the *mqo* and *phrA genes* were seen to occur in both cohorts. However, the location and synonymity of *mqo* mutation appeared to differ while the identical, synonymous *phrA* mutation occurred in both the *in vitro* cohort as well as volunteers HE and TN. Synonymous and non-synonymous SNPs were not observed in volunteer EB or volunteer AN week 2 samples. Of the sixteen detected non-synonymous SNPs among the *in vitro* cohort; eleven changed the amino acid side chain polarity from polar to non-polar and vice versa, three changed the side chain charge from positive, negative or neutral, one changed both side chain polarity and charge and one changed to an amino acid which had no effect on net side chain polarity and charge. Conversely, of the seven detected non-synonymous SNPs among the *in vivo* cohort; one changed amino acid side chain polarity, one changed side chain charge and five changed to amino acids that

possessed the same side chain charge and polarity as the amino acid they originally substituted.

Figure 4-5 Comparison of coding sequence mutation types between the *in vivo* and *in vitro* cohorts



Single nucleotide polymorphisms (**SNPs**) affect codons which may be non-synonymous (**NS**) changing the amino acid encoded or introducing a premature stop codon (**NS, stop codon**). Conversely, they may maintain the encoded amino acid (**synonymous**). Frame shifts can be introduced into genes via the addition of superfluous bases (**insertion**) or the subtraction of bases (**deletion**).

Figure 4-6 Coding sequence mutations across time detected in the *in vitro* cohort

Control	position	Mutation Type	Day 1	Day2	Day 3	Week 2	Week 4	Annotation	Gene	Description
113	221,924	SNP NS: polarity	0	0	1	0	0	Y64C (TAC→TGC)	mqo	Malate:quinone oxidoreductase
113	261,175	SNP Synonymous	0	0	0	0	1	G45G (GGT→GGG)	PROKKA_00276	hypothetical protein
113	394,982	SNP NS: charge	1	0	0	0	0	D70N (GAT→AAT)	rfbA	Glucose-1-phosphate thymidylyltransferase 2
113	551,325	SNP Synonymous	0	0	2	0	0	P170P (CCA→CCC)	phrA	Deoxyribodipyrimidine photo-lyase
113	1,214,369	SNP NS: polarity	0	0	0	5	4	A254T (GCC→ACC)	gltA	Citrate synthase
113	1,325,826	SNP NS: no change	0	0	0	0	1	N195T (AAC→ACC)	PROKKA_01319	Putative lipoprotein/NMB1164 precursor
115	221,175	Insertion +CC	0	0	0	2	4	coding (940/1467 nt)	mqo	Malate:quinone oxidoreductase
115	551,325	SNP Synonymous	0	0	0	0	1	P170P (CCA→CCC)	phrA	Deoxyribodipyrimidine photolyase
115	788,384	SNP NS: *	0	0	0	2	5	W154* (TGG→TGA)	pilP	Pilus assembly protein
115	1,213,733	SNP NS: polarity	0	0	0	1	0	S42P (TCC→CCC)	gltA	Citrate synthase
115	1,321,222	SNP NS: polarity, charge	0	0	0	0	1	D188A (GAC→GCC)	PROKKA_01314	haloacid dehalogenase-like hydrolase
115	1,842,835	SNP NS: charge	0	0	0	2	0	D213Y (GAC→TAC)	PROKKA_01787	putative FAD-linked oxidoreductase

The **position** column shows the base the mutation affected (out of a potential 2,146,723 bases). A SNP is short for single nucleotide polymorphism. Non-synonymous SNPs (**NS**) have the potential to introduce amino acids which vary side chain polarity (**NS: polarity**), side chain charge (**NS: charge**), introduce a premature stop codons (**NS: ***) or maintain polarity and charge (**NS: no change**). The **day** and **week** columns show the number of detected mutations among samples isolated from those time-points. In the **Annotation** column the unbracketed letters and numbers refer to amino acid codes and the position of the where substituted amino acid is found in the protein translation of the gene. The bracketed letters describe the codon change that led to this outcome. This column also shows the position of the nucleotide affected by insertion events. The **gene** and **description** column list information on the gene the mutation occurred in.

Figure 4-7 Coding sequence mutations across time in the *in vivo* cohort

Volunteer	position	Mutation Type	Day 1	Day2	Day 3	Week 2	Week 4	Annotation	Gene	Description
AN	500,676	SNP NS: no change	0	0	0	4	0	A637V (GCG→GTG)	nuoL	NADH-quinone oxidoreductase subunit L
HE	551,325	SNP Synonymous	0	0	1	0	0	P170P (CCA→CCC)	phrA	Deoxyribodipyrimidine photo-lyase
HE	1,399,418	SNP NS: polarity	0	1	0	0	0	A26T (GCA→ACA)	PROKKA_01401	hypothetical protein
HE	1,438,953	Deletion	0	0	0	1	0	Δ37 bp	PROKKA_01446	hypothetical protein
HE	1,672,226	SNP NS: no change	0	0	1	0	0	R422H (CGC→CAC)	gyrA	DNA gyrase subunit A
SW	118,481	SNP Synonymous	0	0	0	0	1	S328S (TCC→TCT)	pilG	Type IV pilus biogenesis protein
SW	220,798	SNP Synonymous	0	0	0	0	1	I439I (ATC→ATT)	mqo	Malate:quinone oxidoreductase
SW	551,325	SNP Synonymous	0	0	0	1	0	P170P (CCA→CCC)	phrA	Deoxyribodipyrimidine photo-lyase
SW	717,208	SNP Synonymous	0	0	0	1	0	T54T (ACA→ACC)	accB	Carrier protein of acetyl-CoA carboxylase
SW	1,252,506	SNP NS: charge	0	0	0	0	1	D120N (GAC→AAC)	PROKKA_01249	AMP protein trasnferase Sofic
TN	551,325	SNP Synonymous	0	0	0	0	1	P170P (CCA→CCC)	phrA	Deoxyribodipyrimidine photo-lyase

The **position** column shows the base the mutation affected (out of a potential 2,146,723 bases.) In the **mutation type** column **SNP** is short for a single nucleotide polymorphism; these can be synonymous or non-synonymous (**NS**). Non-synonymous SNPs have the potential to introduce amino acids which vary side chain polarity (**NS: polarity**), side chain charge (**NS: charge**), or maintain polarity and charge (**NS: no change**). The **day** and **week** columns show the number of detected mutations among samples isolated from those time-points In the **Annotation** column the unbracketed letters and numbers refer to amino acid codes and the position of the where substituted amino acid is found in the protein translation of the gene. The bracketed letters describe the codon change that led to this outcome. This column also shows the number of base pairs affected by the deletion event. The **gene** and **description** column list information on the gene the mutation occurred in.

4.3.2 Changes in the repetitive sequences are more prevalent among the *in vivo* rather than *in vitro* genome group in the study

Due to the accuracy of mutational analysis based off a single, contiguous reference genome, it was possible to accurately analyse changes in the repetitive sequence structure in the coding sequence and intergenic regions surrounding potential contingency loci. The *in vivo* cohort showed elevated levels of this type of mutation (phase variation occurring in CDS tracts) (**Figure 4-8**). In order to determine which repetitive sequence changes may be more indicative of host adaptation than stochastic mutation, tract changes in the *in vivo* cohort were discriminated by mutational position and genes affected compared to the *in vitro* cohort. It was noted that there were no instances of unique repetitive sequence change in the *in vitro* cohort. Any that were found were identical in location and repeat number to a repetitive sequence change found in the *in vivo* cohort. **Figure 4-9** summarises novel changes in repetitive sequence structure targeting genes undergoing phase variation in both cohorts While the gene targets of said mutation were the same in this table, it also indicated variants of the mutation that were only detected in the *in vivo* cohort. Two of these genes were primarily associated with modifying structures (LOS) mediating host pathogen response (glycosyl trasnferases; *lgt* gene.) There was also phase variation detected in the specificity subunit (*hsdS*) of the type 1, dual gene, NgoAV restriction modification system. In addition, **Figure 4-10** displays phase variation seen only in the *in vivo* group. Three of these four novel phase variants occurred in genes of biological interest that have affected iron acquisition in *Neisseria spp.* (*fetA, hpuA* and rubredoxin.)

Figure 4-8 Comparison of repetitive sequence changes in the *in vivo* and *in vitro* cohorts.

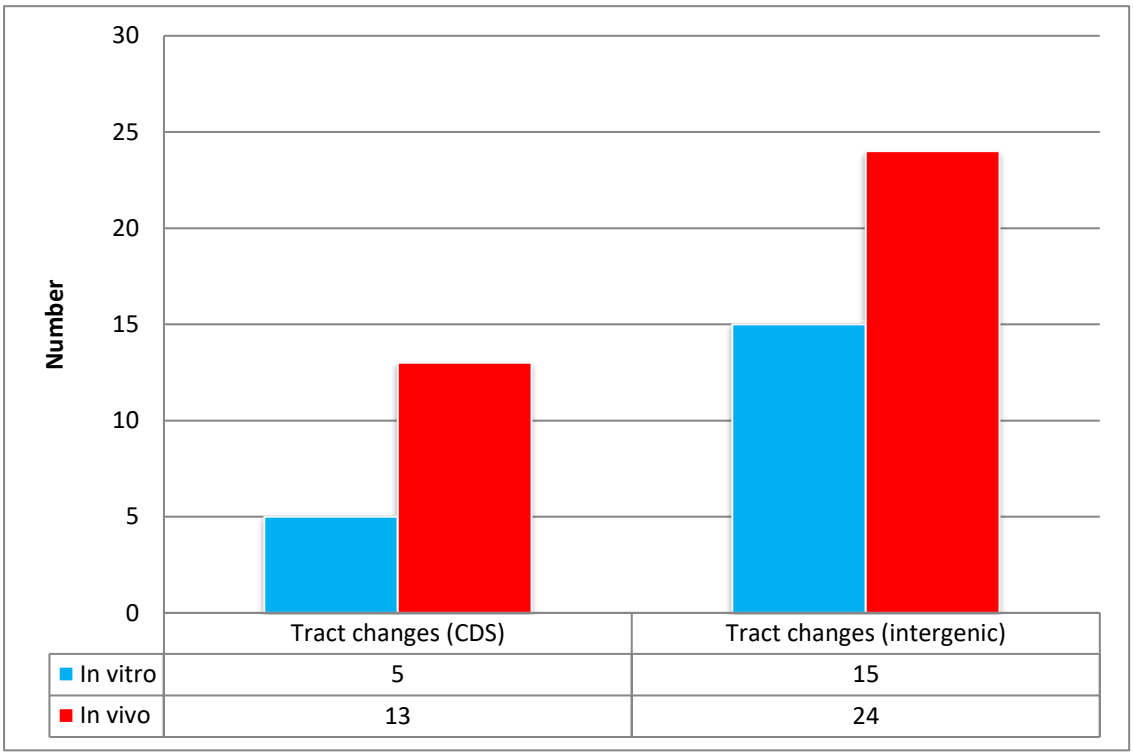


Figure 4-9 : List of mutations affecting polymeric tracts in the *in vivo* cohort which occurred in the *in vitro* cohort.

Position	shared mutation	in vivo only variant	incidence	volunteer/timepoint	location of mutation	gene(s)	details
161,341	$(C)_{12\rightarrow13}$	$(C)_{12\rightarrow11}$	1	AN/W4	coding sequence	lgtG	O3 linked glucosyltransferase LOS subunit modifier
1,124,158	$(G)_{9\rightarrow10}$	$(TACGCTGGAAGC)_{1\rightarrow2}$	1	SW/W2	coding sequence	hsdS	type I restriction modification system, specificity subunit S
1,124,237	$(G)_{9\rightarrow10}$	$(G)_{9\rightarrow11}$	1	HE/W2	coding sequence	hsdS	type I restriction modification system, specificity subunit S
1,483,216	$(G)_{10\rightarrow9}$	$(G)_{10\rightarrow11}$	2	AN/W2	coding sequence	conserved hp	glycosyl transferase
189,023	$(A)_{20\rightarrow19}$	$(A)_{20\rightarrow17}$	1	HE/D3	intergenic	NEIS2043: 2 (Thif) → / ← NEIS2044: 28	enzyme activation/hp
593,527	$(G)_{10\rightarrow9}$	$(G)_{10\rightarrow12}$	1	AN/W4	intergenic	FetA VR F4-8 ← / → NEIS1949	finetyping antigen/ GroS chaperone protein folding
837,301	$(CTTG)_{9\rightarrow10}$	$(CTTG)_{9\rightarrow11}$	1	TN/W4	intergenic	*hp* ← / → *NEIS0842:42*	YadA-like C-terminal region/ conserved hypothetical protein
893,705	$(T)_{18\rightarrow17}$	$(T)_{18\rightarrow15}$	2	HE/D3	intergenic	NEIS0978 (396) ← / ← NEIS0528(112)	putative surface fibrial protein/putative periplasmic binding protein
893,707	$(T)_{18\rightarrow17}$	$(T)_{18\rightarrow19}$	1	EB/W2	intergenic	NEIS0978 (396) ← / ← NEIS0528(112)	putative surface fibrial protein/putative periplasmic binding protein
1,933,087	$(A)_{28\rightarrow27}$	$(A)_{28\rightarrow26}$	1	AN/W2	intergenic	AutA /hp	autotransporter A

Grey cells indicate changes to intergenic repetitive tracts while **white cells** indicate CDS repetitive tract changes. The **shared mutation column** displays a change in the repetitive sequence tract of a given gene that is shared between the *in vivo* and *in vitro* cohort. The corresponding "***in vivo* only**" column shows a novel tract change that occurs exclusively but also targeting the same gene, as a mutation found in the *in vitro* cohort. The base or bases in brackets in both of these columns refers to the genetic sequence or base that is repeated. For example "$(A)_{20)\rightarrow17}$" means that a sequence of 20 consecutive adenine bases has been shortened to 17.

Key: **D** = Day**, W**= Week**, hp**= hypothetical protein. **AN**, **HE**, **EB**, **SW** & **TN** are separate volunteers

Figure 4-10 List of unique mutations affecting genes only the *in vivo* cohort.

Position	mutation	incidence	volunteer/timepoint	location of mutation	gene(s)	details
617,862	$(C)_{10\rightarrow9}$	1	SW/W2	coding sequence	hpuA	haemoglobin-haptoglobin utilisation protein
1,483,216	$(C)_{10\rightarrow11}$	1	AN/W4	coding sequence	hp	glycosyl transferase activity
1,688,136	$(G)_{13\rightarrow12}$	7	AN/W2(4) & W4(2) & SW/W4(1)	intergenic	l-hp/rubredoxin	large multi domain protein / iron-ion binding protein
1,717,988	$(C)_{10\rightarrow9}$	1	SW W4	intergenic	ybaB/MDP	DNA binding protein /multi domain protein

Grey cells indicate changes to intergenic repetitive tracts while **white cells** indicate CDS repetitive tract changes. The **mutation** column displays a change in the repetitive sequence tract of a given gene. The bracketed base in this column refers to the composition of the polymeric tract. For example "$(G)_{13)\rightarrow12}$" means that a sequence of 13 consecutive guanine bases has been shortened to 12.

Key: **D = Day, W= Week, AN** & **SW** are separate volunteers.

.

4.3.3 Allelic profiling of *in vitro* and *in vivo* cohorts reveals discrepancies in allele type

As mentioned in **4.2.6**, the genes affected by coding sequence mutations were matched with NEIS loci on www.pubmlst.org/Neisseria. This allowed the use of the genome comparator tool to determine whether allelic variation occurred in the same isolates found to be the source of mutation. Mutations occurring in PROKKA_00276 (83% identity), PROKKA_01446 (89% identity) & PROKKA_01401 (27% identity) genes were unable to be matched via BLAST with NEIS loci and were excluded from this analysis. **Figure 4-11** shows the genes affected by coding sequence mutation in this chapter, their NEIS locus equivalents and other aliases. The table also shows the reference alleles present among unmutated isolates alongside the variable alleles.

Allelic variation was perfectly correlated with isolates with mutable genes in the case of *rfbA* (NEIS0046, n=1), *pilP* (NEIS0409, n=7), PROKKA_01787 (NEIS1453, n=2), PROKKA_01319 (NEISp1066, n=1) and *pilG* (NEIS1838, n=1). Allelic variants were found for n=7/9 isolates containing mutable *gltA* (NEIS0930). And while all isolates containing mutable *mqo* (NEIS2076, n=8) and PROKKA_01249 (NEIS2498, n=1) genes were identified as containing variable alleles, the analysis identified an additional isolate in both cases with a variable allele which did not correlate with detected mutations.

Due to the large presence of incomplete alleles, genome comparator was unable to correlate allelic variation among isolates with mutable *phrA* (NEIS1892, n=6) or *accB* (NEIS0358, n=1). Allelic variation was identified among only 1/4 isolates with mutable *nuoL* (NEIS0251) and no variation was found in the sole isolate with a detected PROKKA_01314 (NEIS1039) mutation. None of the isolate genes affected by coding sequence phase variation were discriminated by allelic analysis due to the presence of incomplete or non-differing alleles. These included the conserved hypothetical protein (NEIS1156), *lgtG* (NEIS2011), *hsdS* (NEIS2362) and *hpuA* (NEIS1946)

Figure 4-11 Mutable genes: NEIS loci counterparts, Aliases and Genome comparator results

Gene	Mutation type (s)	Cohort	NEIS locus	Alias	Reference allele	Variant allele(s)
mqo	SNP S (X2), INS	Both	NEIS2076	NMB2096	2	100, 101, 102 "New#1"
accB	SNP S	In Vivo	NEIS0358	NMB1860	Incomplete	N/A
pilG	SNP S	In Vivo	NEIS1838	NMB0333	25	35
phrA	SNP S	Both	NEIS1892	NMC1892	3	82, 83, 84
gltA	SNP NS (X2)	In Vitro	NEIS0930	NMB0954	30	105
nuoL	SNP NS	In Vivo	NEIS0251	NMB0257	29	133
pilP	SNP NS	In Vitro	NEIS0409	NMB1811	32	"New#1"
PROKKA_01249	SNP NS	In Vivo	NEIS2498	N/A	16	36, 37
PROKKA_01314	SNP NS	In Vitro	NEIS1039	NMB1075	3	no variants
PROKKA_01319	SNP NS	In Vitro	NEISp1066	NMB1164	3	56
PROKKA_01787	SNP NS	In Vitro	NEIS1453	NMB1524	2	112
rfbA	SNP NS	In Vitro	NEIS0046	NMB0080	62	158
c-hp	PV	Both	NEIS1156	N/A	18	"New#1", "New#2"
hpuA	PV	In Vivo	NEIS1946	NMC1946	45	"New#1"
hsdS	PV	In Vivo	NEIS2362	NGO_02155	"New#1"	"New#2,#3,#4,#5"
lgtG	PV	Both	NEIS2011	NMC2011	16	"New#1,#2"

Caption for Table 4-9

In the **mutation type (s)** column **SNP** is short for a single nucleotide polymorphism; these can be synonymous **(S)** non-synonymous (**NS**). SNPs were sometimes found affecting the same gene but in two different locations **(X2)**. **Ins** refers to Insertion events and **PV (**short for phase variable) describes insertions and deletions affecting polymeric tracts. The **cohort** column indicates whether a gene was found to mutate in the ***in vitro, in vivo*** or **both** cohorts. The **reference allele** column shows the baseline allele in the majority of unmutated isolates while the **variant allele** column lists the alleles shown to deviate from this majority. The result of a **"New#"** allele occurred where genome comparator encountered a sequence not designated an allele number on its database. In the case of the *hsdS* gene, both the reference allele ("New#1") of the majority and the varying alleles ("New#2, New#3, New#4 & New#5) were sequences currently uncatalogued on PubMLST Neisseria. Allele numbers and NEIS loci can be utilised together to search for sequence information on www.pubmlst.org/Neisseria.

4.4 Discussion

4.4.1 Comparisons of SNP analysis between the *in vivo* and *in vitro* cohorts

This short-term study represents a brief window into the *in vivo* and *in vitro* microevolution of *N. lactamica* Y92-1009. The minimal selective pressure introduced by passaging the strain on blood agar as opposed to the human nasopharynx demonstrated both a greater number of non-synonymous/ synonymous mutations as well as their recurrence in the *in vitro* cohort. A study (Koskiniemi *et al.*, 2012) found that 25% of their examined deletions improved the growth rates of the *Salmonella enterica* grown in nutrient rich media. They postulate that deletions improved metabolic fitness in this medium by removing the energy burden associated with synthesizing potentially unrequired proteins. Non-synonymous SNPs have a greater potential of altering gene transcriptions than synonymous mutation. But among non-synonymous mutations, those that substitute amino acids with differing side chain polarity and/or charge have a greater effect on the tertiary structure of the protein once translated. The non-synonymous SNPS detected and maintained among the *in vitro* cohort were more likely to alter side chain polarity, charge and introduce a premature stop codon than those observed among the *in vivo* cohort. This finding mirrors *E. coli* passaged *in vitro* during the long term evolution experiment in which a continual, positive selection of mutations were found among evolving isolates (Barrick *et al.*, 2009).
The mutations detected *in vivo* isolates were lower in number despite the the cohort size being marginally higher. Non-synonymous mutations were fewer in number and less likely to persist or introduce non-synonymous SNPS compared to *in vitro.* However, the number and variety of mutations in phase variable genes targeting polymeric tracts were larger in the *in vivo* cohort. The mutations observed targeted genes broadly involved in synthesising modifiers of subunits involved in the host-pathogen interaction (hypothetical protein with glycosyl transferase activity & *lgtG*), iron acquisition (*hpuA*) and self-regulating, restriction/modification (*hsdS*). This could allow *in vivo* adapted isolates to offer a broader range of phenotypic diversity in their niche without a large mutation burden. These findings were similar to data seen in longer term studies of *Pseudomonas aeruginosa.* The successfully-colonising, *in vivo*-acquired isolates described were subject to purifying selection post rapid adaptation via a small number of mutations (Yang *et al.*, 2011).

A recurring mutation in serially passaged, *in vitro* control 115 involved the insertion of a stop codon located 154bp/774bp into the *pilP* encoding sequence. This would have a greater chance of negatively affecting expression of the PilP protein and subsequently the formation of a PilQ

multimer, an essential component of type IV pilus biogenesis (Drake, Sandstedt and Koomey, 1997). Alongside motile functions, the Neisserial pilli are known to be essential contributors to the initial attachment of *Neisseria* species to human cells (Stephens, Krebs and McGee, 1984) after which they are subsequently lost to promote more intimate attachment (Pujol *et al.*, 1999). This loss of pili is known to occur rapidly during *in vitro* culture (Devoe and Gilchrist, 1975) but at different rates dependant on the media used (McGee *et al.*, 1979).

The selective pressure involved in adapting to colonise the nutrient-limited nasopharynx compared to a rich medium may allow greater tolerance for the persistence of potentially disruptive mutations in a population. The *in vitro* cohort displayed non-synonymous, recurrent mutation in *gltA* and both cohorts displayed mutation in *mqo*. These genes are both essential components of the tricarboxylic acid cycle. However, the *mqo* mutation observed singularly *in vivo* was synonymous while the mutations observed *in vitro* were either nonsynonymous or frameshifting. This means that *in vitro* passaged *N. lactamica* isolates may have potentially lost the ability to synthesize isocitrate from oxaloacetate and Acetyl Co-A (*gltA* inactivation) or generate oxaloacetate from malate (*mqo* inactivation). The TCA cycle in *Neisseria* is used to synthesise metabolic precursors while essential glucose catabolism can be achieved alternatively using the Entner-Doudoroff and pentose phosphate pathways (Schoen *et al.*, 2014). This may therefore be a case of the *in vitro* passaged bacterium evolving to better suit it's new environment by deactivating the expression of a non-essential gene. This effect has been observed among *in vitro* passaged *E. coli* isolates in the long-term evolution experiment which were found to eventually lower citrate synthase activity of *gltA* through mutation as they gained the ability to metabolise citrate (Quandt *et al.*, 2015).

The breseq pipeline (Barrick *et* al., 2009) was successfully used in this study to detect SNPs in two studies (This chapter & chapter 5) utilising a Prokka annotated, highly accurate *N. lactamica* genome as a reference. Breseq includes support for a variety of short read sequencing technology. Raw sample data can be inputted in the form of unpaired or paired. fastq files. While Breseq does support the use of these files it does not utilise the additional positional data present in paired end sequenced reads and instead converts everything to unpaired read data to map to the reference assembly. The developer's (Jeff Barrick, University of Texas) explanation for this lack of utilisation is that while using paired end data is already achievable and would technically be more effective, the issue is a trade-off between program run time and marginal, elevated accuracy of SNP calls. The Breseq manual advises to consider the types of mutations this pipeline cannot discriminate. Namely, novel sequences or inversions, not existing in the reference. This is the inherent issue of using a reference-based approach to SNP call and necessitates reference genome accuracy. In addition to SNP calling software, this chapter sought to use allelic variation

via the genome comparator tool hosted on https://pubmlst.org/neisseria/ to validate mutation detection. Excluding examples containing incomplete alleles, genome comparator was able to detect all instances of non-synonymous SNPs and insertion events. The tool was less capable of distinguishing synonymous SNPs and not capable at all of identifying allelic variation among isolates in which coding sequence phase variation was detected. Taken as a whole, genome comparator was an excellent way of quickly validating data among isolate samples with detected mutations.

Bacteria can lose or gain genetic material through a variety of mechanisms influenced by natural selection; however, this may also occur stochastically. Any new variant, especially a beneficial one, can become fixed in a rapidly diversifying population. The most important finding of this chapter was to discriminate differences in the types of mutation seen in bacteria evolving in an *in vivo* environment versus an *in vitro* one. This allowed us to expand our study from one month and five volunteers (this chapter) into a longer-term study utilising a greater number of volunteers (chapter 5).

Chapter 5: The long-term, *in vivo* microevolution of *N. lactamica* Y92-1009 in a student cohort of longitudinally sampled & colonised individuals.

5.1 Introduction

The fall of costs associated with genome sequencing has led to non-environmental microbiologists expanding their sequencing repertoire to encompass not just pathogens but greater varieties of asymptomatic and commensal bacteria. This has allowed for ever more accurate and customised analyses. In the past, a study may have described how global isolates of the same species differ in a gene. It is now possible to sequence and compare every ORF in every isolate from the same strain under investigation.

At the time of writing, this type of experiment has so far not been attempted for the *Neisseria* although a longitudinal study has assayed ten, phase variable, meningococcal genes for evidence of microevolution (Alamro *et al.*, 2014). Among other bacteria, WGS has already been used to understand the effects of within-host selection pressure and microevolution ((reviewed in (Didelot, Walker, *et al.*, 2016)) Examples include the asymptomatic carriage and within host evolution of another nasopharyngeal coloniser *Staphylococcus.aureus* (Golubchik *et al.*, 2013) as well as convergent evolution detected among hypermutating *Pseudomonas.aeruginosa* infecting patients suffering from cystic fibrosis (Marvig *et al.*, 2015) These papers contain evidence of bacterial pathogens mutating in similar ways across multiple subjects which suggests the presence of an inherent mechanism activated by the bacterium when transitioning to human carriage

Due to the accuracy of the reference used in SNP calling, any evidence of a volunteer unspecific, adaptive change with regards to the consistent accumulation of similar mutations across multiple volunteers will be able to be reported in a similar way as tested in **chapter 4**.

A study (Deasy *et al.*, 2015) noted differences in whether meningococci, persisted or were displaced by *N. lactamica* carriage. All *N. lactamica* genomes from volunteers co-colonised by meningococci will be sequenced. Their subsequent mutation results following SNP analysis will be compared as a group against isolates against volunteers in which no meningococcal isolates were recovered. Pathogenic and commensal *Neisseria* species have been known to exchange genetic information, this is a driver of diversity for the genus (Linz *et al.*, 2000). Assuming these interspecific interactions occur, it could mean that the genomes of *N. lactamica* isolates from co-

carrying volunteers should be more plastic and demonstrate larger and more frequent insertions and deletions caused by horizontal gene transfer.

In the previous chapter, I discovered variation in the number and type of mutations from the same bacterium either passaged *in vitro* or collected from colonised volunteers *in vivo* over the course of a month. Although I was unable to detect a pattern of adaptive carriage seen consistently among five volunteers and approximately 30 genomes in the *in* vivo cohort, I demonstrated that the majority of microevolution in host dwelling *N. lactamica* Y92-1009 occurs in phase variable genes containing polymeric tracts. The use of this mechanism could potentially allow rapid changes in bacterial expression governed by natural selection.

The experimental human challenge study contained *N. lactamica* isolates colonising many more volunteers and collected up to six months after the initial inoculation. This would therefore be a good basis to study the longer-term *in vivo* evolution of a commensal *Neisseria* species using as much genetic information as possible. As before, isolate sequenced reads will be mapped onto a highly accurate reference genome, any mutations detected will be statistically confirmed by SNP calling software and re-confirmed manually.

5.2 Methods

5.2.1 Controlled Human Infection with *N. lactamica*

The inoculated strain *N. lactamica* Y92-1009 was originally isolated from a 1992 carriage study of school pupils in Londonderry, Northern Ireland. The controlled human infection study has been reported previously (Deasy *et al.*, 2015). 149 participants received nasal challenge with 10^4 CFU of *N. lactamica*. In this group, natural *N. meningitidis* carriage at baseline was 36 of 149 (24.2% [95% CI, 17.5%–31.8%]). Natural carriage of *N. lactamica* prior to inoculation was observed in 3 of 149 (1.9% [95% CI, .4%–3.5%]). Oropharyngeal swabs were taken 2, 4, 8, 16, and 26 weeks after inoculation and were plated directly onto gonococcus (GC)-selective media (E&O Laboratories, Scotland) and incubated at 37°C in 5% CO_2. After 48 hours, possible *Neisseria spp.* colonies were subjected to API-NH strip testing (bioMérieux, France) and PCR (Deasy *et al.*, 2015). Two weeks after inoculation, oropharyngeal swabs from 48 individuals challenged with *N. lactamica* yielded cultivable *N. lactamica* (33.6% [95% CI, 25.9%–41.9%]), but colonization with *N. lactamica* was detected as late as week 26, at which point colonisation with *N. lactamica* at some point in the study had been confirmed in 61 of 149 (41.0% [95% CI, 33.0%–49.3%]) of the challenge group. Of the 61 volunteers, the isolates from 35 were selected for sequencing following the rationale described just below. All isolates of *N. lactamica* and *N. meningitidis* were cryostored.
Clinical Trial Registration: NCT02249598.

5.2.2 Sequencing isolates

The sequencing-selection strategy prioritised a) isolates from volunteers that were present for the most timepoints; b) isolates from volunteers co-colonised with *N. meningitidis* and; c) isolates recovered either earlier (week 2) in the study or later (week 26) with the rationale that the earlier time point had undergone adaption and selection more recently and the isolates at the latest time point would have a greater chance of demonstrating signs, if any, of convergent evolution having colonised their respective hosts for the longest period of time.

5.2.3 Mutational analysis

Experimental procedures for mutational analyses were performed as outlined in **section 2.4.2** using the Breseq pipeline. All mutations observed and reported were manually checked using genome browsers for false positives arising because of low read coverage and subsequent statistical bias.

5.2.4 Phase variable loci detection

Putative ON/OFF expression states were assigned to isolates based on the seven contingency loci characterised in previous *Neisseria* studies (Snyder and Saunders, 2006; Marri *et al.*, 2010)). The numbers of ON or OFF states per time point using the statistical method outlined above were used to determine statistical significance. A count of these loci per isolate was used to generate a phasotype score where an ON-expression state was equal to a value of 2 and an OFF expression state was equal to 0 (Alamro *et al.*, 2014). An isolate expressing putative ON states for all seven contingency loci would therefore have a combined phasotype score of 14.

5.2.5 Protein Analysis

Where applicable hypothetical proteins were investigated using the conserved domain architecture tool (Geer *et al.*, 2002) which was used to meta-analyze the predicted protein domains. These domains alongside any mutations that were seen to occur during longitudinal carriage were used as input for plot protein ((version unknown(Turner, 2013)to generate a protein diagram. *In silico* translations on the knock on effects of mutation on protein size were examined using ExPASy translate (Gasteiger *et al.*, 2003).

5.2.6 Mutation rate calculations

Mutation rate was also determined with regards to the number of SNPs discovered per site (bases in *N. lactamica* genome multiplied by volunteers examined) per year. The isolates in the dataset was sub-clustered into volunteers presenting with carriage of the inoculated *N. lactamica* Y92-1009 (n=27 volunteers) and volunteers displaying co-carriage of the inoculated *N. lactamica* with wild type, *N. meningitidis* (n=8 volunteers). The mutation rate was determined and contrasted between both of these subsets

5.3 Results

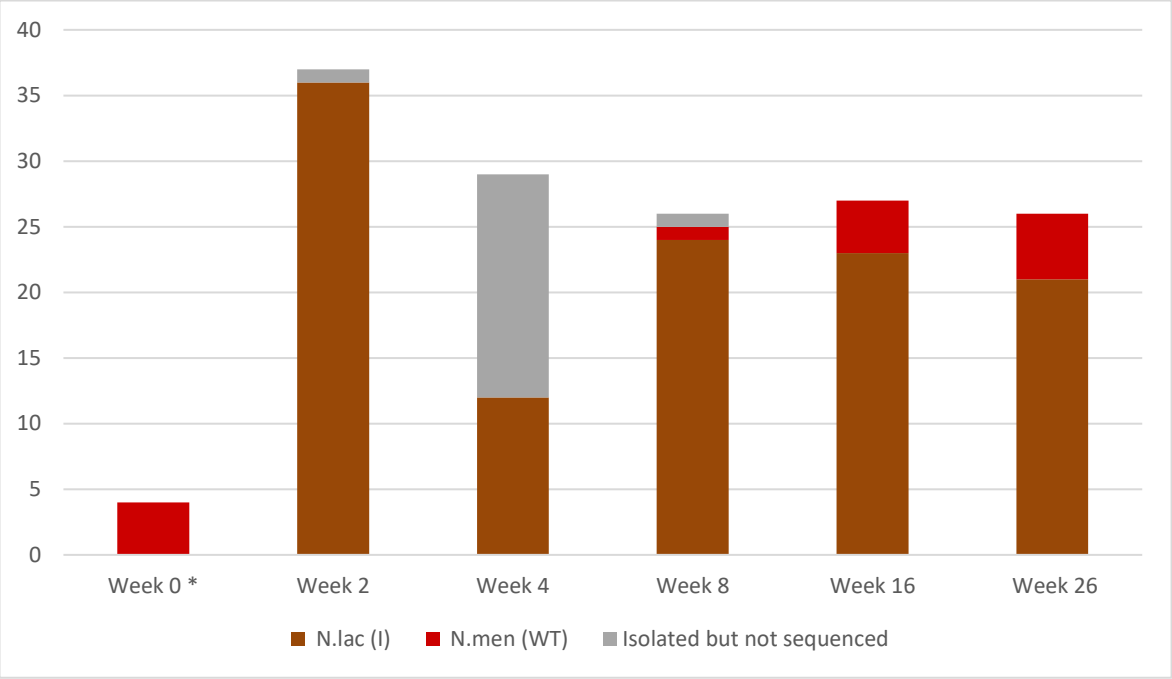
Having noted a difference in mutational profiles in short term (1 month) *in vivo* and *in vitro* carriage of *N. lactamica.* I utilized isolates from a longer (6 months) experimental human challenge study to study the long-term evolution of *N. lactamica* Y92-1009.

In this section I will first describe the dataset of the study in terms of recovered isolates, sequenced isolates and species of *Neisseria* sequenced. Following this, I will describe the phase variable events that formed the bulk of detected mutations in this study. Then I will describe the mutation type and consequence affecting the only non-phase variable gene in which mutation was detected across different volunteers. After this I will catalogue the SNPs that were seen to persist volunteer specifically before then detailing the transient mutations (SNPs, insertions and deletions) that occurred once per volunteer per timepoint and were not detected again. Finally, I will compare mutation rate estimates of the inoculated, sole-colonising *N. lactamica* Y92-1009 and inoculated *N. lactamica* Y92-1009 that co-colonised with wild type, baseline *N. meningitidis.*

5.3.1 Lactamica long term evolution study: dataset metrics

A total of 95 isolates of *N. lactamica* and 14 isolates of *N. meningitidis* from 35 participants were isolated, sequenced and uploaded to PubMLST neisseria. **Figure 5-1** shows the sampling points at which *N. lactamica* Y92-1009 and/or *N. meningitidis* was isolated. Appendix Table 0-6 lists the meningococcal isolates and their PubMLST IDs. Appendix Tables 0-7 and 0-8 list the inoculated, *N. lactamica Y92-1009* isolates and PubMLST IDs. An additional 2 individuals carried wild type *N. lactamica* at baseline and belonged to sequence types other than the inoculum strain (ST 4192 & 11524) and are discussed further in **chapter 6**. In 91% (n=32/35) of participants, *N. lactamica* was isolated from more than one sampling point over the course of 6 months. In 10 individuals both *N. lactamica* and *N. meningitidis* was obtained with 7 individuals having simultaneous carriage of isolates of both species that could be isolated.

Figure 5-1 Barchart summarizing sequenced isolates of *N. meningitidis*, *N. lactamica* and recovered but unsequenced isolates for each time point



Across 35 volunteers a total of 95 isolates of inoculated (I) *N. lactamica* (N.lac) Y92-1009, 14 isolates of wild type (WT) *N. meningitidis* (N.men). 19 isolates of inoculated *N. lactamica* remained unsequenced.

5.3.2 Phase variation

The study detected a total of 268 mutations across 35 volunteers carrying *N. lactamica***.** The most numerous and recurrent mutation type were phase variants and accounted for 71.6% (n=192/268) of the mutations observed. Phase variants are insertion/deletion mutations (indels) that affect change in repeat tracts that have been shown previously to affect gene expression via frameshift or changing transcriptional efficacy. A number of putatively phase variable regions based on long, homopolymeric tracts in *N. lactamica,* are detailed in **Appendix tables 0-9** and **0-10**. Out of this list of 98 candidate genes, this study found 11 contingency loci in which phase variation recurs longitudinally and across volunteers (**Figure 5-2)**. Genes affected by these mutations were involved in LOS subunit modification; *lgtG* (NEIS2011) *lgtC* (NEIS2154), pilin subunit modification; *pglA* (NEIS0213), *pglH* (NEIS0400), *pglL* (NEIS0539), iron acquisition: *hpuA* (NEIS1946) and type I restriction-modification: *hsdS* (NEIS0795). The hypothetical proteins were assigned putative function as glycosyl transferases ((hp and conserved-hp (NEIS1156)) while hp2 was a 100% identity, 45 amino acid matched fragment of a 470-amino acid, transferrin-binding like protein (accession: WP_003706847). While most of these mutations occurred in polymeric tracts located within coding sequence, 28/192 phase variable mutations were detected in an upstream, intergenic promotor region bordering the *fetA* gene(NEIS1963). As this homopolymeric tract was found between the -10 and -35 transcriptional start sites it could potentially modulate transcriptional efficacy.

All contingency loci with assigned NEIS loci were checked for presence and absence among other *N. lactamica* isolates in PubMLST Neisseria. The contingency loci were found widely disseminated among many *N. lactamica* isolates including among reference genomes *N. lactamica* ATCC_23970 (possessed all loci) and *N. lactamica* 020-06 (possessed all loci bar *lgtC*). The hypothetical proteins which were not assigned a NEIS locus were matched via BLASTn and the nr database firstly against *Neisseria lactamica* and secondly against the genus *Neisseria* overall. Among *N. lactamica* the first hypothetical protein was only found in *N. lactamica* Y92-1009 and at 90% identity among *N. meningitidis* genomes. The second hypothetical protein (hp2) wasn't found at high identity among *N. lactamica* or other members of the *Neisseria.* Neither hypothetical protein was found to contain conserved domains.

All contingency loci were found to contain homopolymeric G/ C tracts despite evidence of A/T tracts among the 98 candidate loci. All isolates contained changes in the repeat tracts of at least two of these contingency loci. Only phase variants with acceptable coverage (30X) were listed, as applied to all other mutations but there were numerous occurrences of low to medium coverage

mutation data (not shown, 5-20X) associated with type III res/mod subunit *modA*. This gene contained multimeric repeats of four ("-CGGA") bases.

Both the number of tract change deviations and phase ON and OFF expression states for each contingency locus was tested using contingency tables and fisher's twin tailed exact test (GraphPad calculator). This was used to determine if there was any statistically significant gain/loss in numbers of phase variants/ ON-OFF states longitudinally. There was a statistically significant increase in the numbers of tract changes of *lgtC* when comparing week 2 results with week 4 (p=0.0186) but no other significant data in any of the other genes across comparing any of the other time points. Phasotype profiles were calculated as described in section 5.2.4 were analysed to see if any SSR content was linked to expression profiles. There was no statistically significant association found between phasotype score and time point using fisher's twin tailed exact test

.

Figure 5-2 summary of mutations occurring among contingency loci during long term *N. lactamica* carriage

Gene	Mutation Type	Position	Mutations	Volunteers	reference tract	variant tract(s)	Description
c-hp	G tract	1483216	18	15	G10	G9, G10, G12	Conserved hypothetical protein (glycosyl transferase)
fetA	C tract	1674201	28	23	C10	C9, C11, C12	Iron regulated outer membrane protein
hp	G Tract	386523	4	3	G10	G9	Hypothetical protein (putative transferase activity)
hp2	G Tract	2004035	14	10	G9	G8	Hypothetical protein (putative transferrin binding activity)
hpuA	C Tract	617862	7	7	C10	C9, C11, C12	Hemoglobin-haptoglobin utilization protein A
hsDS	G tract	1124237	21	17	G9	G10, G11, G12	Type-1 restriction enzyme specificity protein MPN_089
lgtC	G Tract	659619	15	11	G13	G12	Lipooligosaccharide glycosyl transferase C
lgtG	G Tract	161341	36	23	G12	G11, G13	Lipooligosaccharide glycosyl transferase G
pglA	G Tract	256710	6	6	G11	G10, G12	Pilin glycosylase A
pglH	C tract	776735	24	16	C13	C10, C11, C12	Pilin glycosylase H
pglL	G Tract	742750	19	7	G11	G10, G12	Pilin glycosylase L

Combining the **mutation type** with the **reference tract** and **variant tract(s)** columns gives an indication of the results on a gene by gene basis. For example, the *lgtG* gene in the day zero inoculum (reference) contained a 12G tract or (“GGGGGGGGGGGG”). 36 isolates from 23 volunteers contained mutations in this gene. These mutations were insertions or deletions to the original 12G tract and they were G11 and G13

Figure 5-3 A summary of which of the putative contingency loci were phased ON in the reference and their functional GO

Gene	Mutation Type	Reference tract	Tract (Phase on)	% detected switched on
c-hp	G tract	G10	Unknown	N/A
fetA	C tract	C10	C11	15
hp	G Tract	G10	Unknown	N/A
hp2	G Tract	G9	Unknown	N/A
hpuA	G Tract	G10	G10	92
hsDS	G tract	G9	G7, G10	14
lgtC	G Tract	G13	G11, G14	0
lgtG	C Tract	G12	G11	25
pglA	G Tract	G11	G11	41
pglH	C tract	C13	C10, C13	75
pglL	G Tract	G11	G10, G13	2

Genes and repeat tracts are **highlighted** in cases where the **phase ON state is present in the reference and** therefore, introduced into volunteers in the ON phase. The column **time point Phase on %** is calculated for a gene by dividing the number of time points the gene was switched ON by the total number of time points sampled. This value was converted to a percentage. Because the conserved hypothetical protein (c-hp) and hypothetical proteins 1 and 2 have not been examined in the literature to date. It was not possible to look up a Phase ON/OFF homopolymeric tract value for them.

5.3.3 One example of mutations occurring across multiple volunteers in a non-phase variable, large hypothetical protein

Twelve mutations (out of a total of 268) were detected disseminated among eight of the volunteers and targeted a large (~10kb) hypothetical protein (hereafter dubbed "L-hp", **genbank accession number**: **ARB05049.1**) located at position 1,719,272 in the *N. lactamica* Y92-1009 genome (**Figure 5-4)**. This was the only coding sequence devoid of phase variable tracts found to be targeted by mutation across multiple volunteers. Conserved domain hits are shown alongside mutations which were applied *in silico* to the sequence of the L-hp to predict the impact on the translated protein (**Figure 5-4** + Companion **Figure 5-5)**. Five of these mutations persisted in volunteers 158 & 227 reducing protein size by 84-93%. Only two of twelve mutations observed were seen to occur within the protein domains and both targeted YadA-like C terminal regions. L-hp contained 8 conserved protein domains, these included; an extended signal peptide of the type V secretion system (pfam13018), a tryptophan-ring motif (pfam15401), supporting a HiaBD2 trimeric autotransporter adhesin (pfam15403), a Yad-A like, left-handed beta roll (cl17507) and four YadA-like C terminal regions (3 copies of cl27224 and 1 pfam03895). CDART analysis revealed significant homology with other *N. lactamica* strains (score =5/5, three sequences) and *Haemophilus influenzae* (score 4/5, thirteen sequences) and *N. cinerea* ATCC 14685 hep/hag repeat protein (score=4/5, two sequences) but only marginal or non-existent matches to other members of the *Neisseriacae*. This indicates that three other *N. lactamica* strains possess this protein with all five domains present but that fifteen sequences from both *H. influenzae* and *N. cinerea* genomes possess proteins with similar domain architecture. The protein content of the L-hp was matched via blastP against the nr and PubMLST *Neisseria* databases. The search found significant matches (>95% identity) only among other *N. lactamica* Y92-1009 genomes, suggesting the protein was strain specific. The GC content of the protein (44.6%) skewed negatively compared to GC content calculated for the *N. lactamica* Y92-1009 genome (52.2%.).

Figure 5-4 Domains, mutations & mutation areas of effect introduced in Large-Hypothetical Protein (ARB05049.1)

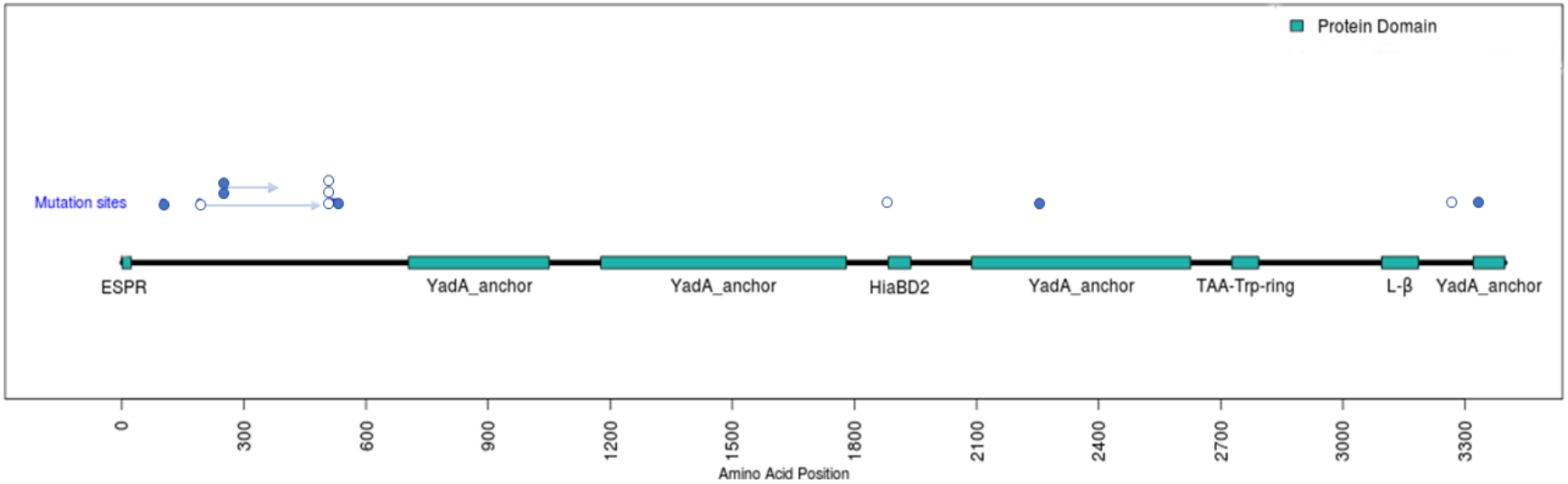


Figure 5-5 Tabulation of 12 mutations occurring among eight volunteers targeting the L-HP

volunteer	Timepoint (week)	mutation type	mutation annotation	position	protein size reduction (%)	Methionine start sites introduced
172	26	SNP	E102K (GAA→AAA)	coding (304/10197 nt)	0	0
172	16	deletion	Δ997 bp	coding (0574-1570/10197 nt)	94	1
158	8	deletion	Δ461 bp	coding (0747-1207/10197 nt)	93	0
158	26	deletion	Δ461 bp	coding (0747-1207/10197 nt)	93	0
227	4	deletion	Δ1 bp	coding (1539/10197 nt)	84	1
227	8	deletion	Δ1 bp	coding (1539/10197 nt)	84	1
227	16	deletion	Δ1 bp	coding (1539/10197 nt)	84	1
315	16	insertion	+G(GGGGG→GGGGGG)	coding (1577/10197 nt)	84	1
113	16	deletion	Δ1 bp (AAAAA→AAAA)	coding (5638/10197 nt)	45	1
221	26	deletion	Δ1 bp	coding (6762/10197 nt)	34	0
201	26	insertion	+GG	coding (9803/10197 nt)	2	5
104	16	deletion	Δ1 bp	coding (9990/10197 nt)	2	5

Caption for Figure 5-2 and Companion Table 5-3

Figure 5-2 displays the conserved domains detected in the L-hp. Above the domains are the mutations detected and their approximate range of effect in this protein. These were colour-coded blue and white to correspond with Table 5-3 and improve visual clarity.
Table 5-3 lists the detected mutations and displays the order they appear 5' to 3' across 10,197 nucleotides. In the **mutation type** column, **SNP** is short for single nucleotide polymorphism. In the **mutation annotation** column, the sole SNP that occurs has the amino acid change (E-K) caused by the bracketed, transitioning codons. Also in this column, "Δ" refers to a deletion event , "+" refers to an insertion event and the other bracketed insertion and deletion events were modifications to small polymeric tracts. The **protein size reduction** and **methionine start sites introduced** columns were generated with by applying the mutations to the protein coding sequence *in silico* and translating these.

5.3.4 Recurring SNPs

Our study defined a persistent/recurring mutation as one which was sampled in consecutive, successful isolations. A total of 8 SNPs (2 synonymous and 6 non-synonymous) disseminated among six of the volunteers were found to persist longitudinally as volunteer 264 was found to have three recurring SNPs in two hypothetical proteins and *atpC*. The recurring mutations were found in lipid-hydrolase protein *rssA*, oxidoreductase *nuoN*, energy synthase *atpC,* regulatory/recombination protein *recX* and CRISPR associated endonuclease *cas1*. These persistent mutations were tabulated in **Figure 5-6.** Of the six detected non-synonymous SNPs found to recur; four changed the side chain charge from positive, negative or neutral, one changed both side chain polarity and charge and one changed to an amino acid which had no effect on net side chain polarity and charge. No recurrent SNPS changed the amino acid side chain polarity from polar to non-polar and vice versa

Figure 5-6 Recurring mutations in long term study of *N. lactamica* Y92-1009 microevolution

volunteer	position	mutation	mutation type	Week 2	Week 4	Week 8	Week 16	Week 26	annotation	gene	description
63	760996	C→T	SNP: NS charge	0	1	ND	ND	1	D126N (GAT→AAT)	cas1	CRISPR-associated endonuclease Cas1
160	468817	A→C	SNP: NS charge	0	ND	1	ND	1	K131T (AAA→ACA)	rssA	NTE family protein RssA
190	505463	C→T	SNP: NS charge	1	1	ND	1	ND	H269Y (CAT→TAT)	nuoN	NADH-quinone oxidoreductase subunit N
264	1779889	G→A	SNP: NS charge	0	1	ND	1	ND	D62N (GAC→AAC)	hp	large multi domain protein
264	2123920	G→A	SNP: NS no change	1	1	ND	1	ND	A47V (GCG→GTG)	hp	Roadblock/LC7 domain protein
181	2126208	G→T	SNP: NS polarity, charge	0	1	ND	ND	1	A55E (GCG→GAG)	hp	hypothetical protein
88	1804594	G→A	Synonymous SNP	ND	1	ND	1	ND	L55L (TTG→TTA)	recX	Regulatory protein RecX
264	653511	G→A	Synonymous SNP	0	1	ND	1	ND	A93A (GCG→GCA)	atpC	ATP synthase epsilon chain

The **position** column shows the base the mutation affected (out of a potential 2,146,723 bases.) In the **mutation type** column **SNP** is short for a single nucleotide polymorphism; these can be synonymous or non-synonymous (**NS**). Non-synonymous SNPs have the potential to introduce amino acids which vary side chain polarity (**NS: polarity**), side chain charge (**NS: charge**), or maintain polarity and charge (**NS: no change**). The **week** columns show the number of detected mutations among samples isolated from those time-points where **1**= detected, **0**= undetected and **ND** = not done (either unsequenced or unisolated). In the **annotation** column the unbracketed letters and numbers refer to amino acid codes and the position of the where substituted amino acid is found in the protein translation of the gene. The bracketed letters describe the codon change that led to this outcome. The **gene** and **description** column list information on the gene the mutation occurred in.

5.3.5 Transient mutations

Transient mutations in non-phase varying genes were the second most abundant mutations (52/268). These occurred at a specific time point within a specific volunteer and were not detected again. These comprised of 15 synonymous SNPs, 3 insertions, 2 multiple base substitutions and 3 deletion events (**Figure 5-7**) in addition to 29 non-synonymous SNPs (**Figure 5-8**). The amino acid substitutions introduced by the non-synonymous SNPs were found to alter side chain polarity in 14 mutations, alter side chain charge in 2 mutations, alter both polarity and charge in 3 mutations, maintain the initial side chain and polarity in 7 mutations and introduce premature stop codons in 3 mutations.

Figure 5-7 Transient mutations (Part 1 of 2): Synonymous SNPs, multiple base substitutions, insertions and deletions

volunteer	position	mutation	Mutation Type	Week 2	Week 4	Week 8	Week 16	Week 26	annotation	gene	description
260	31096	G→A	SNP Synonymous	0	0	ND	1	0	G126G (GGC→GGT)	mraY	Phospho-N-acetylmuramoyl-pentapeptide- transferase
72	464946	C→T	SNP Synonymous	ND	1	ND	0	0	P241P (CCG→CCA)	PROKKA_00487	hypothetical protein
166	528744	G→A	SNP Synonymous	0	ND	ND	ND	1	P536P (CCG→CCA)	lptD	LPS-assembly protein LptD precursor
313	550170	T→C	SNP Synonymous	0	ND	ND	1	0	L86L (CTA→CTG)	PROKKA_00561	Transposase DDE domain protein
82	689859	C→T	SNP Synonymous	0	0	ND	ND	1	I226I (ATC→ATT)	leuS	Leucine--tRNA ligase
315	706251	A→G	SNP Synonymous	0	ND	ND	1	ND	V47V (GTT→GTC)	fda	Fructose-bisphosphate aldolase
264	726291	G→A	SNP Synonymous	0	0	ND	1	ND	I71I (ATC→ATT)	PROKKA_00729	MtN3/saliva family protein
118	908306	G→A	SNP Synonymous	ND	ND	ND	ND	1	I85I (ATC→ATT)	PROKKA_00895	HIT-like protein
166	1378450	C→T	SNP Synonymous	0	ND	ND	ND	1	G42G (GGC→GGT)	PROKKA_01375	Phage terminase large subunit
221	1550904	C→T	SNP Synonymous	1	ND	ND	0	0	R121R (CGC→CGT)	mfd	Transcription-repair-coupling factor
113	1592247	G→A	SNP Synonymous	0	0	ND	0	1	R62R (CGC→CGT)	rpsR	30S ribosomal protein S18
227	1599663	G→A	SNP Synonymous	0	0	0	0	1	A360A (GCG→GCA)	hcpC_2	Putative beta-lactamase HcpC precursor
181	1765065	C→T	SNP Synonymous	0	0	ND	ND	1	D2011D (GAC→GAT)	fhaB_3	Filamentous hemagglutinin
113	2041277	G→T	SNP Synonymous	0	0	ND	0	1	G76G (GGC→GGA)	PROKKA_01964	holo-(acyl carrier protein) synthase 2
227	2060759	G→A	SNP Synonymous	0	0	1	0	0	P58P (CCG→CCA)	cirA	Colicin I receptor precursor
213	1409564	2 bp→TT	multi base substitution	1	ND	ND	ND	ND	coding (335-336/849 nt)	hp	hypothetical protein
213	1409567	4 bp→TCCG	multi base substitution	1	ND	ND	ND	ND	coding (330-333/849 nt)	hp	hypothetical protein
170	1821617	+A	insertion	0	0	0	1	0	coding (198/207 nt)	hp	hypothetical protein
170	2126153	+A	insertion	1	0	0	0	0	coding (219/1098 nt)	hp	hypothetical protein
222	2127038	+A	insertion	0	1	ND	ND	ND	coding (772/1419 nt)	hp	hypothetical protein
172	904196	Δ615 bp	deletion	ND	0	ND	1	0	coding (690-1304/1815 nt)	PROKKA_00889	O-Antigen ligase
290	2009491	Δ2 bp	deletion	0	0	ND	0	1	coding (218-219/642 nt)	tesA	Acyl-CoA thioesterase I precursor
227	2125964	Δ2 bp	deletion	0	0	0	0	1	coding (407-408/1098 nt)	hp	hypothetical protein

The **position** column shows the base the mutation affected (out of a potential 2,146,723 bases.) In the **mutation type** column **SNP** is short for a single nucleotide polymorphism. The **week** columns show the number of detected mutations among samples isolated from those time-points where **1**= detected, **0**= undetected and **ND** = not done (either unsequenced or unisolated). In the **annotation** column the unbracketed letters and numbers refer to amino acid codes and the position of the where substituted amino acid is found in the protein translation of the gene. The bracketed letters describe the codon change that led to this outcome. The **gene** and **description** column list information on the gene the mutation occurred in.

Figure 5-8 Transient mutations (Part 2 of 2): Non-Synonymous SNPs

volunteer	position	mutation	Mutation Type	Week 2	Week 4	Week 8	Week 16	Week 26	annotation	gene	description
227	1213958	C→T	SNP NS: *	0	0	1	0	0	Q117* (CAG→TAG)	gltA	Citrate synthase
260	2127062	G→A	SNP NS: *	0	0	ND	1	0	Q250* (CAA→TAA)	PROKKA_02039	hypothetical protein
158	904781	C→T	SNP NS: *	0	0	ND	ND	1	W240* (TGG→TAG)	PROKKA_00889	O-Antigen ligase
222	776648	T→C	SNP NS: charge	0	1	ND	ND	ND	K236E (AAA→GAA)	PROKKA_00768	Glycosyl transferases group 1
190	1473005	C→T	SNP NS: charge	0	0	ND	1	ND	R81Q (CGG→CAG)	mobA	Molybdenum cofactor guanylyltransferase
264	658669	C→T	SNP NS: no change	0	0	ND	1	ND	A24V (GCG→GTG)	lex1_1	Lipooligosaccharide biosynthesis protein lex-1
313	817659	C→T	SNP NS: no change	0	ND	ND	0	1	A87V (GCC→GTC)	queC	7-cyano-7-deazaguanine synthase
118	69679	C→A	SNP NS: no change	ND	ND	ND	ND	1	L170F (TTG→TTT)	rmpM	Outer membrane protein class 4 precursor
172	499696	G→T	SNP NS: no change	ND	0	ND	1	0	L310F (TTG→TTT)	nuoL	NADH-quinone oxidoreductase subunit L
160	1806920	G→A	SNP NS: no change	0	ND	0	ND	1	P94L (CCT→CTT)	nlpD	Murein hydrolase activator NlpD precursor
312	867145	C→T	SNP NS: no change	ND	0	ND	1	ND	S88N (AGC→AAC)	apbE_1	Thiamine biosynthesis lipoprotein ApbE precursor
221	293046	T→G	SNP NS: no change	1	ND	ND	0	0	V227G (GTT→GGT)	mnmG	tRNA uridine 5-carboxymethylaminomethyl modification enzyme MnmG
166	895514	C→A	SNP NS: polarity	0	ND	ND	ND	1	A111S (GCC→TCC)	mntB	Manganese transport system membrane protein MntB
227	604496	G→A	SNP NS: polarity	0	0	0	1	0	A516T (GCG→ACG)	polA	DNA polymerase I
279	396787	C→T	SNP NS: polarity	0	1	ND	ND	ND	D538N (GAT→AAT)	PROKKA_00421	OPT oligopeptide transporter protein
309	1481653	G→A	SNP NS: polarity	ND	1	ND	ND	ND	G112S (GGT→AGT)	PROKKA_01487	hypothetical protein
172	793934	C→T	SNP NS: polarity	ND	0	ND	1	0	G202S (GGC→AGC)	engB_2	putative GTP-binding protein EngB
227	803071	C→T	SNP NS: polarity	1	0	0	0	0	G261S (GGC→AGC)	dacB	D-alanyl-D-alanine carboxypeptidase DacB precursor
158	788120	C→T	SNP NS: polarity	0	0	ND	ND	1	G55S (GGC→AGC)	pilQ	Type IV pilus biogenesis and competence protein PilQ precursor
72	1027495	G→A	SNP NS: polarity	ND	1	ND	0	0	G69S (GGC→AGC)	ihfA	Integration host factor subunit alpha
222	262313	G→A	SNP NS: polarity	0	1	ND	ND	ND	G91S (GGC→AGC)	pilE_1	Fimbrial protein precursor
315	1921366	C→T	SNP NS: polarity	0	ND	ND	1	ND	S172F (TCC→TTC)	fixL	Sensor protein FixL
190	230281	A→G	SNP NS: polarity	0	0	ND	1	ND	T12A (ACC→GCC)	PROKKA_00244	hypothetical protein
158	1645804	G→A	SNP NS: polarity	0	1	ND	ND	0	T261M (ACG→ATG)	xseA	Exodeoxyribonuclease 7 large subunit
72	1685720	G→A	SNP NS: polarity	ND	0	ND	0	1	T434I (ACA→ATA)	PROKKA_01654	hemagglutinin
72	1226461	T→C	SNP NS: polarity	ND	0	ND	1	0	T869A (ACC→GCC)	uvrA	UvrABC system protein A
313	550174	C→T	SNP NS: polarity, charge	0	ND	ND	1	0	G85D (GGT→GAT)	PROKKA_00561	Transposase DDE domain protein
201	1150450	T→A	SNP NS: polarity, charge	0	0	ND	0	1	K259M (AAG→ATG)	speE	Spermidine synthase
63	382198	C→T	SNP NS: polarity, charge	0	0	ND	ND	1	R215C (CGC→TGC)	pykA	Pyruvate kinase II

Caption for Figure 5-8

The **position** column shows the base the mutation affected (out of a potential 2,146,723 bases.) In the **mutation type** column **SNP** is short for a single nucleotide polymorphism; these can be synonymous or non-synonymous (**NS**). Non-synonymous SNPs have the potential to introduce amino acids which vary side chain polarity (**NS: polarity**), side chain charge (**NS: charge**), maintain polarity and charge (**NS: no change**) or introduce a frameshifting, premature stop codon into the ORF **(NS: *)**. The **week** columns show the number of detected mutations among samples isolated from those time-points where **1**= detected, **0**= undetected and **ND** = not done. In the **annotation** column the unbracketed letters and numbers refer to amino acid codes and the position of the where substituted amino acid is found in the protein translation of the gene. The bracketed letters describe the codon change that led to this outcome. The **gene** and **description** column list information on the gene the mutation occurred in.

5.3.6 Mutation rate estimates for solo-carrying vs co-carrying volunteers

SNP detection was used to infer a mutation rate in terms of SNPs detected per site per year. Sixty-four SNPs were detected among 27 volunteers carrying only *N. lactamica* Y92-1009. This yielded a mutation rate of $1.57E^{-06}$ SNPs per site per year. Seven SNPs were detected among 7 volunteers co-colonised with *N. meningitidis* and were used to infer a mutation rate of $9.32E^{-07}$ SNPs per site per year. Therefore, no evidence of a higher mutation rate was observed in volunteers co-carrying the inoculated *N. lactamica* and strains of *N. meningitidis.*

5.4 Discussion

Longitudinal mutation analysis was used to accurately monitor minute base changes to the genome of *N. lactamica* Y92-109 across multiple *in vivo* participants over the course of 26 weeks. This analysis revealed no consistent patterns of adaptive mutation transitioning from broth culture to human nasopharynx but did demonstrate an abundance of modifications to repeat tracts in genes associated with phase variation.

5.4.1 Phase variation

A number of putatively phase variable genes (~100 genes) based on homopolymeric tract possession in *N. lactamica* Y92-1009 genome were discovered. This study found 11 genes in which phase variation recurred longitudinally and between volunteers. Based on the results of previous studies identifying genes undergoing phase variation in *Neisseria* (Snyder and Saunders, 2006; Marri *et al.*, 2010) this study found a number of genes demonstrating phase variation in the *N. lactamica* species for the first time. These include *pglAHL* and *lgtCG* genes *(*pilin and LOS subunit modifiers) and *hsdS* the NGoAV1 type I restriction apparatus. Taken together, the hypermutation observed among the contingency loci would allow the commensal *Neisseria* to demonstrate a broad range of stochastically varying phenotypes to allow portions of a given population to survive in response to malignant host or environmental pressures (Palmer *et al.*, 2013). Phase variation has recently been demonstrated to correlate positively with the increased tract sizes (Alamro *et al.*, 2014). This finding was confirmed in our study as no shorter tracts under 9 bp demonstrated any variation. Alarmo *et al* studied phase variation in eight loci; outer membrane porin *porA*, opacity protein *opc*, haemoglobin receptor *hmbR*, adhesin *nadA*, autotransporters *mspA* and *nalP* in addition to two loci detected to undergo phase variation in this study *fetA* and *hpuA.* In their findings they noted that phase variable genes were driven into lower expression states (phase OFF) during the course of persistent carriage. This was not a finding that is mirrored in our data as phase variations were observed in higher numbers (but not enough to be statistically significant). This may be due to the specific genes examined or the species-level difference in pathogenic potential and lifestyle between *Neisseria meningitidis* and *N. lactamica* supported by the roles of structures like type IV pili (Trivedi, Tang and Exley, 2011) and polysaccharide capsules(Hill *et al.*, 2010).

5.4.2 SNPs and microevolution

For the most part, the distribution of SNPs among was found to be unequal among the volunteers of our study; with no patterns of general adaptive evolution across all isolates emerging as seen in other bacteria (Marvig *et al.*, 2015) or within microbiomes (Zhao *et al.*, 2017). A far greater number of SNPs detected were non-synonymous, either altering translation through amino acid modification or halting it completely by the introduction of a stop codon. A population generally acts to remove unbeneficial mutations from its genetic repertoire (Hughes *et al.*, 2008), and this was reflected in our data by the fact that most SNPs were transient, non-recurrent events. Despite this, some non-synonymous SNPs were seen to persist longitudinally in some volunteers. Suggesting that stochastic mutations still have some chance of effecting change that is either beneficial or evolutionarily neutral. With this in mind, there was some evidence of minimal adaptive evolution occurring in one volunteer as a synonymous substitution was shown to be absent in initial sampling but become fixed during the course of the longitudinal study. Synonymous substitutions represent neutral changes to a bacterium, preserving amino acid structure and thus are more likely to remain unchanged once established (Nielsen, 2005; Cambray and Mazel, 2008). The gene affected by the synonymous substitution in question, *recX*, facilitates RecA in providing pilus antigenic variation via homologous recombination. This has been demonstrated in gonococci (Greunig *et al.,* 2010.) Although It is currently unknown what effect this may have on a species such as *N. lactamica* that do not contain any *pilS* genes and therefore do not participate in pilin antigenic variation via the mechanism of homologous recombination. A substitution rate estimate of 1.91×10^{-7} per genetic site per year based on the only synonymous SNPs found to recur in the same volunteer was inferred and contrasted with results from a study using BEAST analysis to calculate mutation rate among gonococci (Grad *et al.*, 2014). The study found that 236 gonococcal isolates microevolved at a mean rate of $2{\cdot}5 \times 10^{-6}$ substitutions per site per year, a factor of ten less than our study.

Among volunteers carrying only *N. lactamica* Y92-1009, a mutation rate estimate of $1.57E^{-06}$ SNPs per site per year was higher than that found among volunteers co-colonised with *N. meningitidis* ($9.32E^{-07}$ SNPs per site per year). It was hypothesized that *N. lactamica* colonising in the presence of the related pathobiont would elicit a difference in mutation type or rate on the *N. lactamica* genome due to a history of interspecific gene transfer between the species (Linz *et al.*, 2000; Corander *et al.*, 2012). This mutation difference would potentially manifest as multiple SNPs/ multiple base substitution within a gene compared to the reference genome. However, this did not prove to be the case.

5.4.3 Large hypothetical protein

Only one non-phase variable gene was found to be the target of multiple mutations across multiple volunteers. This gene was a large, (~10,000 bp) multi domain containing hypothetical protein (L-hp). An *in-silico* simulation of how these mutations may affect protein coding revealed frame-shifted introductions of interproteomic methionine residues and overall protein size reductions. The low GC content of this gene suggests that it was horizontally acquired by *N. lactamica* from another nasopharyngeal coloniser; most likely *H. influenzae* based on the results from CDART and the possession of a *H.influenzae* autotransporter domain . In the literature, there are other examples of genes postulated to have been acquired by *N. lactamica* 020-06 from *H. influenza,* including phosphroylcholine biogenesis genes *licABCD*, ATPase gene *slpA*, protease gene *slpB* and putative surface fibril protein NLA12600 (Bennett *et al.*, 2010). Horizontal gene transfer has also been proposed to have previously occurred between the *Haemophilus* and *Neisseria* families (Kroll *et al.*, 1998; Schoen *et al.*, 2008). The presence of yadA -like genes and trimeric autoadhesins suggests this protein may have a role in initial attachment of the bacterium to a human host (Łyskowski, Leo and Goldman, 2011). But post-colonisation, *N. lactamica* Y92-1009 may not require the presence of this strain-specific gene any longer, possibly either allowing deleterious mutations to persist and lower the metabolic burden of expressing such a large protein or removing it as a potential site of host-recognition. Without further examination of the expression and biological activity of this product it is impossible to predict what allows the accumulation of these deleterious mutations.

Limitations

The method of counting phase variable loci from WGS data necessitates read counting. A number of potentially false-positive results were detected for *modA*, a gene known to be phase variable and highly influential among *Neisseria* as a global regulator of gene expression (Tan *et al.*, 2016). This may be due partly to the length and nature of the repeat but also, to the algorithm used to detect contingency loci. Tract lengths varied from 13-18 repeats of 4 base pairs (CGGA-CGGA) which led to the exclusion of most reads spanning this area and biasing the statistical determination. While short-read data is highly accurate per base and has been successfully used in this study to determine mutations, including changes in tract length, this method may be unsuitable for evaluating longer multimeric repeat regions. A cost-effective alternative may be to amplify the gene from every isolate via PCR and resequence using a long-read sequencing method with high per base accuracy (i.e. Sanger sequencing.). The use of SMRT cell sequencing technology paralleled with short read technology to re-create accurate assemblies of every isolate may also be an option that is more viable in the future.

Chapter 5:

Taken as a whole, this chapter demonstrates that inoculated *N. lactamica* Y92-1009 possess a stable genome over 6 months of *in vivo* carriage showing no signs of convergent evolution and minimal signs of within-host adaptation. This is probably due to the mechanism of phase variation which allows the bacterium to generate stochastic diversity primarily in loci associated with host-pathogen interaction and adapt *ad hominem*.

Chapter 6: Recombination among inoculated and wild type *N. lactamica* co-colonised with *N. meningitidis*

6.1 Introduction

A significant bulk of this body of work has dealt with the implications of mutations as a source of genetic variation in bacteria. Homologous recombination can also act as another driver of adaptive evolution, particularly for genetically-promiscuous organisms like *Neisseria* which are surrounded by an abundance of donor DNA and mechanisms facilitating exogenous uptake (DUS sequences) as well as genetic copy-editing potential. Interspecific homologous recombination has been previously shown to occur between the pathogenic and commensal *Neisseria* (Didelot and Maiden, 2010) but almost never in the context of longitudinal co-carriage (Mulhall *et al.*, 2016).

Techniques such as genome hybridisation microarrays and MLST have previously been used to detect and quantify recombination pressure in *Neisseria* datasets. In 2011, a dataset comprising of 29 meningococcal isolates from the same country of origin (Germany) and including reference genomes of *N. meningitidis* strains FAM18, MC58 & Z2491 was used to detect recombination pressure in the species (Joseph *et al.*, 2011). These 29 genomes covered over 98% of the observed genetic diversity of the meningococcus at the time of publication and included all major hyperinvasive lineages associated with meningococcal morbidity. The study, found no significant association between the presence of virulence-associated genes and the source of the strain. 64 of the 98 virulence genes present in the microarray were present in all meningococcal strains while the remaining 34 genes were differentially distributed between hyperinvasive and carriage isolated meningococci, mirroring the finding that genetic content doesn't vary greatly among the *Neisseria* (Marri *et al*., 2010, Bennet *et al*., 2010.) A counter balance exists among *Neisseria meningitidis,* an organism both structured in specific clades but able to diversify beyond a neutral model of evolution through substantial recombination events. The chromosomal rearrangements and gene conversion introduced by homologous recombination are tightly regulated by multiple restriction modification systems and are more likely to occur within clades than between them (Budroni *et al.*, 2011).

Genetic variation coupled with a high reproductive turnover is essential for a bacterial pathogen to survive and thrive in the environment it inhabits. The patterns of this variation observed through analysis by NGS technologies can broadly categorise bacteria of interest into three groups:

1) Monomorphic or "clonal" bacteria maintain very low levels of sequence diversity amongst their descendants (Achtman and Wagner, 2008). Examples of this group include *Yersinia pestis, Bacillus anthracis* and *Mycobacterium tuberculosis.* Very low levels of SNPs and point mutations have been observed in the comparative genomics studies of these organisms and (in the case of the latter organism) until recently, very little or no recombination has been ascribed to them (Namouchi *et al.*, 2012). Many of the most virulent pathogens known, fall into this category.
2) Clonal bacteria that behave like the group above but can horizontally exchange large sections of sequences through the rare dissemination of mobile genetic elements. Examples of this group include *Staphylococcus aureus* and *Salmonella enterica* (Holt *et al.*, 2008).
3) Naturally transformable bacteria possess systems that actively import DNA from the environment. This DNA is then integrated into the pathogen's genome in a region that shares high homology with the sequence. Examples of this group include *Haemophilus influenzae, Streptococcus pneumoniae* and *Neisseria meningitidis*. Variation in naturally transformable species can be the result of a complex combination of point mutation, other mechanisms of horizontal gene transfer and homologous recombination (Croucher *et al.*, 2014)).

Seven meningococcus-carrying volunteers from the human challenge study described in this thesis were inoculated with commensal *Neisseria lactamica Y92-1009*. Two volunteers who were not inoculated instead presented with wild type *N. lactamica,* naturally co-carried with meningococci. The recombination workflow described here was used to determine whether there was any evidence of horizontal gene transfer via the mechanism of homologous recombination occurring among the co-carried meningococcal and *N. lactamica* strains sampled. The results of this chapter will be used to inform future carriage studies and answer the following research questions:
1) Is there enhanced recombination measured in co-colonised *N.lac* Y92-1009 inoculants compared to solo colonising isolates of the same strain?
2) Is there enhanced recombination observed between wild type *N.lactamica* isolates and meningococci compared to artificially inoculated, co-colonised *N.lac* Y92-1009 isolates?
3) If allelic change mediated by homologous recombination occurs in any longitudinal *N.lactamica or N. meningitidis* samples, is this a change from a previous allele or is this an importation from the co-colonising species?

6.2 Methods

The following section defines a pipeline devised to utilize output from the genome comparator alignment tool (available on https://pubmlst.org/neisseria/) as input into the recombination detection program ClonalFrameML. In a broader context, this pipeline will work for any bacterial alignments which are capable of being converted into .fasta format and is outlined in **Figure 6-1**.

6.2.1 Screening for paralogous loci (BLAT) and generating NEIS loci list for alignment

Bigsdb analyses (Jolley and Maiden, 2010) hosted on PubMLST Neisseria use a gene by gene approach. Therefore, there are gene specific assemblies (NEIS loci) containing sequence information for every gene able to be sequenced from every isolate in this study (Bratcher *et al.*, 2014). In order to avoid subsequent false positives in recombination detection, it was important to remove any paralogues at the outset of the experiment. BLAT (Kent, 2002) was used to perform an "all vs all" comparison of gene sequences in order to determine and remove the existence of paralogues among every sequence type of every *Neisseria* species examined in this chapter

A .xmfa file outputted by genome comparator containing all of the annotated genes and sequence information (All non-absent NEIS loci from NEIS001-NEIS2494) was used to do this. Using BLAT via default parameters for DNA sequences, the xmfa file was screened for the presence of paralogous loci by using itself as both query and subject. The criteria for paralogous loci with closely matching homology (>95% homology) and large amounts of identical sequence. The BLAT output.psl file was checked and any paralogous genes were excluded from a list of NEIS loci to be aligned. This process was repeated for every sequence type of *N. lactamica* and *N. meningitidis* examined in this chapter (n=5 ST's examined). A best representative genome for every sequence type was chosen as input based on possession of the highest N50 value and the lowest number of assembly contigs.

6.2.2 Sequencing, Isolate hosting, alignment and allelic analysis (BIGSdb)

The experimental isolates were hosted on PubMLST Neisseria. Any sequences corresponding to loci on the database were manually curated and tagged. Isolates were analysed using BIGSdb's genome comparator tool and aligned via MAFFT (Katoh and Standley, 2013). Using a list of all present loci between NEIS001-NEIS2494, minus paralogues (detected as described above), the genome comparator tool was used to generate a coding sequence alignment of the loci using the MAFFT aligner under default parameters. The wild type isolates under analysis were subsetted to be aligned in the following way: a) wild type *N. lactamica* belonging to volunteer 36, b) wild type *N. meningitidis* belonging to volunteer 36, c) wild type *N. lactamica* belonging to volunteer 291 &

d) wild type *N. meningitidis* belonging to volunteer 291.

The inoculated *N. lactamica* Y92-1009 isolates were subsetted into: a) isolates from volunteers co-colonised with *N. meningitidis* and b) isolates from volunteers in which *N. lactamica* Y92-1009 was the only detected *Neisseria* species for the duration of the carriage study. The alignment outputted files in .fasta and .XMFA alignment formats.

Alignment correction and conversion (GBlocks 0.91b/EMBOSS v6.0)

The fasta alignment was trimmed to remove low confidence parts of the alignment (for example poorly aligned positions) using Gblocks (Castresana, 2000) under default settings. Automatically calculated block parameters were checked for accuracy (100%) before the ouput alignment was used. The alignment was converted from fasta to phylip format using the EMBOSS package's seqret tool (Peter Rice, Longden and Bleasby, 2000).

6.2.3 Maximum likelihood Tree Building (PhyML)

A maximum likelihood tree using was computed using a source code installation of PhyML using the HKY model (Guindon *et al.*, 2009) and the .phylip format alignment as input. The generated tree was in newick format and the transition/transversion values calculated were noted for future use as inputs for the "-kappa" parameter of ClonalFrameML.

6.2.4 Tree checking and editing (R package ape and phangorn)

The ape (Popescu, Huber and Paradis, 2012)and phangorn (Schliep, 2011) packages were downloaded and installed in R v 3.2.1. Newick trees from PhyML were converted to an object of class "phylo". The tree was tested to see if it contained any multichotomy among its branches using the "is.binary.tree" function. If any were found, the phylo object was resolved so that any multichotomies were converted into dichotonomies (binary format) in the order that they appear in the tree using the "Multi2di" function with the "random=TRUE" option. The phylo object was then converted back into a newick tree. The newick tree was edited in a text editor to remove any internal node labels.

6.2.5 Recombination analysis (ClonalframeML)

ClonalframeML (Didelot and Wilson, 2015) was used to perform the recombination analysis on the dataset. The .xmfa alignment file and maximum-likelihood newick tree were used as input files. The program was run per branch with the: "-xmfa_file" option, the "emsim 100"option to perform bootstrap iterations, and the "-kappa" option containing the transition/transversion ratio

calculated by PhyML. An R script provided by the authors was used to create a graphical output of the results.

Following the CFML run, the results in output "importation_status.txt" file were examined to find the positions of any potential recombinant regions. These regions were then compared to the input .xmfa file to ascertain the identity of the NEIS loci implicated. Using a xmfa file in this pipeline introduced 1000bp spacer regions between each NEIS locus. These were accounted for when inferring recombination results from the importation status output back to NEIS loci in the xmfa file.

6.2.6 Alignment visualization and locus homology analysis

In order to double-check that recombiant regions differed in terms of SNPs and not sequence absence. Alignments of loci identified as recombinant in their respective isolates were outputted by genome comparator to be visualized with MView (Brown, Leroy and Sander, 1998) .

6.2.7 Effect of recombination equation

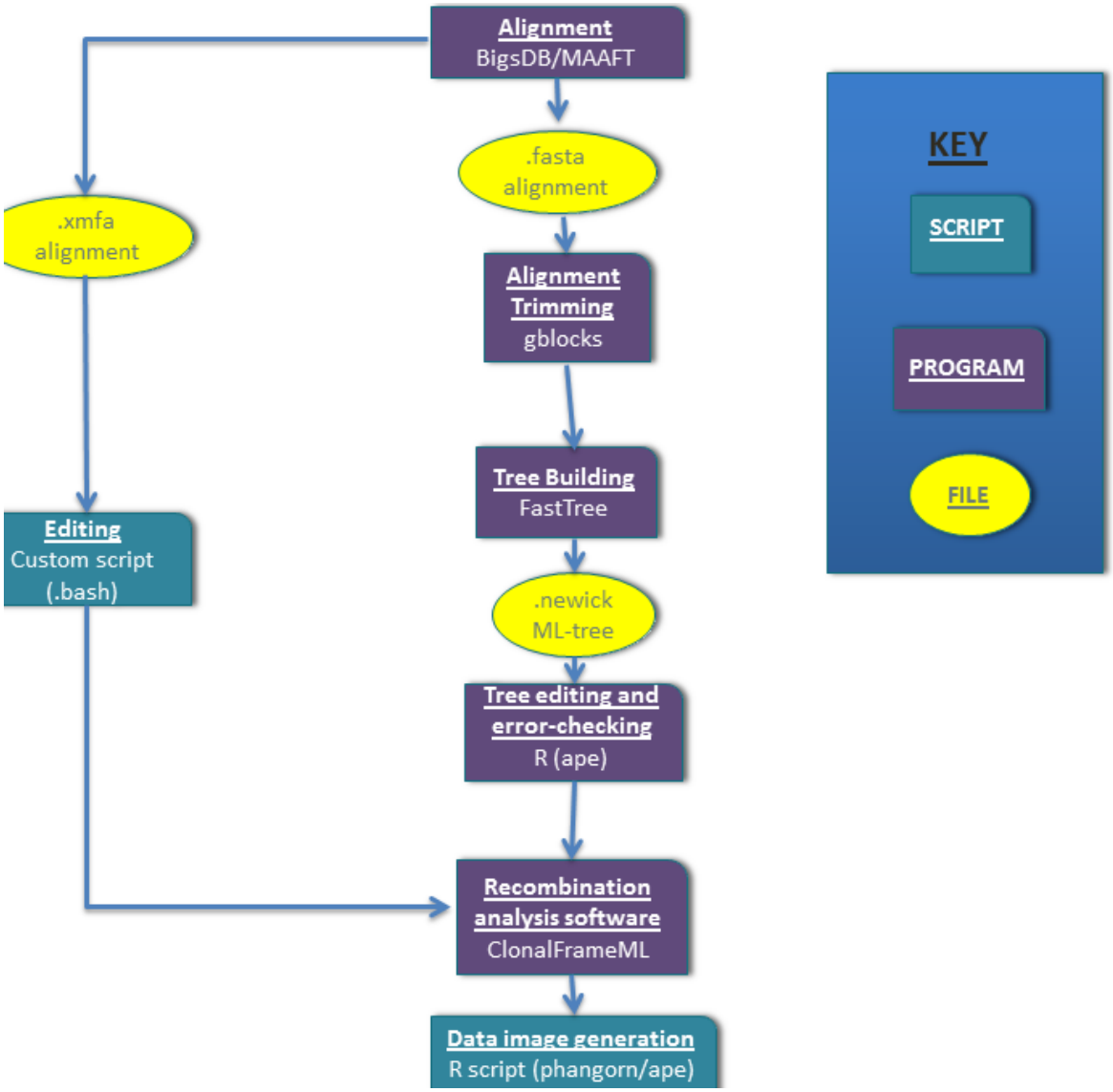
The effect of mutation (r/m) = the relative rate of recombination to mutation (R/theta) * mean importation DNA length (delta) * nu (mean divergence of DNA).
This means that the relative effect of recombination observed was greater where there wass a larger relative rate of recombination importing larger and more diverse (compared to what was there orginally) DNA sequences to/from the host organism.

This formula was used to calculate the relative rate of recombination for all *Neisseria* isolates.

6.2.8 DNA pattern motif searching

Where applicable, DNA repeat motifs were searched using the same method outlined in chapter 3 section **3.2.4.** In the four groups examined for repeat motifs (wild type *N. lactamica* and *N. meningitidis* from volunteers 36 and 291), the best quality genome assembly was chosen from all the isolates examined with regards to lowest contig size and highest N50 value.

Figure 6-1 An overview of the recombination detection workflow



6.3 Results

As discussed previously, there were a number of volunteers in the experimental challenge study who demonstrated co-colonization of *N. meningitidis* and *N. lactamica* detected over the range of 26 weeks. Most volunteers carried the inoculated strain (*N.lac* Y92-1009) and some volunteers co-carried the inoculum strain alongside wild type meningococci However two volunteers were found to carry wild type meningococci and wild type *N. lactamica*. I will first quantify the metrics of recombination detected among these six groups of species (1: Solo-carried *Neisseria lactamica* Y92-1009, 2: co-carried *N. lactamica* Y92-1009, 3-4: wild type, co-carried *N. lactamica* in volunteers 36 and 291, 5-6: wild type, co-carried *N. meningitidis* in volunteers 36 and 291) before describing the genes detected as undergoing recombination in each species, among each volunteer. Lastly, I will describe the disparity in repeat sequence content between species shown to undergo interspecific recombination and species in which it was not found.

6.3.1 Dataset description

In most of the successfully colonised volunteers described in the previous chapter, the inoculated *N. lactamica* was the sole *Neisseria* species detected during the 26 weeks of the experiment. However, although a rare event the inoculum strain was delivered to volunteers 64, 77, 104, 181 221, 255 & 279 who were either pre-colonised with *N. meningitidis* or became colonised with *N. meningitidis* (**Figure 6-2**). Appendix table **0-6** lists the meningococcal isolates and their PubMLST IDs. Appendixl tables **0-7** & **0-8** list the inoculated, N. *lactamica Y92-1009* isolates and their PubMLST IDs. It should be noted that there was one volunteer (309) who co-carried both organisms but was excluded from recombination analysis because of only one successful isolation of either organism. As such there would be no comparator genome to detect recombination using ClonalFrameML in either the *N. lactamica* Y92-1009 or *N. meningitidis* from this participant.

In addition to the above volunteers, there were two volunteers who presented with natural co-colonisation of wild type *N. lactamica* and *N. meningitidis.* Volunteers **36** and **291** did not receive experimental challenge with *N. lactamica* Y92-1009 and are naturally colonised by different sequence types of *N. lactamica* (**Figure 6-2**). **Figure 6-3** displays the PubMLST ID's and some assembly metrics of the wild type N. *lactamica* and *N. meningitidis* sequences.

Figure 6-2 The lactamica 2 dataset: co-colonised volunteers

Wild type?	Volunteer	Week 0	Week2	Week 4	Week 8	Week 16	Week 26	Lactamica ST
Yes	**36**	10457	10457				10457	ST-4192
Yes	**291**	41	41				41	ST-11524
No	**64**					839	839	ST-613
No	**77**					1423	1423	ST-613
No	**104**	1655						ST-613
No	**181**						2394	ST-613
No	**221**	466				466		ST-613
No	**255**	10478				213	213	ST-613
No	**279**	60					60	ST-613

This table is ordered numerically starting with the wild type *N. lactamica* co-colonisers (from volunteers 36 & 291, **Wild type?** =Yes) before listing the *N.lac* Y92-1009 (ST 613) artificial inoculants (volunteers 64-279, **Wild type?**=No). **Dark green** cells indicate sequenced *N.lactamica* isolates. **Light green** cells indicate isolation but not sequencing of *N. lactamica*. **Red** cells indicate the presence of a sequenced *N. meningitidis* isolate only at a given time point. **Purple Cells** indicate both *N. meningitidis* and *N. lactamica* isolation at a singular time point. A **white** cell indicates an absence of isolation of either organism. The **numbers within the cells** refer to **meningococcal sequence types**; the *N. lactamica* ST's are listed in the final column **(Lactamica ST).**

Figure 6-3 Wild Type *N. lactamica* and *N. meningitidis* isolates from volunteers 36 &291: PubMLST ID's, strain designations and Contigs

Wild type *spp.*	BIGSDB id	time point	volunteer	Capsule/ST/cc	contigs
N.lactamica	36855	week 0	36	ND/ ST-4192/(ccX)	78
N.lactamica	36862	week 02	36	ND/ ST-4192 (ccX)	104
N.lactamica	36874	week 04	36	ND/ ST-4192/(ccX)	84
N.lactamica	36908	week 08	36	ND/ ST-4192/(ccX)	95
N.lactamica	36946	week 16	36	ND/ST-4192/(ccX)	83
N.lactamica	36969	week 26	36	ND/ST-4192/(ccX)	85
N.lactamica	36856	week 02	291	ND/ ST-11524/(ccX)	97
N.lactamica	36869	week 04	291	ND/ ST-11524/(ccX)	98
N.lactamica	36938	week 16	291	ND/ ST-11524/(ccX)	113
N. meningitidis	26411	Week 0	36	E/ ST-10457/(cc60)	230
N. meningitidis	26297	Week 02	36	E/ ST-10457/(cc60)	202
N. meningitidis	26352	Week 26	36	E/ ST-10457/(cc60)	248
N. meningitidis	26463	Week 0	291	Cnl/ ST-41/(cc41/44)	176
N. meningitidis	26492	Week 02	291	Cnl/ ST-41 /(cc41/44)	137
N. meningitidis	26378	Week 26	291	Cnl/ST-41/ (cc41/44)	198

In the **Capsule/ST/cc** column **ND** refers to “Not Described” as *Neisseria lactamica spp.* do not possess a capsule. **Cnl** refers to a capsule null strain while **E** means that the meningococcal organism has an E-Type Capsule and is serogroup E. **ST** indicates sequence type and **cc** is clonal complex. Where a clonal complex has not been assigned the value **ccX** is given.

6.3.2 Recombination metrics

The effect of recombination leveraged against the effect of mutation (r/m) was calculated for all *N. lactamica* and *N. meningitidis* isolates. This information is displayed in **Figure 6-4.**

r/m values were higher for volunteers colonised with wild type *Neisseria spp.*as opposed to those artificially inoculated with *N. lactamica* Y92-1009 regardless of solo or co-carriage status. The highest detected r/m value (18.06) was in isolates recovered from volunteer 291 wild type *N. lactamica* (ST-11524). The effect of recombination seen among these isolates was ~5 fold greater than that observed among artificially inoculated *N. lactamica* and ~2 fold more than the mean r/m value observed among both groups of meningococcal isolates (r/m: 9.51= (10.69 + 8.32)/2))

Figure 6-4 the effect of recombination leveraged against the effect of mutation (r/m) and the parameters used to calculate this value averaged across all volunteers.

Colonisation	Organism	ST	Isolates	R/theta	Delta*	Nu	r/m
Sole *Neisseria*	*N. lactamica* (I) (Y92-1009)	613	97	0.48 (5.7 E^{-4})	35 (2.8 E^{-2})	0.22 (1.6E^{-5})	**3.61**
Co-colonised	*N. lactamica* (I) (Y92-1009)	613	19	0.40 (9.1 E^{-3})	138 (4.5 E^{-6})	0.11 (9.1 E^{-4})	**3.84**
	N. lactamica (WT)	11524	3	0.68 (3.0 E^{-2})	585 (2.1 E^{-7})	0.05 (6.4 E^{-6})	**18.06**
	N. meningitidis (WT)	41	3	0.38 (1.8 E^{-3})	403 (7.4 E^{-7})	0.07 (1.2 E^{-6})	**10.69**
	N. lactamica (WT)	4192	6	0.39 (9.1 E^{-3})	123 (3.5 E^{-6})	0.19 (1.4 E^{-4})	**9.21**
	N. meningitidis (WT)	10457	3	0.28 (1.2 E^{-2})	420 (9.4 E^{-7})	0.07 (4.3 E^{-5})	**8.32**

Relative recombination effect **(r/m)** of isolates in colonised volunteers and the parameters used to infer them. The table displays recombination parameters among wild type **(WT)** N. lactamica and N. meningitidis in addition to inoculated **(I)** N. lactamica Y92-1009 in solo and co-colonised volunteers. The parameters include the relative rate of recombination to mutation **(R/theta)**, the mean length of detected recombinant regions **(Delta)** and mean divergence of imported DNA from what was present before recombination **(Nu)**. Standard deviations are indicated for these values in the accompanying brackets (X.X E^{-X}). The parameters were multiplied together to generate the relative effect of recombination compared to the relative effect of mutation (r/m). The sequence type (**ST**) of each collection of organisms is also given.

*Delta is calculated by ClonalFrameML as 1/Delta. This has been resolved in the table to show clearly how r/m was calculated. Standard deviations for this value are for 1/Delta. This equation is described in **section 6.2.7**

6.3.3 Recombination in the artificially inoculated, solo *Neisseria spp.* coloniser *N. lactamica* Y92-1009

Despite being the dataset with the greatest number of isolates tested (n=97 isolates), solo-colonising *N. lactamica* Y92-1009 possessed the lowest r/m value among all *Neisseria spp.* tested. Only three importations were detected targeting three loci (out of a total 1595 loci) in *N. lactamica* Y92-1009 (**Figure 6-10**). These loci included; a ~885 nt, hypothetical protein (NEIS1708) detected in volunteer 166, a ~720 nt, putative mafS3 cassette (NEIS1794) detected in volunteer 133 and a ~550 nt, intracellcular septation (cell division) protein (NEIS1828) detected in volunteer 38.

6.3.4 Recombination in the artificially inoculated, co-colonised *N. lactamica* Y92-1009

Very little to potentially no importations were detected in volunteers 64, 77, 104, 181, 221, 255 and 279 co-carrying *N. lactamica* Y92-1009 and *N. meningitidis.* As shown in **Figure 6-5**, the average size and number of the detectable fragments were low and any large importations (recombinant regions, >150bp) were always inferred from loci that were later found to be incomplete on inspection with mView. Running genome comparator between the meningococci and *N. lactamica* co-carried isolates of these volunteers showed no evidence of any recombinant allele sharing (n=1595 total loci)

Figure 6-5 Average size and range of recombined fragments (importations) given in base pairs for each co-colonised *N. lactamica* Y92-1009 per volunteer

Strain	Volunteer	Importations	Average Size (bp)	Range (bp)	r/m
ST-613 (Y92-1009)	64	8	28.5	1-153	2.06
ST-613 (Y92-1009)	77	10	6.7	1-18	7.32
ST-613 (Y92-1009)	104	11	9.1	2-19	7.64
ST-613 (Y92-1009)	181	7	31.8	2-153	1.95
ST-613 (Y92-1009)	221	10	8.4	1-18	3.10
ST-613 (Y92-1009)	255	4	37.25	5-119	1.28
ST-613 (Y92-1009)	279	7	5.7	1-13	3.55

6.3.5 Homologous recombination in wild type *N. lactamica* (Volunteer 36)

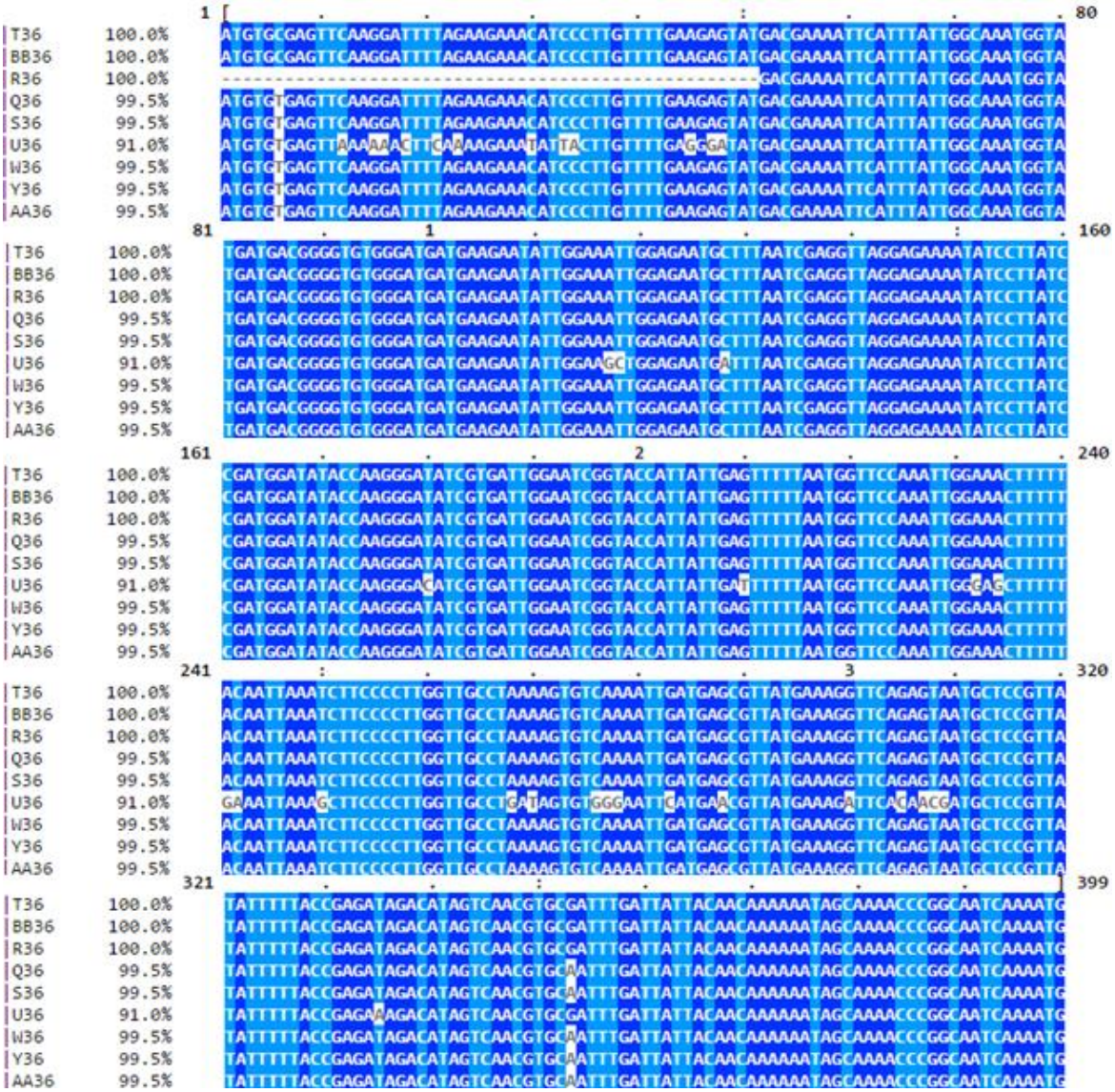
In volunteer 36, a 41bp importation was detected at the locus NEIS1795 (mafl immunity gene; type o2MGI-2, 399nt size) on the week 4 isolate. As shown in **Figure 6-6** this prompted a change in the allele (new#2 to new#3) for this time point only, as the allele reverted back to new#2 and stayed that way for the duration of the study. Visualising the locus alignment on MView (**Figure 6-7**) revealed that the week 0 meningococcal isolate was missing the first 50bp of the sequence but otherwise was 100% homologous to the sequence observed at week 2 and week 26. The new#2 lactamica allele observed at 5 separate time points was different by 2bp (position 7 and position 353) to the meningococcal allele. The new #3 allele from week 4 *N. lactamica* (U36 in figure) was more divergent, displaying variation at 37 sites from the other *Neisseria spp.* alleles for this locus. Looking at the figure, the variant sites from week 4 *N. lactamica* isolate were interspersed, irregularly along the length of the gene.

Figure 6-6 Allelic variation of recombinant locus in volunteer 36.

Locus	N.men Week 0	N.men Week 2	N.men Week 26	N.lac Week 0	N.lac Week 2	N.lac Week 4	N.lac Week 8	N.lac Week 16	N.lac Week 26
NEIS1795	I	new#1	new#1	new#2	new#2	new#3	new#2	new#2	new#2

This highlights changes in allelic variation across one locus (NEIS1795) in *N. lactamica and N.men* carried isolates for one volunteer (volunteer 36.) **Dark green cells** indicate an incomplete allele, **cyan/dark blue/purple** cells differentiate the new#1, #2 and #3 alleles from each other

Figure 6-7 Multiple sequence alignment of locus NEIS01795 shows allelic disparity among co-colonising *Neisseria spp.* in volunteer 36.



Isolates **R**, **T** and **BB** correspond to **meningococci** sampled at weeks **0**, **4** and **26**. Isolates **Q, S, U, W, Y** and **AA** correspond to ***N. lactamica*** isolated at weeks **0, 2, 4, 8, 16** and **26** respectively.

6.3.6 Homologous Recombination in wild type *N. lactamica* (Volunteer 291)

ClonalframeML identified 7 regions of importation ranging in size from 149 bp (NEIS2437) to 4106 bp (which encompassed three loci; NEIS0069, NEIS0071 and NEIS0072) these were all detected in the *Neisseria lactamica* isolate sampled at week 16 of the study. The pubMLST genome comparator and locus explorer tools were used to determine the level of allelic similarity between any genes found in the regions of importation. Out of the 1627 total loci aligned for use in this analysis, 10 were discovered in recombinant regions in volunteer 291. Of these 10 recombinant loci, 4 displayed allelic variation in *N. lactamica* independent of alleles displayed by their co-colonising meningococcal isolates (**Figure 6-8**). These loci included; sodium glutamate symport carrier protein (NEIS0069), an outer membrane transport protein (NEIS0073), *pgi2:* a glycolysis pathway enzyme (NEIS1837) and a putative lipoprotein (NEIS2437). The difference between the alleles that changed for these four loci was always equal to or greater than 28 polymorphic sites.

Furthermore, there was evidence of a co-carried *N. lactamica* isolate sharing exactly the same allele as *N. meningitidis* in 6 of the 10 recombinant loci. Of these 6 loci, four displayed examples of *N. lactamica* isolates sampled at weeks 2 and 4 sharing the same allele found in the meningococci while the week 16 *N. lactamica* isolate showed a divergence (**Figure 6-9**). These loci included; a putative lipoprotein (NEIS0071), hypothetical proteins (NEIS0072, NEIS2514) and *lst:* a siayltransferase (NEIS0899). This pattern of allelic variation was found to be reversed for the final two loci as week 2 and 4 *N. lactamica* isolates displayed alleles different from those of the meningococcal isolates while the final week 16 isolate shared the same allele (**Figure 6-9**). These loci included; an enoyl acyl carrier reductase (NEIS1834) and a 2,3,4,5-tetrahydropyridine-2,6-carboxylate N-succinyl transferase (NEIS1835). The difference between the alleles that changed for 5 of these 6 was always equal to or greater than 19 polymorphic sites. The allelic variation in NEIS0072 was only different by 3 polymorphic sites but these were found to be interspersed along the length of the coding sequence.

Figure 6-8 Allelic change detected within volunteer 291 *N. lactamica* isolates occurring independently of *N. meningitidis* isolates

Locus	Locus annotation and size (nt)	N.men Week 0	N.men Week 2	N.men Week 26	N.lac Week 2	N.lac Week 4	N.lac Week 16
NEIS0069	sodium/glutamate symport carrier (1217)	21	21	21	363	363	365
NEIS0073	outer membrane transport protein (1403)	4	4	4	400	400	671
NEIS1837	glucose-6-phosphate isomerase 2 (1648)	20	20	20	587	587	1057
NEIS2437	putative lipoprotein (371)	X	X	X	new#1	new#1	new#2

This table is modified from the output from the genome comparator tool. Isolate **alleles** are coloured **cyan**, **purple** & **violet** to identify them as distinct for each new allele identified for a given locus. Alleles are coloured **black** to indicate an absence of locus. N**.men** refers to *Neisseria meningitidis*; **N.lac** refers to *Neisseria lactamica.* In the **locus annotation and size** column the value given in brackets refers to the size of the locus in nucleotides (**nt**). Any alleles with the prefix "**new#"** refer to novel alleles not currently manually curated and assigned an allele number by the moderators of pubMLST.

In locus NEIS0069 allele 363 that changed to 365 was found to differ in 46 polymorphic sites. In locus NEIS0073, allele 400 was found to differ from allele 671 in 44 sites. In locus NEIS1837, allele 587 was found to differ from allele 1057 in 34 sites. In locus NEIS2437 the "new#1" allele was found to differ from "new#2" in 28 sites.

Figure 6-9 Allelic change detected within volunteer 291 *N. lactamica* isolates displaying similarity with *N. meningitidis* isolate alleles

Locus	Locus annotation and size (nt)	N.men Week 0	N.men Week 2	N.men Week 26	N.lac Week 2	N.lac Week 4	N.lac Week 16
NEIS0071	putative lipoprotein (1019)	18	18	18	18	18	284
NEIS0072	hypothetical protein (220)	8	8	8	8	8	27
NEIS0899	alpha-2,3-sialyltransferase (1043)	82	82	82	82	82	127
NEIS2514	hypothetical protein (295)	28	28	28	28	28	48
NEIS1834	enoyl reductase (797)	84	84	84	305	305	84
NEIS1835	N-succinyltransferase (825)	26	26	26	347	347	26

This table is modified from the output from the genome comparator tool. Isolate **alleles** are coloured **cyan** and **purple** to identify them as distinct for each new allele identified for a given locus. N**.men** refers to *Neisseria meningitidis*; **N.lac** refers to *Neisseria lactamica.* In the **locus annotation and size** column the value given in brackets refers to the size of the locus in nucleotides (**nt**).

In locus NEIS0071 allele 18 allele that changed to 284 was found to differ in 32 positions. In locus NEIS0072, allele 8 was found to differ from allele 27 in 2 positions. In locus NEIS0899, allele 82 was found to differ from allele 127 in 29 positions. In locus NEIS2514, allele 28 was found to differ from allele 48 in 28 positions. In locus NEIS1834, allele 84 was found to differ from allele 305 in 19 positions. In locus NEIS1835, allele 26 was found to differ from allele 347 in 35 positions.

6.3.7 Recombination among the co-colonising meningococci in volunteers 36 and 291

Among meningococcal isolates in volunteer 36, five recombinant regions were detected by ClonalFrameML. Three of the regions were linked to loci with incomplete alleles and were subsequently disregarded however two corresponded with *murA*, (NEIS2149, UDP-N-acetylglucosamine 1 carboxyvinyltransferase) involved in peptidoglycan biosynthesis and a hypothetical protein (NEIS2859). The isolates were sampled at week 0, week 2 and week 16. Allelic variation was detected among these alleles in genome comparator but in the case of *murA*, the meningococcal alleles (68 and 288) did not match those found among the co-colonising *N. lactamica* isolates for this volunteer (allele 207). NEIS2859 was not detected among co-colonising *N. lactamica* isolates.

Among meningococcal isolates in volunteer 291, 4 recombinant regions were detected and matched to two complete loci. Two of these regions affected two pars of a conserved hypothetical protein with transferase activity (NEIS1402) as well as *pilE* (NEIS0210) an integral membrane component. The isolates were sampled at week 0, week 2 and week 16. Allelic variation was detected among these loci but none of the *pilE* alleles observed among the meningococcal isolates (alleles 733 and 686) were found among the co-colonising *N. lactamica* (novel allele) whereas NEIS1402 was absent from *N. lactamica* entirely.

The recombinant loci observed among the meningococcal isolates were amalgamated alongside those detected among the *N. lactamica* isolates to give a summary of all recombiant loci detected among all *Neisseria spp.* examined in this chapter (**Figure 6-10**).

Figure 6-10 Summary of genes undergoing recombination among *N. lactamica* and *N. meningitidis*

Organism	Volunteer	Recombinant NEIS locus	Gene Alias	Annotation	Direction of recombination
N. lactamica (I)	166	NEIS1708	NMB0444	hp	unknown
N. lactamica (I)	133	NEIS1794	N/A	putative mafS3 cassette	unknown
N. lactamica (I)	38	NEIS1828	NMB0342	intracellular septation protein A	unknown
N. lactamica (WT)	36	NEIS1795	NMC1795	mafl immunity gene: type o2MGI-2	unknown
N. lactamica (WT)	291	NEIS0069	NMB0085	sodium glutamate transport	unknown
N. lactamica (WT)	291	NEIS0073	NMB0088	outer membrane transport	unknown
N. lactamica (WT)	291	NEIS1837	NMB0334 (pgi2)	glycolysis pathway enzyme	unknown
N. lactamica (WT)	291	NEIS2437	NMB0336	putative lipoprotein	unknown
N. lactamica (WT)	291	NEIS0071	NMB0071	putative lipoprotein	interspecific
N. lactamica (WT)	291	NEIS0072	NMB0087	hp	interspecific
N. lactamica (WT)	291	NEIS0899	NMB0922 (lst)	LOS alpha-2,3-sialyltransferase	interspecific
N. lactamica (WT)	291	NEIS2514	hp	-	interspecific
N. lactamica (WT)	291	NEIS1834	NMB0336	enoyl-(acyl carrier protein) reductase	interspecific
N. lactamica (WT)	291	NEIS1835	NMB0335	N-succinyltransferase	interspecific
N. meningitidis (WT)	36	NEIS2149	NMB0011	peptidoglycan (cell wall) biosynthesis	unknown
N. meningitidis (WT)	36	NEIS2859	hp	-	unknown
N. meningitidis (WT)	291	NEIS0210	NMB0018 (pilE)	pilus component synthesis	unknown
N. meningitidis (WT)	291	NEIS1402	hp	-	unknown

The table lists genes identified as recombinant in the dataset following analysis by ClonalFrameML. Isolates of N. lactamica are differentiated with the labels wild type **(WT)** and artificially-inoculated **(I)**. Gene names are given as found on PubMLST Neisseria **(Recombinant NEIS locus)**, in addition to more commonly recognised **gene aliases**. The label **interspecific**, refers to loci that shared the same recombinant allele with their co-colonising species during the course of the study.

6.3.8 Repetitive sequence comparison between recombining *Neisseria spp.*

Using the same method outlined in chapter 3. Repetitive DNA sequences were scanned in *N. meningitidis* and *N. lactamica* isolates in volunteers 291 and 36. The searched sequences included DNA uptake sequence (DUS) dialects, DRS3 and four types of Correia repeat sequences and is shown in **Figure 6-11**.

In comparing the repeat sequences between *N. lactamica* Y92-1009 (from chapter 3) an organism not detected to undergo recombination in co-carriage with the volunteer 36 and 291 sourced meningococcal isolates which did engage in recombination the following trends were noted. The meningococcal isolates both demonstrated lower levels of AT- DUS, higher levels of AG-mucDUS and higher levels of Correia repeat type 4. However, the greatest difference was in the values for DRS3 and Correia repeat type 1 which were ~4 fold greater among the volunteer 36 and 291 meningococci compared to the innoculated *N. lactamica*.

The co-colonising wild type *Neisseria lactamica* possessed approximately the same amount of AT-DUS as *N. lactamica* Y92-1009 (~300 higher than the meningococci) and slightly lower amounts of Correia repeat type 4. However, the levels of AGmucDUS and correia repeat 1 were found to be higher than that detected in *N. lactamica* Y92-1009 and the number of DRS3 repeats were ~10% higher (WT *N. lactamica* DRS3 n=492 and 512 VS *N. lactamica* Y92-1009 DRS3 n=454) correlating the 4-fold increase observed in their meningococcal co-colonisers.

Figure 6-11 DNA repeat sequence patterns among *Neisseria spp.* undergoing recombination

Repeat	Sequence	N. lac Y92-1009	(36) WT N.lac	(36) WT N. men	(291) WT N.lac	(291) WT N. men
AT-DUS	ATGCCGTCTGAA	1718	1708	1472	1720	1472
TG-wadDUS	TGCCTGTCTGAA	0	0	0	1	0
AG-DUS	AGGCCGTCTGAA	262	259	356	281	184
AG-mucDUS	AGGTCGTCTGAA	45	63	186	55	81
AG-simDUS	AGGCTGCCTGAA	22	21	42	24	13
AG-kingDUS	AGGCAGCCTGAA	29	26	32	25	11
AG-king3DUS	AAGCAGCCTGCA	14	14	34	15	16
AG-eikDUS	AGGCTACCTGAA	0	3	6	1	4
DRS3	ATTCCCNNNNNNNNGGGAAT	454	492	1158	512	1110
Correia rpt 1	ATAG[CT]GGATTAACAAAAATCAGGAC	50	53	193	58	173
Correia rpt 2	TATAG[CT]GGATTAAATTTAAACCGGTAC	1	0	1	0	3
Correia rpt 3	TATAG[CT]GGATTAACAAAAACCGGTAC	17	17	8	17	10
Correia rpt 4	TATAG[CT]GGATTAAATTTAAATCAGGAC	17	15	30	13	20

This table lists the number of times repeat motifs were found in the genome assemblies of the above groups. The value for *N. lactamica* Y92-1009 were imported from chapter 3. The column header brackets **(XXX)** indicate the volunteer a given bacterium was recovered from. The following abbreviations were used; **WT** is wild type, ***N. lactamica*** *is Neisseria lactamica* and **N.men** is *Neisseria meningitidis.*

6.4 Discussion

This chapter of work has shown that homologous recombination enabling allelic exchange has been detected interspecifically among wild type *N. lactamica* co-colonised with meningococci. This finding was not observed in volunteers co-colonised with meningococci and *N. lactamica* Y92-1009 as no homologous recombination was detected in this strain despite greater numbers of isolates sequenced. The likeliest reason this interspecific interaction was observed among the wild type bacteria instead of the recent *N. lactamica* Y92-1009 inoculants was the duration of carriage among wild type co-colonised volunteers. Studies of recombination among *Streptococcus pneumoniae* isolates have detected a correlation between carriage length and effect of recombination (r/m) (Chaguza *et al.*, 2016).

Despite greater numbers of samples and genomes tested there was no evidence to suggest that locus-wide homologous recombination with meningococci occurred over 26 weeks of carriage in the *N. lactamica* Y92-1009 subset. The recombination metrics show that five out of seven of longitudinally carried *N. lactamica* do not show elevated r/m value compared to the average in the presence of meningococcal co-carriage. And for the two volunteers that exceed this average, r/m is maintained towards the lower thresholds (r/m= 6-14) for that previously reported in *N. lactamica* species overall, (Didelot & Maiden., 2009 Jolley *et al*., 2005). All of the importations detected by CFML were small and located within incomplete loci which would inflate r/m. This finding complements data in the chapter 5 which suggested *N. lactamica* Y92-1009 possesses a stable genome undergoing stochastic mutations over this time period with little evidence of positive selection.

For broader context, a meta-study amalgamated recombination rates sourced from a variety of bacterial phyla using MLST loci (Table 1, (Vos and Didelot, 2008)). Among other nasopharyngeal colonisers they re-reported r/m rates of 23.1 (*S. pneumoniae*), 10.1 (*Moraxella catarrhalis*), 7.1 (*N. meningitidis*), 3.7 (*H.influenzae*).and 0.1(*Staphylococcus aureus*). These estimates from MLST housekeeping loci have been improved upon by the newer technique of whole-genome recombination analysis. Using this method *S. pneumoniae* demonstrated a r/m rate of ~7, lower than that observed when just testing MLST loci (r/m: ~22-65, (Croucher *et al.*, 2011)) and *S. aureus* demonstrated a higher r/m: ~0.7, (Everitt *et al.*, 2014)). Therefore, the findings of r/m rates among commensal *Neisseria spp.* in this chapter are currently the most accurate available due to the inclusion of many more loci.

Co-carriage of *Neisseria spp.* is a rare event (Deasy *et al.*, 2015). In volunteer 36, the number and spacing of mutations affecting the new#3 allele of the *mafl* type O2MGI-2 immunity gene with no wobble base pattern (every third base) evident suggests a locus importation from an external and non-meningococcal source. Conversely it could be argued that the new #3 allele may be closer to that observed in this species naturally and that the new#2 allele observed in *N. lactamica* week 0 and week 2 isolates was imported from meningococci, positively selected and mutated slightly (2bp) to best suit the requirements of its new host. The section describing locus variation in volunteer 291 shows putative evidence of a dynamic inter/intraspecific, recombination interplay between wild type *N. lactamica* ST 11156 and the co-colonising meningococci. Except for locus NEIS0072 (alleles 8 and 27 of which demonstrated a 2-base pair/SNP difference when compared to one another), all other alleles were shown to be divergent (>19 SNPS) in terms of SNP difference. This reduces the likelihood that these mutations occurred by point mutations as opposed to a whole gene being imported from another source.

Because *Neisseria spp.* with similar types (dialects) of DNA uptake sequence (DUS) are more likely to interact and interspecifically recombine (Frye *et al.*, 2013). DUS dialects, DRS3 repeats and Correia repeats were reported among the recombining, wild type *N. lactamica* and *N. meningitidis* co-colonisers and contrasted against the recently inoculated *N. lactamica* Y92-1009 which did not recombine with its meningococcal co-colonisers. Wild type *N. lactamica* retained some species-specific traits like an elevated AT-DUS value (Pandey *et al.*, 2017) but also demonstrated elevated values of AGmucDUS and Correia repeat 1 which followed the trends observed in comparing their meningococcal co-colonisers to *N. lactamica* Y92-1009. However, the largest repeat type disparity between the inoculated and wild type *N. lactamica* was found in DRS3 repeats which were ~10% higher, mirroring the 4-fold increase observed in their meningococcal co-colonisers. This is of interest as a link between DRS3 repeat sequence possession and enhanced likelihood of exogenous DNA incorporation via homologous recombination has been suggested in *N. meningitidis* (Van Der Ende, Hopman and Dankert, 1999; Bentley *et al.*, 2007) and Neisseria gonnorhoeae (Vélez Acevedo *et al.*, 2014). However, the possession of larger amounts of DRS3 sequences among *N. lactamica* species shown to undergo recombination is currently unreported in the literature.

The experimental human challenge study described in the previous chapter (Deasy *et al.*, 2015) used single colony sampling per time point, and while this has allowed us to test a large dataset both with regards to sampling time and volunteer number, due to the fact that the volunteers were screened as having no *N. lactamica* carriage before nasal inoculation of *N.lactamica* Y92-1009 we can be reasonably sure that any recombination observed occurred during the course of this study. There was more uncertainty surrounding the provenance and evolutionary path of the

wild type *N. lactamica* that recombination was detected in. This leaves us unable to prove beyond a doubt that the recombination which was captured during the course of the study actually occurred during the course of the study. This is because minor variants of the alleles measured later in the study could have been present in the population to begin or new colonisers could have arisen having undergone recombination elsewhere. We observed the allelic profile of WT *N. lactamica* to alter alleles identified in co-colonising meningococci at earlier time-points. However, the reverse (lactamica allele transfer to *N. meningitidis)* was not seen. While it is impossible to determine whether the ancestral source of the recombinant alleles was *N. lactamica* or *N. meningitidis*. An example of interspecific, allelic transfer from *N. lactamica* to *N. meningitidis* isolates was demonstrated recently (Mulhall *et al.*, 2016) indicating its possibility.

The source of sequence data for this chapter was pubMLST Neisseria; a repository containing over 2500 Neisseria loci but biased more towards those of a more gonococcal and meningococcal origin. As a result, the locus cohort available to test (~1600) will be smaller than that naturally found in *N. lactamica* (*N. lactamica* typically contain ~2000 putative coding sequences) but is also likely to contain all potential sites of locus allele exchange via recombination between the meningococci and *Neisseria lactamica*. Using alignments of these loci as the basis of this recombination workflow has necessitated a paralogue removal step. This was caused by the incorrect assignment of lower confidence sequence information by the automatic velvet assembler. The advantages of this workflow are locus comparison against an extensive, manually-curated meningococcal database, automatic alignment generation.

The evidence in this chapter suggests that importation events have at one point occurred in the wild type *N.lac* population but the data described was not enough to conclusively comment on the nature and direction of the detected genetic transfer. In theory, this information could be obtained by examining the GC ratios of all new alleles to determine whether they represent a significant skew from the average GC ratio of the donor genome. Most of the loci affected by recombination appear to be associated with metabolism, outer membrane proteins and modification of structures known to be involved in the host-bacteria response. However, the original direction of recombination between donor and recipient and the nature of what originally was “a meningococcal allele” vs “a lactamica allele” is difficult to determine without further information on the origin of these bacteria. Nevertheless, comparative genomics have statistically validated the highly likely event of recombination between longitudinally-isolated, co-colonising *Neisseria* species. In the past, these recombination events were theorised to occur and validated by statistics. Currently, there is a high possibility of observation of recombination within a longitudinal study. But in the near future, as sequencing and isolation methods improve and

reference databases grow larger, it may become possible to exactly detect, quantify and ancestrally source the nature of homologous recombination among genetically promiscuous organisms like the *Neisseria*.

Chapter 7: The *N. lactamica* Pan genome: understanding *N. lactamica* Y92-1009 as a strain within a species.

7.1 Introduction

The term "pan-genome" was first used in the bacterial context in over a decade ago (Medini *et al.*, 2005; Tettelin *et al.*, 2005). The term was used to define a list of genes sequenced from and shared by six *Streptococcus agalactiae* isolates. In retrospect, this was the first core genome by the modern definition, a collection of genes shared by a group of isolates. The core genome is named to discriminate from the accessory genome, a collection of genes that are not shared in all the genomes examined. Combined, these gene collections constitute a pan genome.
A pan genome is only as large as the number of genes contained in the genomes that constitute it and both studies found that a total list of genes shared by the genomes is seen to gain fewer novel genes as more genomes of the same taxonomic group are continually added to the analysis. Furthermore, core genome size decreases as greater numbers and diversity are introduced into the pan-genome as fewer numbers of genes are shared by all constituents. The construction of pan-genomes is a rapidly evolving discipline made possible with a fall in the costs of WGS. Studies have attempted to utilise pan genomics to supersede the use of reference genomes in identifying diagnostic targets among bacterial pathogens. Examples of this include work on *Escherichia coli* and *Shigella spp.* (Rouli *et al.*, 2015) *Burkholderia spp.* (Sahl *et al.*, 2016) and *Campylobacter spp.* (Méric *et al.*, 2014). A study also used the pan-genome of *Staphylococcus aureus* as a basis for transcriptomic read mapping to discriminate between *in vivo* versus *in vitro* expression profiles (Chaves-Moreno *et al.*, 2015). The advantage of this being that RNA sequencing data was utilised from samples which did not possess whole genome information.

The commensal, *Neisseria lactamica* has been studied alongside its occasionally pathogenic cousin, *Neisseria meningitidis* for a number of years now. *N. lactamica* shares a large degree of genetic similarity with the meningococcus (Tobiason and Seifert, 2010). And while it does not possess certain major meningococcal virulence factors such as *porA* and capsule loci (*cps*), *N. lactamica,* along with most *Neisseria* commensals such as *Neisseria mucosa, Neisseria cinerea* and *Neisseria polysaccherea* share a large common gene pool (Marri *et al.*, 2010) with the more pathogenically associated *Neisseria spp.* Despite its status as arguably the third most notable *Neisseria* species (after the pathogens *Neisseria gonnorhoeae* and *N. meningitidis)* due to an idiosyncratic, inverse-carriage relationship with the meningococcus (Cartwright *et al.*, 1987; Bakir

et al., 2001; Deasy *et al.*, 2015), there has been no recent attempt made to understand the genetic diversity of the species *N. lactamica*.

Neisseria spp. remain a paradigm in that they are both simultaneously similar in terms of gene content and highly variable in terms of allelic variation and protein expression. This has ruled out sequence limited approaches such as microarrays with a view to determining gene presence or absence differences and necessitated the use of annotated, WGS comparison (Maiden and Harrison, 2016) Among the genus *Neisseria*, 246 core genes were found to be shared collectively. This pan-*Neisseria* analysis included thirteen species (Bennett *et al.*, 2012). An analysis identifying shared gene content between *N. lactamica, N. meningitidis* and *N. gonorrhoeae* estimated a core genome of 1190 genes (Bennett *et al.*, 2010). This value was unusually high for a multiple species core-genome and signifies a shared ancestry between the three *Neisseria spp.* A study analysing multiple serogroups among hyperinvasive, clonal complex 11 found a core-genome of 1546 genes (Lucidarme *et al.*, 2015). Lastly, an analysis of MLEE electrophoretic type 5, ST-32, serogroup B meningococci (denoted lineage 5) found 1752 core loci out of a pan-genome total of 1940 loci using disease-causing isolates sourced globally (Harrison *et al.*, 2015). These studies demonstrate an increase in core-genome estimates as the relative genetic diversity of the pan-genome falls. Among *N. gonorrhoeae,* a recombination study found 1,189 genes were core to the species (Ezewudo *et al.*, 2015).

In the work described in this chapter I had the objective to understand the extent of genetic diversity of the species *Neisseria lactamica.* An experiment of this type has not been performed since 2010 (Bennet *et al*., 2010). Since then, more than ten times the number of *N. lactamica* whole genome sequences have been made available for analysis. A species encompassing, *N. lactamica* pan genome has never been determined up until now. In performing this analysis, I will aim to quantify both the genomic diversity that currently exists within this species and the genomic content that distinguishes *Neisseria lactamica* Y92-1009 as a strain within this species. This data is crucial for investigators using this strain in experimental human challenge studies. This is because the data would inform of any potentially “harmful” (i.e. virulence factors, antibiotic resistance, restriction modification systems) genes unique to this bacterium.

I set out to achieve these goals in four steps

1) Collect a sample size as representative at possible of the species *Neisseria lactamica*. I annotated all of these genome assemblies using a pre-defined and manually curated list of *N. lactamica* proteins to make downstream homology analysis easier.

2) Construct a pan-genome in order to determine the diversity of putatively coded proteins by every sequence type in this species.

3) Analyse the pan genome to uncover any phylogenetic trends and novel genes for strain *N.*

lactamica Y92-1009 in order to understand its place within the species.
4) Evaluate the difference between the older shotgun sequenced *N. lactamica* Y92-1009 and my PacBio RS I assembly of the same organism to determine what changes if any have occurred between these events.

I expected to find more in the way of new alleles/ novel gene variants than actual novel genes unique to *N. lactamica* Y92-1009 during the course of this analysis. This is due to the fact that a "pan-neisseria", genome including *N. lactamica*, *N. meningitidis* and *N. gonorrhoeae* found few unique *N. lactamica* genes (Bennett *et al*., 2010.) This was a genus-encompassing pan genome representing greater diversity than the species-specific pan-genome I planned to construct. Nevertheless, the experiment only utilised one available *N. lactamica* genome (strain 020-06) and a much greater diversity is publicly available for the interrogation of this species.

7.2 Methods

7.2.1 Sample Collection

An investigation of gene content and diversity is only as good as its representative sample. The genomes included in this study were either

1) Downloaded from pubMLST Neisseria (http://pubmlst.org/neisseria/). On the isolate search page sequences

a) larger than 2 mega bases according to the sequence bin

b) species typed as "*Neisseria lactamica*" were searched.

Isolates were ordered by sequence type and an assembly was selected for pan-genomic analysis from every available sequence type. Where there were multiple genomes available for a given sequence type, the "best representative assembly" was chosen with regards to the lowest overall contig number and highest N50 value.

2) Reference genome assemblies were downloaded from the ncbi:refseq database using the Bio::RetrieveAssemblies perl module [http://search.cpan.org/dist/Bio-RetrieveAssemblies/].

The three reference genomes (*N. lactamica* strains ATCC 23700, 020-06 and the public health England (PHE) Y92-1009 shotgun (Vaughan *et al.*, 2006) sequence assembly were included alongside draft assemblies' representative of every sequence type of *N. lactamica* currently known and described just above. A total of 25 assemblies were included and are tabulated in **Error! Reference source not found.**.

Roary (Page *et al.*, 2015) is a pan-genomic calculation pipeline consisting of a number of sub-processes and sub-programs. These are split into four sections outlined below; annotation, quality control, alignment and protein clustering.

7.2.2 Roary: Annotation

Prokka (Seemann, 2014)was used to identify coding sequences (CDS) in long, contiguous stretches of sequence (i.e. genome assemblies) and putatively assign genetic function. Prodigal (Hyatt *et al.*, 2010) is initially used to identify all co-ordinates of CDSs from input data (genome assembly file) but does not assign a putative gene product. Once all CDSs are detected, gene prediction is normally inferred by comparing an unknown protein to a database containing known protein sequences. To ensure maximum possible accuracy, this sequence-database homology comparison is staggered hierarchically in the following way by Prokka.

Firstly, all putative CDSs are matched with a trusted list of proteins. In this study's case, these proteins are from the only available and characterised *N. lactamica* reference genome (Bennett *et al.*, 2010) this option was selected by using the" –proteins" flag.

Following this, all unannotated proteins are then compared to the highly curated uniprot bacterial database, a *Neisseria* specific RefSeq database (enabled with the " –genus" and "–usegenus" flags) and finally the PFAM database (Finn *et al.*, 2014). A given coding sequence may be able to be successfully matched by all of these but the gene function for a neisserial protein is likely to be more accurate from a *Neisseria*-specific database than a general bacterial database and so Prokka's method ensures that this given protein is more likely to be identified by a more bespoke source. Matches between CDS and protein are made using BLAST+. After these four rounds of database searches, any remaining unannotated proteins are labelled as hypothetical in function.

Example: Without the use of the trusted protein list, the *N. lactamica fetA* gene is misannotated as "generic recombinant outer membrane protein PiiC".

7.2.3 Roary Quality Control: Kraken

Kraken is a taxonomic identifier of sequence data (Wood and Salzberg, 2014). Its resolution can reliably detect at a species level but not strain level. It does this through the use of reading and querying *k*-mers, DNA or protein sequences of length = *k*. The sequence "ACTG" is an example of a 4-mer. K-mers within an input DNA or amino acid sequence can be extracted and cross-referenced against a standard genus or customised database which is built as part of kraken's installation. This database consists of *k*-mer sequences assigned to most recent common ancestors (MRCA.) A MRCA of two similar sequences is the shared ancestral node furthest from the root. In the case of identifying *N. lactamica* sequences, amino acids sequences in gff v 3.0

format are the input. The *k*-mer value used for this analysis was 31bp, as this was found through trial and error to be a good compromise between memory usage and sufficient accuracy in discerning sequences containing microsatellites or repeats.
Kraken's results contained a list of input-files submitted to it and their matched genus and species. This step was used for QC purposes as the diversity of sequences submitted to this pan-genomic study cannot exceed species level. All assemblies used in this study were downloaded from either from the RefSEQ or pubMLST Neisseria repositories. As a result, metrics in terms of coverage, contig number and N50/L50 values were available for each one.

7.2.4 Roary Protein clustering: CD-HIT, BlastP and MCL

All proteins sequences were extracted from all Prokka-annotated files and pre-clustered using CD-HIT (Fu *et al.*, 2012). Any remaining proteins from the dataset were compared all vs all, using BLASTP (Camacho *et al.*, 2009)(95% homology) and clustered using MCL (Enright, Van Dongen and Ouzounis, 2002). The clustering results from the two programs are merged together and ordered starting with gene clusters present in all genomes of the dataset (**core genes**) and ending in gene clusters in less than all of the strains analysed (**accessory genes**). Roary subdivided these core and accessory assignations into four sub-categories in the order of most to least homology among the genomes tested. These sub-categories are called the core, soft-core, shell and cloud genes. Roary described protein clusters found in 99% of the assemblies tested as **core genes**. If a given protein was found in 95-98% of all the assemblies tested, this was known as a **soft core gene**. Proteins found in 15-94% isolates were described as **shell genes** and below this threshold (0-14%) were known as **cloud genes**. Roary was run under default settings as outlined in the manual (https://sanger-pathogens.github.io/Roary/)

7.2.5 Roary Alignment (MAFFT)

The PRANK aligner (Löytynoja, 2014) is used by default but the MAFFT aligner (Katoh and Standley, 2013) can be specified by the –mafft flag. MAFFT is less computationally demanding, faster but with the caveat of less accuracy than prank. However, this study found the alignment results from MAFFT to be identical to PRANK and take less time (data not shown). For this reason, the" –mafft" flag was used in all instances of pan-genomic work

7.2.6 Roary Data-Visualisation: Phandango and R/ggplot2

The Roary output files; gene_presence_and_absence.csv and a maximum likelihood tree (gtr model) calculated from the MAFFT core gene alignment using FastTree V2.1 recompiled with duse-double were submitted to phandango (http://jameshadfield.github.io/phandango/), an interactive tool for visualising phylogenetic data.
R version 3.2.5.was used in conjunction with the ggplot2 package (Wickham, 2009) and a script provided by the Roary authors to generate dynamic plots of gene metrics as further genomic content is added to the pan-genome. An updated list of visualisations for this software can be found in the following URL. [https://github.com/sanger-pathogens/Roary/tree/master/contrib]

7.2.7 Roary Analysis

Evidence of genes unique to any given isolate/sub-group of the dataset mentioned forthwith were discerned by using the “query_pan_genome” series of commands as outlined in the manual Comparisons can be made between two groups in the pan genome by using the “difference” command, “--input_set_one/--input_set_two” flags and specifying the gff files to be discriminated following each flag. The net result for a two-group comparison were three spreadsheets displaying protein clusters unique to either group and a shared protein list.

7.3 Results

In the following section I will first describe the dataset used to construct the *N. lactamica* pan-genome. I will then comment on the composition of the pan genome, quantify the differences between each strain with regards to genes/gene variants and finally determine whether the pan-genome can be grouped phylogenetically by the country of original sample isolation. An application of the pan-genome was already described in chapter 3 (**section 3.3.5**). In this instance, the pan genome was already used to decipher the extent of genomic differences between a previously sequenced *N. lactamica* Y92-1009 assembly (done by the PHE) and our new long read sequenced PacBio genome.

7.3.1 Pan genomic dataset description

This dataset was built to span a species wide diversity for *N.* lactamica. As such, a representative genome has been selected for every distinct sequence type currently identified. The sequence types, accession numbers and assembly metrics for every genome included (and excluded) are summarised in **Error! Reference source not found.**. A total of 25 strains were able to utilised for this analysis as the best, representative assemblies for a further five sequence types (ST-585, 586, 601, 604 and 11730) failed to meet one or both of the cut-off thresholds for assembly quality (Contigs <200, N50 <50,000bp)

7.3.2 Kraken QC results

Quality control analysis on the assemblies was undertaken using Kraken enabled with the miniKraken database. Analysis on all pan-genomic genomes revealed them to be correctly classified as *N. lactamica* (S2.1.) Six biosamples hosted on NCBI's refseq database (http://www.ncbi.nlm.nih.gov/refseq/) were found to be *N. meningitidis* isolates mislabelled as "*N. lactamica*". These were found under the accession numbers AEPI01, JUNU01, JUOC01, JVAU01 & and were excluded from further analysis.

Table 7-1 Epidemiology and assembly metrics for Pan-genome dataset

Accession No.or PubMLST ID	Strain/ST	clonal complex	contigs	N50	Average coverage	CoO	Year of Isolation
020-06	020-06/ 640	640	1	NA	Unknown	UK	2005
ACEQ02	ATCC 23970		101	55069	28.4	USA	1969
AEPI01	NS-19		132	58892	95	Unknown	2010
CACL01	Y92-1009/ ST-3493	613	44	85732	Unknown	UK	1992
8917	582		104	56615	30	UK	1997
1770	584		123	55383	30	UK	1997
1777	591		158	66792	50	UK	1998
1780	594		100	69350	30	UK	1998
8778	595	595	106	99780	30	UK	1997
8837	608		89	66557	30	UK	1997
1809	624	624	104	72081	30	UK	2000
8790	631		91	78586	30	UK	1997
1827	642	1494	106	51868	30	UK	1997
29271	1209		80	120780	25	Greece	1996
36173	1494	1494	95	77373	30	Greece	1997
27623	6206	624	100	83434	50	Unknown	Unknown
27622	10326	640	95	116447	50	Unknown	Unknown
36190	10984	640	75	127610	40	Ireland	2013
29274	11143	640	66	102277	40	Greece	2013
36174	11383	624	126	57498	30	Ireland	2013
38985	11653	613	81	58148	200	France	Unknown
36897	1495	1494	94	80026	30	UK	2012
36874	4192		84	88963	30	UK	2012
36856	11524		97	75913	30	UK	2012
36901	11526		126	65137	30	UK	2012
Unpublished/ PacBio	Y92-1009/ ST-3493	613	1	NA	~500X	UK	2015
1771	585		489	24773	30	UK	1999
1772	586		568	21036	30	UK	1998
1787	601		555	27420	40	UK	1999
1790	604		698	18443	30	UK	2000
38975	11730	640	247	18851	150	France	Unknown

The top four rows contain *N. lactamica* genomes hosted on genbank/refseq. All other assemblies were obtained from PubMLST Neisseria (and therefore only have 4-5-digit ID numbers). The assemblies highlighted in bold **(36897, 36874, 36856, 36901, Unpublished/PacBio**) were isolated and sequenced as part of the experimental human challenge study described in chapter 5. Of these five isolates, three were novel sequence types (ST-4192, 11524 & 11526) in the pubMLST database. The assemblies highlighted in **red** were excluded from further analysis due to quality control issues. Where assemblies are complete the value of "1" is given in contigs column and "NA" is given in the N50 column.

Abbreviations include **ST**: Sequence Type, **CoO:** the country of origin of the isolate

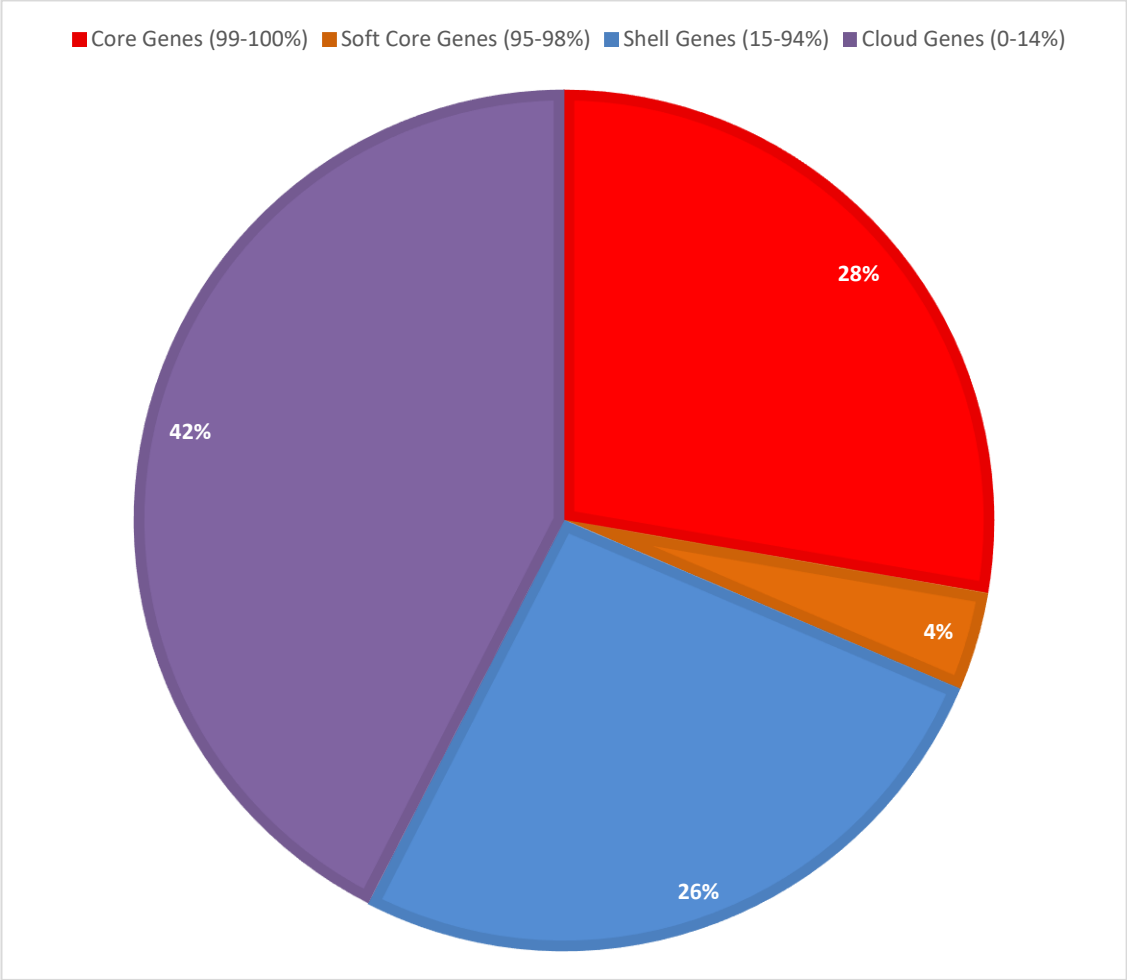
7.3.3 Pan Genome metrics reveal genetic diversity of *N. lactamica* species

The pan genome revealed a total of 4,407 genes disseminated across the representative 25 *N. lactamica* isolates. Of these, 1221 genes were present in 99-100% of all genomes tested and constituted the core genome, 163 genes were present in nearly all (95-99%) of the genomes tested (soft core.) The accessory genome was calculated by adding the numbers of shell and cloud genes together and contained 3023 genes (**Figure 7-1**).

As genomes were added to the pan-genome, plots were generated detailing the fall in core genome size correlating with the inclusion of more genomes (and therefore variation) and the subsequent steady increase of the number of unique genes **(Figure 7-2)**. The values of both novel (7-2A) and conserved genes (7-2B) discovered in conjunction with genome inclusion fell sharply at first before steadily trending downwards These observations are in contrast with the values of unique genes (7-2A) and total gene content (7-2B), both of which increased at consistent rates

Qualitative gene presence and absence analysis of the pan-genome using phandango revealed that the sequence type with the largest number of unique genes was ST-11526 followed by ST-595, ST-591, reference genome ATCC_23970 and ST-11383 (**Figure 7-3**). *N. lactamica* Y92-1009 genomes, represented in the tree as “PacBio” (our assembly) and “Y92-1009” (porton down assembly), seem to share a small block of genes unique to the strain, this can be observed above the 2nd to last plateau (n=2) although this block is smaller than those mentioned above the n=1 plateau. This qualitatively indicates that there are fewer unique genes in *N. lactamica* Y92-1009 than those of the strains. Furthermore, there was a region lacking genes in the Porton Down assembly above the soft-core genome (Strain Y92-1009). This was unique among the *N. lactamica* strains analysed. These accessory genome differences were quantified by examining the gene presence and absence .csv file outputted by Roary and filtering the results by genes/gene variants only detected once per strain (**Figure 7-4**). This graph reveals that the strain of interest *N. lactamica* Y92-1009 has the fourth smallest cloud-genome (n=24 strain-specific genes/ gene variants). The ST with the lowest number of strain-specific genes was ST-10984 (n=18) while the largest belonged to reference genome *N. lactamica* ATCC_23970 (ST-3787 in figure, n=140 strain specific genes/gene variants).

Figure 7-1 Distribution of genes constituting *N. lactamica* pan genome.



Warmer colours (red and orange) illustrate the core genome consisting of the core and the soft core, as defined by Roary. Colder colours (blue and purple) illustrate the accessory genome (shell + cloud genomes combined).

The **%** values in the **pie chart** refer to the proportion of proteins assigned to each sub-group constituting the pan-genome (n= 4,407 total proteins). The **%** values in the **key** (core, soft-core, shell and cloud) refer to the percentage of isolates required to possess a given gene to fall under that pan-genomic sub-category. The core genome size of *N. lactamica* (core + soft core, n=1221 + 163 genes) is 1384 genes while the accessory genome (Shell + Cloud, n=1151 + 1872 genes) size is 3023 genes.

.

Figure 7-2 **Shifts in pan genome metrics upon genome inclusion**

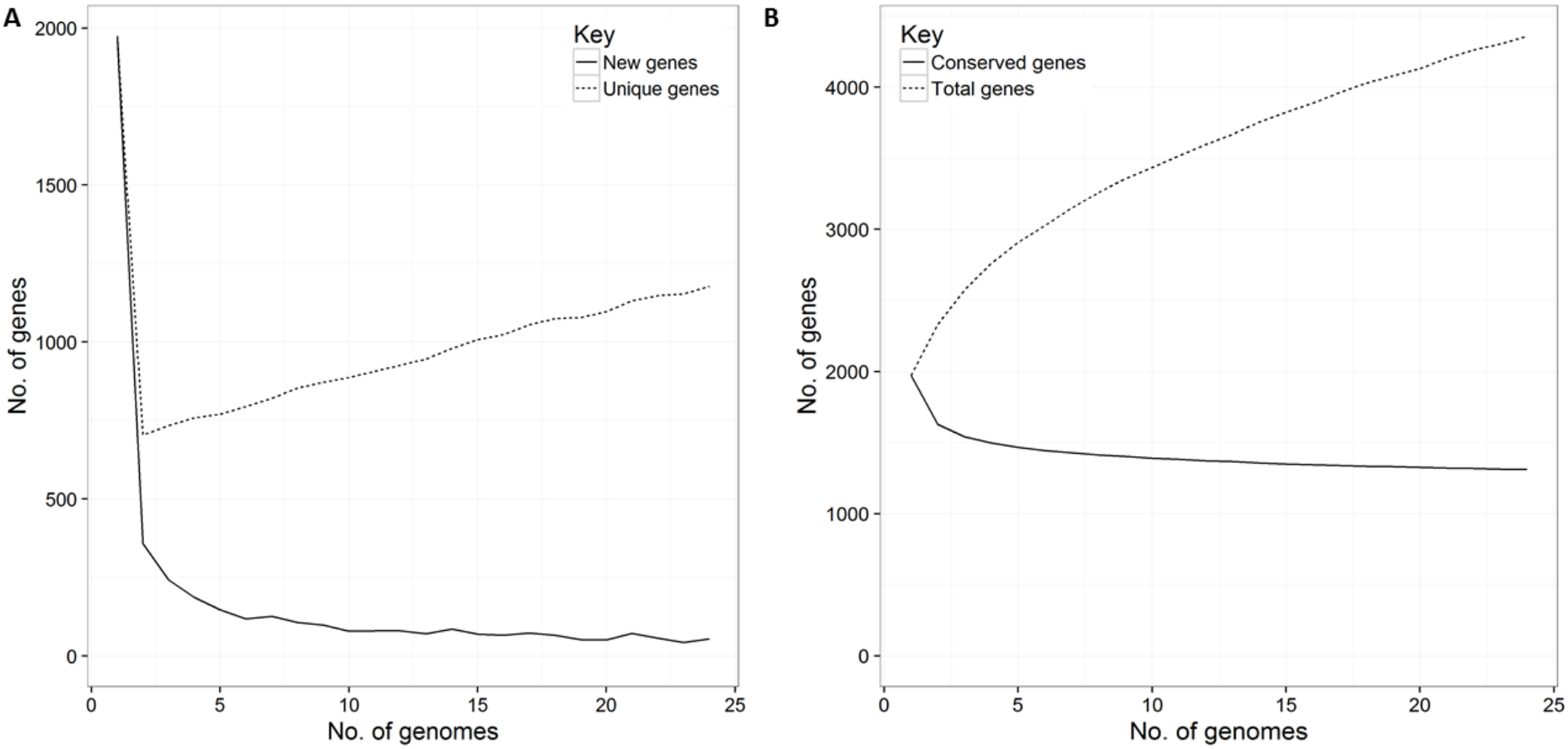


Figure 7-3 Pan Genome: Gene presence and absence of pan genome assemblies ordered against a core genome maximum likelihood phylogeny

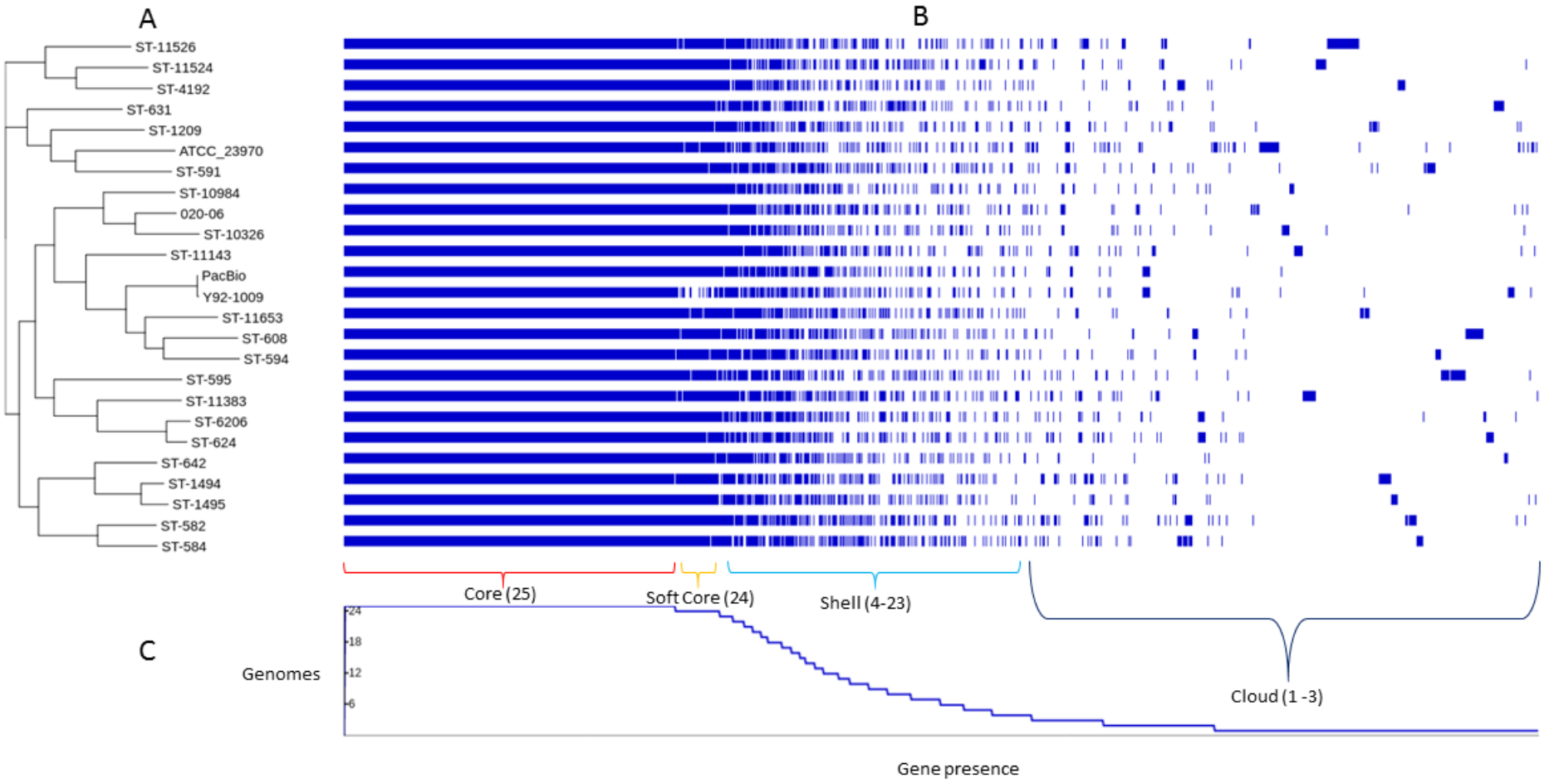
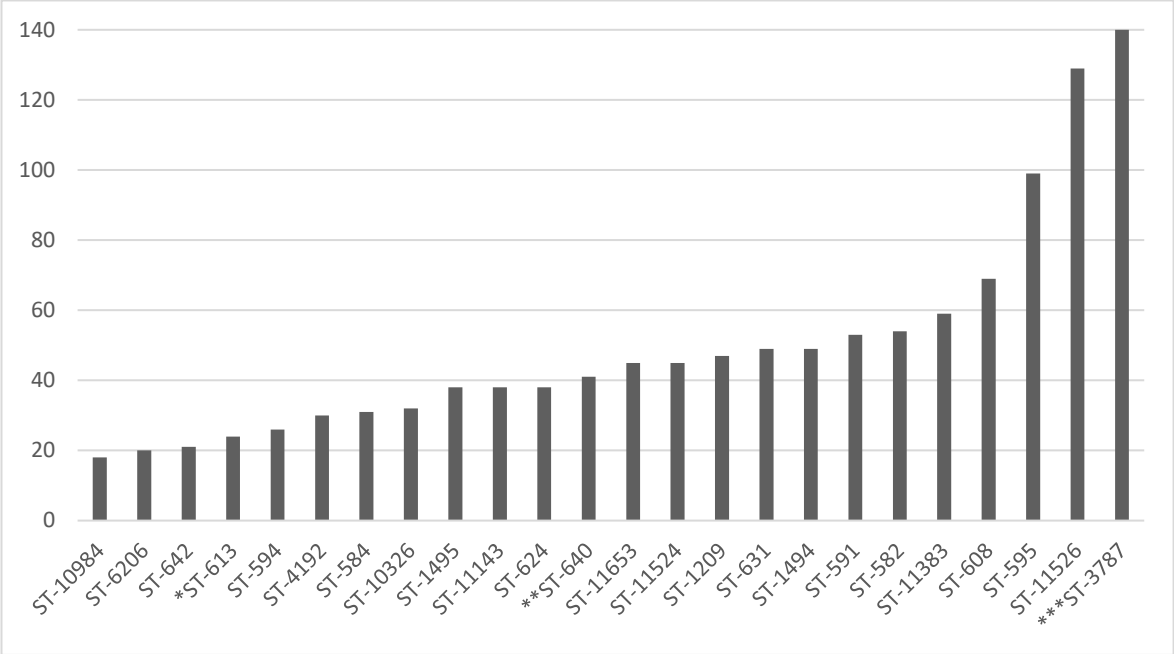


Figure Legend: Figure 7-3

As mentioned in the methods section, this diagram was created on phandango by combining the gene presence and absence csv file generated by Roary with a core maximum likelihood phylogeny. The diagram is split into 4 sections; A, B, C & D.
A: an unrooted, maximum likelihood phylogeny based on a core genome alignment. **B**: Blue segments represented gene presence; white segments represented gene absence. The pan-genome is displayed, starting from the core genome and transitioning into the accessory genome with increasing gene sequence disparity. **C**: A trace representing gene presence or absence among the genomes in the *N. lactamica* pan genome. Similar to B, the trace starts with the core genome (n=total/25 genomes) and falls steadily as it transitions into the accessory genome displaying blocks of genes shared by fewer and fewer genomes. The core transitions to the soft core, then the shell genome (n= genes present in 4-23 genomes). Here, there appear to be larger and larger blocks of blue genes towards the end of the diagram, existing in white space, above the final plateau of the trace (n=1 genome). The blocks present above the cloud genome (n=1-3 genomes) indicate genes unique to a given sequence type of *N. lactamica*.

Figure 7-4 Number of strain specific genes/gene variants detected



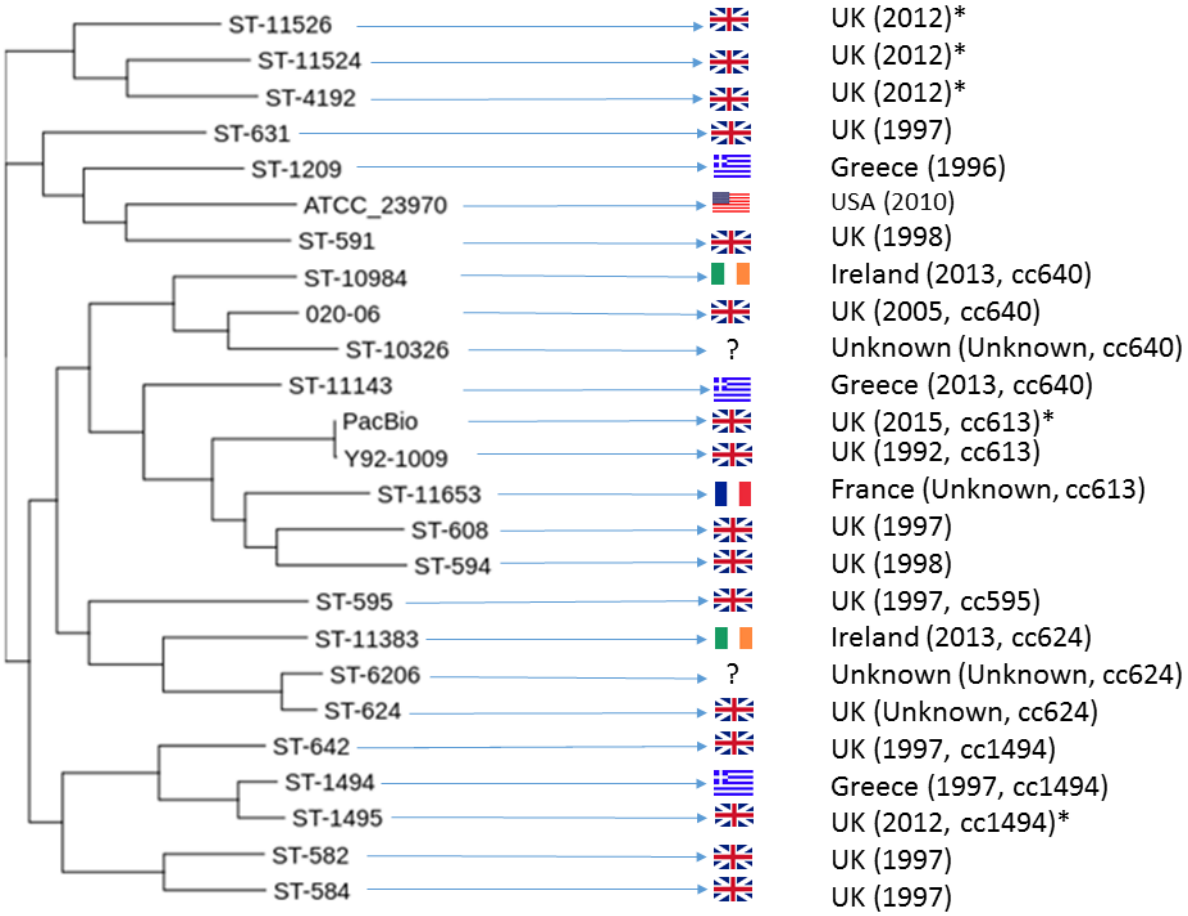
In the graph above, the strains are ordered from left to right by possession of the least to the most strain-specific genes or gene variants.

(*****) Refers to the strain of interest in this thesis (*N. lactamica* Y92-1009), the value was generated by averaging the results of the porton down assembly and our long-read sequenced assembly. (******) Refers to reference genome *N. lactamica* 020-06 and (*******) refers to reference genome *N. lactamica* ATCC_23970. The total number of strain specific genes sorted by this method was n=1195 with the average strain possessing 49.8 specific gene/gene-variants +/- 30.9 standard deviation.

7.3.4 Pan-genomic phylogeny: Country of Origin

Genomes constituting the *N. lactamica* pan genome show no phylogenetic grouping associated with country of isolate origin and year sampled (**Figure 7-5**). While three novel sequence types found in our study (Sheffield, UK) were grouped together phylogenetically (ST-11526, ST-11524 & ST-4192), the fourth, pre-existing sequence type (ST-1495) clustered further down the tree with an isolate of Greek origin.

Figure 7-5 Unrooted maximum-likelihood phylogeny based on the *N. lactamica* core genome



Unrooted core genome, maximum likelihood tree of the core alignment of the *N. lactamica* pan genome. Where possible, country flags have been used to indicate the country of original isolation of the assemblies. Next to the flags, the country is given alongside the bracketed values for original year of isolation and clonal complex. Asterixed (*****) genomes refer to isolates found during the course of the experimental human challenge study (Chapter 5) and our long read sequenced (PacBio) genome assembly. The abbreviation **cc** is short for clonal complex

7.3.5 Using the pan-genome to determine the difference between two assemblies of the same strain

The pan-genome was used to determine the difference between the same strain assembled using shotgun sequencing (Vaughan *et al.,* 2006) and PacBio RS I long-read sequencing. This was described in greater detail in chapter 3 with the raw data available in **Appendix tables 0-1, 0-2, 0-3, 0-4** &**-0-5.** The shotgun sequenced Y92-1009 genome contained little in terms of novel gene possession and variation when contrasted with the rest of the *N. lactamica.* The "query_pan_genome" command was used on the finished pan genome to compare the Y92-1009 genome (Vaughan *et al*., 2006) in terms of unique genes with that of all other *N. lactamica* ST's. The analysis detected 30 novel genes or gene variants once low-quality results were filtered (**Table 6-11**). The majority of the genes unique to the Porton Down assembly of *N. lactamica* Y92-1009 were proteins of unknown function (hypotheticals, n=19) Examples of all putatively annotated proteins (n=11) were detected in other *N. lactamica* genomes (due to the annotation tag [Neisseria lactamica]) suggesting these results were not novel genes to the *N. lactamica* gene pool but instead were novel gene variants.

Table 7-2 List of genes & gene variants unique to *N. lactamica* Y92-1009 PHE assembly

Protein cluster	Annotation	Putative Function
group_1523	hypothetical protein	
group_2039	hypothetical protein	
group_2569	hypothetical protein	
group_2575	hypothetical protein	
group_2608	hypothetical protein	
group_2610	hypothetical protein	
group_2620	hypothetical protein	
group_2622	hypothetical protein	
group_2623	hypothetical protein	
group_2652	hypothetical protein	
group_2660	hypothetical protein	
group_2584	hypothetical protein [Neisseria lactamica]	
group_2619	hypothetical protein [Neisseria lactamica]	
group_2650	hypothetical protein [Neisseria lactamica]	
group_2656	hypothetical protein [Neisseria lactamica]	
group_2657	hypothetical protein [Neisseria lactamica]	
group_2590	lactoferrin-binding protein B [Neisseria lactamica]	iron acquisition
group_2636	large multi domain protein [Neisseria lactamica]	
group_2574	membrane protein [Neisseria lactamica]	
group_2597	membrane protein [Neisseria lactamica]	
group_2635	MULTISPECIES: hypothetical protein [Neisseria]	
group_2651	MULTISPECIES: hypothetical protein [Neisseria]	
group_2655	MULTISPECIES: hypothetical protein [Neisseria]	
group_2599	Opa protein [Neisseria lactamica]	host pathogen interaction
group_2585	para-aminobenzoate synthase [Neisseria lactamica]	folate metabolism intermediary
group_2647	protein Ex13L [Neisseria lactamica]	membrane protein component
group_2662	putative transposase [Neisseria lactamica 020-06]	
group_2659	putative TspB protein [Neisseria lactamica 020-06]	T cell stimulating protein
group_2604	RecT family protein	Recombination
group_2654	transferrin-binding protein B [Neisseria lactamica 020-06]	iron acquisition

The above table was generated using the "query_pan_genome" command comparing the Y92-1009 genome (Vaughan *et al.*, 2006) with that of all other *N. lactamica* ST's. The **protein cluster** column displays the gene groups as ordered by MCL and categorised by Roary. Information on gene identity is found in the annotation column. Additional information on these unique genes/genes variants is available in the **putative function** column for those genes whose function has been studied more extensively than others.

7.4 Discussion

One study (van Tonder *et al.*, 2014) estimated the species-encompassing-core genome size for five commonly studied bacterial pathogens; *Streptococcus pneumoniae* (n=336, STs=163), *Campylobacter jejuni* (n=601, STs= 134), *N.meningitidis* (n=518, STs=198), *Staphylococcus aureus* (n=534, STs=25) and *Helicobacter pylori* (n=107, undeterminable STs). When calculated, the core genome of these bacterial species ranged from 244-851 genes, a smaller value compared to the *N. lactamica* core (core +soft core according to Roary nomenclature) genome size of 1384 genes. The reasons for this could be due to the annotation method, definition of what constitutes a core genome (95% representation among strains tested in our analysis) or the sample size (n=25) used to calculate our pan genome. Core genome size has been shown to decrease as additional samples are added and diversity is introduced (Sahl *et al.*, 2016). This means as more *N. lactamica* genomes are sequenced and included in the pan genome, core genome size could fall to mirror those observed in the van Tonder study. Unfortunately, there remains a paucity of publicly available whole *N. lactamica* genomes when compared to the pathogens above. Despite this, the diversity of the *N. lactamica* pan-genome almost matched that of van Tonder's methicillin-resistant S.*aureus* pan genome. Both analyses used a total of 24 sequence types to represent species wide diversity. The study that calculated this data defined core genes as 99% while Roary defines the core as 95%. This difference in estimation may have led to a more conservative estimate of core genome size in the van Tonder study.

Among *Neisseria meningitidis,* the core-genome was defined as 1605 genes present among 95% of 108 well-characterised, meningococcal isolates analysed (Bratcher *et al.*, 2014). More recently, the *N. gonnorhoeae* core-genome MLST V.1.0 (currently in development at the time of writing) defines 1668 loci as being core to the species (Harrison *et al.*, 2017). Both estimates for meningococcal and gonococcal core genome size exceed the value reported for *N. lactamica* (n=1384) in this chapter. This could be because all these analyses use data hosted on PubMLST Neisseria as a basis and the status of meningococci and gonococci as global pathogens leads to a greater number and diversity of isolates sequenced and therefore novel loci discovered. At the time of writing there are currently >11,500 and >4500 profiles defined by the *N. meningitidis* and *N. gonnorhoeae* cgMLST v1.0 schemes respectively. Meanwhile searching for *N. lactamica* isolates with a sequence bin size of >2mb (i.e. searching for whole genomes not just MLST allele data) returns just 141 public isolates, the majority of which are *N. lactamica* Y92-1009 submitted by our group.

N. lactamica Y92-1009 was seen to possess the fourth smallest accessory genome (with regards to strain-specific unique genes or gene variants) compared to other *N. lactamica* strains. This finding is important in the context of using *N. lactamica* Y92-1009 as a tool in future experimental human challenge studies. Accessory genome size comments on the potential genomic flexibility of a given organism (Mathee *et al.*, 2008). A larger accessory genome constitutes a greater, potential bank of genes that are less conserved by purifying selection (Bohlin *et al.*, 2017) and may mutate unpredictably due to horizontal gene transfer (Segerman, 2012),with unknown repercussions on participant safety.

There was no phylogenetic correlation between the current members of the *N. lactamica* pan genome and their country of origin or year sampled. Sequence types belonging to the same clonal complex did cluster together but all this is indicative of is the success of MLST in classifying relationships between these dynamically evolving organisms. Due to a paucity of global strains, our pan genome is Western Europe centric with most isolates sourced from the UK & Ireland. While the sole strain sourced from outside Europe in this study (*N. lactamica* ATCC_23970, USA) was not an outlier in phylogenetic inference, it was found to contain the greatest diversity in the form of highest number of strain-specific genes/gene-variants. As more genomes of this species are sequenced, country specific phylogenetic clustering such as that detected in African meningococci ((sequence types 7 & 2859, (Lamelas *et al*., 2010) may be possible to be identified. This is because *Neisseria spp.* are more likely to clonally diversify via horizontal gene transfer within clades rather than between them (Budroni *et al.*, 2011). The Lamelas *et* al study also attempted to link phylogeny of the African meningococcal pan genome to year of isolation. While this was unsuccessful at sequence type level, there was evidence suggesting the same clonal complexes (groups of related sequence types) had persisted for a number of years. This in lieu with findings of the clonal complex flux witnessed in the gradual expansion of CC11 serogroup W meningococci over three decades (Mustapha, Marsh and Harrison, 2016). This was a clonal complex that had originated in the Arabian Peninsula and migrated to the African belt (Read, 2014.)

The Roary pipeline contains a quality assurance program called Kraken which is used to species type input genome assemblies. This tool was successfully used to detect sample contamination issues from genomes downloaded via the REFSEQ repositories. These genomes had been falsely classified as *N. meningitidis* not *N. lactamica* and the subsequent pan-genomic analysis with the meningococcal genomes presented an obvious erroneous finding in its core genome size of just 200 loci out of 2500 (data not shown). This finding represents the major disadvantage of using automated, open-source databases and greater efforts must be made to ensure the quality

control of publicly accessible data. A solution to this issue would be to take the genus and species reporting out of the submitter's hands by standardising and automatically screening the submission of bacterial genomes through a metagenomic, sequence-read typer such as Kraken (with the miniKraken database) or a genus specific MLST scheme such as those described on PubMLST.org

One drawback of the Roary pipeline is the loss of annotation data (gff input) after the protein clustering step untaken with CD-HIT. Even though the *N. lactamica* genomes in this analysis were annotated with the same trusted proteins list (derived from the *N. lactamica* 020-06 genome) the vast majority of clustered results are given as numerical protein groups and not as specific genes. Parsing the pan_genome_reference.fasta file given by Roary to output core and accessory sequence data would be the first step in determining the functional groupings of the core and accessory genome of *N. lactamica*. This could be done by matching the loci in the core/accessory genomes against the COG database (Tatusov, 2001), the BLAST2GO typing scheme (Conesa *et al.*, 2005) or the KEGG database (Ogata *et al.*, 1999) as was undertaken for the meningococcal core genome (Bratcher *et al.*, 2014).In addition the core alignment outputted by Roary and used to calculate the maximum-likelihood phylogeny could be used as input into a recombination detection program such as Gubbins (Croucher *et al.*, 2014) or the pipeline described in Chapter 6 using ClonalFrameML. This would determine which loci were subject to the greatest recombination pressure for the *N. lactamica* species as whole. A result which has yet to be determined for this *Neisseria* species.

This chapter has shown that pan genomic analyses are well suited for visualising and grasping the genetic diversity present within a species. This highly dynamic protein network is easily upgradeable and scalable to much larger datasets than the one presented in this chapter. This will become important as more whole genome strains of *N. lactamica* are sequenced in the future. This chapter was the first to use pan-genomic analysis to discriminate *N. lactamica* into a phylogenetic network and quantify the unique genes/gene variants of said network. As a result, there is a significant amount of gene presence and absence data available from every representative sequence type of the species; *Neisseria lactamica.* This could be used next to compile the transcriptome of novel strains of this organism (as demonstrated in work using *S. aureus*, (Chaves-Moreno *et al.*, 2015)) without the pre-requisite of complementary, strain-specific WGS. In addition, the genes constituting the core-genome could be submitted as a typing scheme (cgMLST) on https://pubmlst.org/neisseria/ to supplement pre-existing schemes available for *N. meningitidis* and *N. gonorrhoeae* and aid future research on this organism.

Chapter 8: Final Discussion

This body of work was done to determine the mutational stability of *N. lactamica* Y92-1009, a potential bacterial medicine in combatting serogroup B meningococcal carriage, a pre-requisite of serogroup B invasive meningococcal disease. Understanding the microevolution of the commensal *Neisseria* has become increasingly important because; (a) the meningococcus is typically commensal and hypervirulent lineages represent a small proportion of the overall diversity of this species (Caugant and Maiden, 2009) and (b) the lack of differentiation between *Neisseria* commensals and pathogens at the genomic level (Bennett *et al.*, 2010; Marri *et al.*, 2010).

Despite the findings of meningococcal carriage clearance and prevention of reacquisition following *N. lactamica* inoculation (Deasy *et al.*, 2015) , there was no genomic-wide analysis used during this study to support the reported phenotype. Specifically, there was no difference in mutation number or type between *N. lactamica* Y92-1009 co-colonised with meningococci or in the majority of volunteers where it was the sole *Neisseria* coloniser. The exact mechanism for this carriage suppression remains unknown but could be that *N. lactamica* inoculation leads to the bacteria outcompeting *N. meningitidis* for resources in the nasopharyngeal niche. The possession of genes doesn't correlate with their level of expression, so a follow-up, transcriptomic interrogation between the solo-colonising and co-colonising *N. lactamica* isolates may determine whether there is a difference between which genes were up and down regulated in response to meningococcal co-carriage. Discrepancies in transcriptomic profiles have been observed in other *in vivo* sourced bacterial strains, despite sharing almost identical genetic sequence (Klockgether *et al.*, 2013).

The genome of *N. lactamica* Y92-1009 possessed a mutation rate estimate similar to that observed in other *Neisseria spp* (Grad *et al.*, 2014; Lamelas *et al.*, 2014; Ezewudo *et al.*, 2015; Didelot, Dordel, *et al.*, 2016). In, addition, no interspecific, gene transfer or elevated effect of recombination was detected in eight volunteers co-colonised the with the inoculated *N. lactamica* Y92-1009 and meningococci. This contrasted with the interspecific recombination dynamic observed among two volunteers presenting with wild type co-colonising *N. lactamica* and meningococci. For *N. lactamica* to be viewed as a potential bacterial medicine for the suppression of meningococcal carriage it needs to; a) be confirmed to be harmless residing in the airway, b) remain genetically stable while residing in the human airway and c) establish carriage among a greater proportion of inoculated individuals. The first and second points have been directly addressed with the recombination and mutation analyses in this study. Also, the first example of a pan genomic analysis of the species *N. lactamica* (chapter 7) revealed that *N. lactamica* Y92-1009

possessed one of the lowest numbers of unique genes/gene variants and appeared to possess one of the smallest accessory genomes among the species *N. lactamica*. None of the genes/gene variants unique to this strain were virulence factors. These findings have been submitted as part of a grant for a further carriage study where *N. lactamica* Y92-1009, genetically modified with *nadA* (an adhesin), will be used to improve carriage uptake while still conferring the same carriage-clearance benefit (DEFRA, 2017).

Mutational analysis contrasting the *in vivo* vs *in vitro* cohorts of *N. lactamica* in a short term, multiple colony sampling study (chapter 4) revealed a differing pattern in the types of mutations observed. With larger numbers of SNPs, nonsense and recurring mutations observed in the *in vitro* cohort and more phase variants detected in the *in vivo* cohort. This latter finding was confirmed in the mutational analysis of the long term (6 month), single colony sampled dataset (chapter 5) where modifications to phase variable tracts dominated all other types (SNPs, indels, substitutions) types of mutation. With the nasopharyngeal environment changing significantly due to factors such as temperature, air exchange and nutrient availability, phase variation is a mechanism strongly suggested to afford advantages to an organism colonising this niche (Jerome *et al.*, 2011). The rapid ON-OFF phenotypic switching of expression may allow a given bacterial isolate the best advantage of surviving within a host at a given point in time. Furthermore, the lack of a consistent pattern of mutation observed across multiple volunteers suggest that mutations were observed to more stochastic than adaptive. This suggests that the genes in the *N.lac* Y92-1009 inoculum did not need to be broadly modified when transitioning from liquid culture to *in vivo* carriage. Limitations in the long-term *in vivo* study include the relatively few isolates recovered at early time-points, and the lack of multiple-colony sampling from individual samples. Isolates are difficult to recover soon after colonisation as shown by the short-term (*in vitro* VS *in vivo*) study in which more isolates were recovered 2-4 weeks post colonisation than during days 1, 2 and 3. This early window may be the point at which the majority of adaptation occurs and indicates that a method of directly sampling and whole-genome sequencing from clinical samples must be introduced (Hasman *et al.*, 2014) . The *in vivo* cohort of the short-term, multiple-colony sampled study also showed that the vast majority of SNPs occurred in only one isolate sampled per time point, so whilst multiple colony sampling would undoubtedly result in a greater number of mutations observed in the long term *in vivo* study, this would not likely alter mutation rate estimation. However, the over reliance on one isolate as a proxy representation of a population of organisms, which are known to disperse via micro-colonies to colonise new sites in their habitat (Sigurlásdóttir *et al.*, 2017), means this data represents a snapshot of microevolution rather than a complete picture. In addition, since mutation rate estimates were obtained during this study, follow up work could test for neutral evolution via the statistic of

Tajima's D and calculate the effective population size (as showcased in (Levade *et al.*, 2017; Ghalayini *et al.*, 2018)) by determining Tajima's or Watterson's theta. This could be done by using the pop-genome R package (Pfeifer *et al.*, 2014) and pre-existing alignments generated for recombination analysis and would confirm whether the bacterial populations examined in this study rejected the neutral evolution model and were subject to selective pressures.

A genome-wide association study (GWAS) is defined as the observational study of a genome encompassing set of genetic variants. Since multiple chapters of this thesis deal with microevolution that occurred during a longitudinal study, it would be pertinent to class those chapters as examples of GWAS. Important limitations to consider in the experimental design of such a study include a requirement for stringent and fine-tuned statistics in calling variants (Ribeiro *et al.*, 2015). This is to strike a balance between a) detecting false positives or b) failing to identify rare variants and necessitated the use of both bespoke software designed to statistically test and report direct evidence of mutation in bacterial GWAS (Barrick *et al.*, 2014) and high coverage/per-base accuracy sequencing to increase the number of high quality reads spanning a variant. Another GWAS limitation to consider is genotyping error which occurs when the observed genotype does not correspond to the true genotype (Pompanon *et al.*, 2005). In isolating our organism of interest (*N. lactamica* Y92-1009), the bacteria were screened for beta-galactosidase activity which was previously thought to be unique to this species but has since been identified in *Neisseria oralis* (Bennett, Jolley and Maiden, 2013). To address the potential genotyping error of using data from a non-experimental *Neisseria spp.*, the sequence data was submitted to PubMLST Neisseria which enabled us to species-type samples via MLST and confirm the organism of interest before utilising sample reads in the variant calling step. Another potential genotyping error arises when determining whether observed variation occurred *in vivo* among our human participants or whether microevolution occurred during the minimal passage steps preceding DNA isolation. Chapter 4 of this thesis addressed this potential genotyping error in detail by using a multiple-colony, sampling approach and contrasting mutations observed between *in vivo* vs *in vitro* cohorts. The lack of functional information surrounding genes in which mutations are identified is another limitation of GWAS (Pearson and Manolio, 2008). And mutations occurring in intergenic regions or coded proteins of unknown function (hypothetical proteins) are particularly prone to this limitation. This study addressed this limitation by: predominantly focusing on mutations that occur within the coding sequence, using databased resources to ascribe function to hypothetical proteins of interest and using a trusted proteins list derived from a related reference genome (*N. lactamica* 020-06) in any annotation steps.

This study used a mix of both read mapping and *de novo* assembly-based methods of analysis. *De novo* assembly-based approaches represent a trade-off between accuracy when compared with

reference-based approaches but despite this, they exist as their own entity and can be compared or aligned with a vast and growing library of whole and draft assemblies. This has the advantage that any genome assembly can theoretically be compared to any another genome assembly irrespective of genus or species similarity (Henson, Tischler and Ning, 2012). It would not have been possible to calculate a pan-genome (chapter 7) or detect recombination (chapter 6) without using this approach. Reference based approaches represent a trade-off between *de novo* assemblies' versatility with accuracy and statistical confidence. The strength of this relies on the reference sequence, a genome assembly that allows the mapping of NGS reads and whose accuracy and confidence correlates directly with that of the data generated. Reference based mutational analysis was successfully used in chapters 4 and 5 using a closed genome sequenced and assembled in chapter 3 to detect modifications to all regions of the genome. However, the use of short read sequence mapping in mutation detection did not allow us to determine phase variation in multimeric sequences. One of the genes which was excluded due to our criterion that reads must span the entire length of the varying sequence was *modA*. This is a global regulator of transcription (Tan *et al.*, 2016) in other *Neisseria spp.* but due to its ability to affect methylation, may be studied alongside other regulatory loci by examining the methylome profile generated by our long read sequencing technique. *Neisseria spp.* seem to be able to possess an enhanced ability to alter protein expression without altering the base DNA. Mechanisms such as phase variation (Bidmos & Bayliss, 2014) and global regulation via methylation (Tan *et al*., 2016) by restriction-modification systems have been shown to facilitate this. A keener understanding of the transcriptomic effects of contingency and *mod* loci on the *Neisseria* genome could be coupled with the ability of NGS to monitor changes to the loci with statistical confidence. In the author's opinion, this represents the next key area of required understanding for researchers studying the evolution of this genus.

The study concludes that the *N. lactamica* Y92-1009 genome is a self-curated system with plastic elements (like other *Neisseria spp.*) that could facilitate rapid changes in expression via its phase variable elements. However, it appears to have remained genetically stable during the 6-month course of carriage in human volunteers. Demonstrating little recombination, no interspecific gene transfer with co-colonising meningococci and an average mutation rate for a *Neisseria* species. While efforts need to be made to improve the acquisition and retention of carriage, *N. lactamica* appears to be a safe, naturally competent, potential bacterial therapeutic, capable of a broad-spectrum reduction of meningococcal carriage. This may be a crucial trait in the continual absence of a herd-immunity-inducing, polysaccharide conjugate vaccine for serogroup B meningococci.

List of References

Achtman, M. and Wagner, M. (2008) 'Microbial diversity and the genetic nature of microbial species', *Nature Reviews Microbiology*, pp. 431–440. doi: 10.1038/nrmicro1872.

Aguilera, A. and García-Muse, T. (2013) 'Causes of Genome Instability', *Annual Review of Genetics*, 47(1), pp. 1–32. doi: 10.1146/annurev-genet-111212-133232.

Alamro, M. *et al.* (2014) 'Phase variation mediates reductions in expression of surface proteins during persistent meningococcal carriage', *Infection and Immunity*, 82(6), pp. 2472–2484. doi: 10.1128/IAI.01521-14.

Andrews, S. (2010) *FastQC: A quality control tool for high throughput sequence data*, *Bioinformatics*. doi: citeulike-article-id:11583827.

Araya, P. *et al.* (2015) 'Neisseria meningitidis ST-11 clonal complex, Chile, 2012', *Emerging Infectious Diseases*, 21(2), pp. 339–341. doi: 10.3201/eid2102.140746.

Artenstein, M. S. *et al.* (1970) 'Prevention of Meningococcal Disease by Group C Polysaccharide Vaccine', *New England Journal of Medicine*, 282(8), pp. 417–420. doi: 10.1056/NEJM197002192820803.

Au, K. F. *et al.* (2012) 'Improving PacBio Long Read Accuracy by Short Read Alignment', *PLoS ONE*, 7(10). doi: 10.1371/journal.pone.0046679.

Aziz, R. K. *et al.* (2008) 'The RAST Server: Rapid annotations using subsystems technology', *BMC Genomics*, 9. doi: 10.1186/1471-2164-9-75.

Bakir, M. *et al.* (2001) 'Asymtomatic carriage of Neisseria meningitidis and Neisseria lactamica in relation to Streptococcus pneumoniae and Haemophilus influenzae colonization in healthy children: Apropos of 1400 children sampled', *European Journal of Epidemiology*, 17(11), pp. 1015–1018. doi: 10.1023/A:1020021109462.

Barnett, D. W. *et al.* (2011) 'Bamtools: A C++ API and toolkit for analyzing and managing BAM files', *Bioinformatics*, 27(12), pp. 1691–1692. doi: 10.1093/bioinformatics/btr174.

Barrick, J. E. *et al.* (2009) 'Genome evolution and adaptation in a long-term experiment with Escherichia coli.', *Nature*, 461(7268), pp. 1243–1247. doi: 10.1038/nature08480.

Barrick, J. E. *et al.* (2014) 'Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq.', *BMC genomics*, 15, p. 1039. doi: 10.1186/1471-2164-15-

1039.

Bayliss, C. D. *et al.* (2008) 'Neisseria meningitidis escape from the bactericidal activity of a monoclonal antibody is mediated by phase variation of lgtG and enhanced by a mutator phenotype', *Infection and Immunity*, 76(11), pp. 5038–5048. doi: 10.1128/IAI.00395-08.

Bennett, J. S. *et al.* (2010) 'Independent evolution of the core and accessory gene sets in the genus Neisseria: insights gained from the genome of Neisseria lactamica isolate 020-06', *BMC Genomics*, 11(1), p. 652. doi: 1471-2164-11-652 [pii] 10.1186/1471-2164-11-652.

Bennett, J. S. *et al.* (2012) 'A genomic approach to bacterial taxonomy: An examination and proposed reclassification of species within the genus Neisseria', *Microbiology (United Kingdom)*, 158(6), pp. 1570–1580. doi: 10.1099/mic.0.056077-0.

Bennett, J. S., Jolley, K. A. and Maiden, M. C. J. (2013) 'Genome sequence analyses show that Neisseria oralis is the same species as "Neisseria mucosa var. Heidelbergensis"', *International Journal of Systematic and Evolutionary Microbiology*, 63(PART10), pp. 3920–3926. doi: 10.1099/ijs.0.052431-0.

Bentley, D. R. *et al.* (2008) 'Accurate whole human genome sequencing using reversible terminator chemistry', *Nature*, 456(7218), pp. 53–59. doi: 10.1038/nature07517.

Bentley, S. D. *et al.* (2007) 'Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18', *PLoS Genetics*, 3(2), pp. 0230–0240. doi: 10.1371/journal.pgen.0030023.

Bergval, I. L. *et al.* (2009) 'Resistant mutants of Mycobacterium tuberculosis selected in vitro do not reflect the in vivo mechanism of isoniazid resistance.', *The Journal of antimicrobial chemotherapy*. England, 64(3), pp. 515–523. doi: 10.1093/jac/dkp237.

Bidmos, F. and Bayliss, C. (2014) 'Genomic and Global Approaches to Unravelling How Hypermutable Sequences Influence Bacterial Pathogenesis', *Pathogens*, 3(1), pp. 164–184. doi: 10.3390/pathogens3010164.

Bille, E. *et al.* (2005) 'A chromosomally integrated bacteriophage in invasive meningococci', *The Journal of Experimental Medicine*, 201(12), pp. 1905–1913. doi: 10.1084/jem.20050112.

Bille, E. *et al.* (2017) 'A virulence-associated filamentous bacteriophage of Neisseria meningitidis increases host-cell colonisation', *PLoS Pathogens*, 13(7). doi: 10.1371/journal.ppat.1006495.

Bjune, G. *et al.* (1991) 'Effect of outer membrane vesicle vaccine against group B meningococcal

disease in Norway', *The Lancet*, 338(8775), pp. 1093–1096. doi: 10.1016/0140-6736(91)91961-S.

Black, C. G., Fyfe, J. A. M. and Davies, J. K. (1995) 'A promoter associated with the Neisserial repeat can be used to transcribe the uvrB gene from Neisseria gonorrhoeae', *Journal of Bacteriology*, 177(8), pp. 1952–1958. doi: 10.1128/jb.177.8.1952-1958.1995.

Bleidorn, C. (2015) 'Third generation sequencing: technology and its potential impact on evolutionary biodiversity research', *Systematics and Biodiversity*, 2000(January), pp. 1–8. doi: 10.1080/14772000.2015.1099575.

Blount, Z. D., Borland, C. Z. and Lenski, R. E. (2008) 'Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli', *Proceedings of the National Academy of Sciences*, 105(23), pp. 7899–7906. doi: 10.1073/pnas.0803151105.

Bobay, L. M., Rocha, E. P. C. and Touchon, M. (2013) 'The adaptation of temperate bacteriophages to their host genomes', *Molecular Biology and Evolution*, 30(4), pp. 737–751. doi: 10.1093/molbev/mss279.

De Boeck, I. *et al.* (2017) 'Comparing the healthy nose and nasopharynx microbiota reveals continuity as well as niche-specificity', *Frontiers in Microbiology*, 8(NOV). doi: 10.3389/fmicb.2017.02372.

Bohlin, J. *et al.* (2017) 'The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes', *BMC Genomics*, 18(1). doi: 10.1186/s12864-017-3543-7.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Borrow, R. *et al.* (2013) 'Effectiveness of meningococcal serogroup C vaccine programmes', *Vaccine*, pp. 4477–4486. doi: 10.1016/j.vaccine.2013.07.083.

Børud, B. *et al.* (2010) 'Genetic, structural, and antigenic analyses of glycan diversity in the O-linked protein glycosylation systems of human Neisseria species', *Journal of Bacteriology*, 192(11), pp. 2816–2829. doi: 10.1128/JB.00101-10.

Bratcher, H. B. *et al.* (2014) 'A gene-by-gene population genomics platform: De novo assembly, annotation and genealogical analysis of 108 representative Neisseria meningitidis genomes', *BMC Genomics*, 15(1). doi: 10.1186/1471-2164-15-1138.

Brown, N. P., Leroy, C. and Sander, C. (1998) 'MView: a web-compatible database search or

multiple alignment viewer', *Bioinformatics*, 14(4), pp. 380–381. doi: 10.1093/bioinformatics/14.4.380.

Bruge, J. *et al.* (2004) 'Clinical evaluation of a group B meningococcal N-propionylated polysaccharide conjugate vaccine in adult, male volunteers', *Vaccine*, 22(9–10), pp. 1087–1096. doi: 10.1016/j.vaccine.2003.10.005.

Brunelli, B. *et al.* (2011) 'Influence of sequence variability on bactericidal activity sera induced by Factor H binding protein variant 1.1', *Vaccine*, 29(5), pp. 1072–1081. doi: 10.1016/j.vaccine.2010.11.064.

Bryant, J., Chewapreecha, C. and Bentley, S. D. (2012) 'Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences.', *Future microbiology*, 7(11), pp. 1283–1296. doi: 10.2217/fmb.12.108.

Budroni, S. *et al.* (2011) 'Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination', *Proceedings of the National Academy of Sciences*, 108(11), pp. 4494–4499. doi: 10.1073/pnas.1019751108.

Camacho, C. *et al.* (2009) 'BLAST+: Architecture and applications', *BMC Bioinformatics*, 10. doi: 10.1186/1471-2105-10-421.

Cambray, G. and Mazel, D. (2008) 'Synonymous genes explore different evolutionary landscapes', *PLoS Genetics*, 4(11). doi: 10.1371/journal.pgen.1000256.

Canchaya, C. *et al.* (2003) 'Prophage Genomics', *Microbiology and Molecular Biology Reviews*, 67(2), pp. 238–276. doi: 10.1128/MMBR.67.2.238-276.2003.

Cartwright, K. A. V. *et al.* (1987) 'The Stonehouse survey: nasopharyngeal carriage of meningococci and Neisseria lactamica', *Epidemiology and Infection*, 99(3), pp. 591–601. doi: 10.1017/S0950268800066449.

Castresana, J. (2000) 'Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis', *Molecular Biology and Evolution*, 17(4), pp. 540–552. doi: 10.1093/oxfordjournals.molbev.a026334.

Caugant, D. A. *et al.* (1986) 'Intercontinental spread of a genetically distinctive complex of clones of *Neisseria meningitidis* causing epidemic disease.', *Proceedings of the National Academy of Sciences of the United States of America*, 83(13), pp. 4927–4931. doi: 10.1073/pnas.83.13.4927.

Caugant, D. A. and Maiden, M. C. J. (2009) 'Meningococcal carriage and disease-Population

biology and evolution', *Vaccine*, 27(SUPPL. 2). doi: 10.1016/j.vaccine.2009.04.061.

Chaguza, C. *et al.* (2016) 'Recombination in Streptococcus pneumoniae Lineages Increase with Carriage Duration and Size of the Polysaccharide Capsule', *mBio* , 7(5). doi: 10.1128/mBio.01053-16.

Chang, Q., Tzeng, Y. L. and Stephens, D. S. (2012) 'Meningococcal disease: Changes in epidemiology and prevention', *Clinical Epidemiology*, pp. 237–245. doi: 10.2147/CLEP.S28410.

Charlson, E. S. *et al.* (2010) 'Disordered microbial communities in the upper respiratory tract of cigarette smokers', *PLoS ONE*, 5(12). doi: 10.1371/journal.pone.0015216.

Chaves-Moreno, D. *et al.* (2015) 'Application of a novel "Pan-genome"-based strategy for assigning RNAseq transcript reads to Staphylococcus aureus strains', *PLoS ONE*, 10(12). doi: 10.1371/journal.pone.0145861.

Chonmaitree, T. *et al.* (2017) 'Nasopharyngeal microbiota in infants and changes during viral upper respiratory tract infection and acute otitis media', *PLoS ONE*, 12(7). doi: 10.1371/journal.pone.0180630.

Christensen, H. *et al.* (2014) 'Re-evaluating cost effectiveness of universal meningitis vaccination (Bexsero) in England: Modelling study', *BMJ (Online)*, 349. doi: 10.1136/bmj.g5725.

Cleary, D. W. and Clarke, S. C. (2017) 'The nasopharyngeal microbiome', *Emerging Topics in Life Sciences*, 1(4), p. 297 LP-312. Available at: http://www.emergtoplifesci.org/content/1/4/297.abstract.

Cleary, P. R. *et al.* (2016) 'Variations in Neisseria meningitidis carriage by socioeconomic status: a cross-sectional study', *Journal of public health (Oxford, England)*, 38(1), pp. 61–70. doi: 10.1093/pubmed/fdv015.

Conesa, A. *et al.* (2005) 'Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research', *Bioinformatics*, 21(18), pp. 3674–3676. doi: 10.1093/bioinformatics/bti610.

Corander, J. *et al.* (2012) 'Population structure in the Neisseria, and the biological significance of fuzzy species', *Journal of The Royal Society Interface*, 9(71), pp. 1208–1215. doi: 10.1098/rsif.2011.0601.

Cremers, A. J. H. *et al.* (2014) 'The adult nasopharyngeal microbiome as a determinant of pneumococcal acquisition', *Microbiome*, 2(1). doi: 10.1186/2049-2618-2-44.

References

Criss, A. K. and Seifert, H. S. (2012) 'A bacterial siren song: intimate interactions between Neisseria and neutrophils.', *Nature reviews. Microbiology*, 10(3), pp. 178–190. doi: 10.1038/nrmicro2713.

Croucher, N. J. *et al.* (2011) 'Rapid pneumococcal evolution in response to clinical interventions.', *Science (New York, N.Y.)*, 331(6016), pp. 430–4. doi: 10.1126/science.1198545.

Croucher, N. J. *et al.* (2014) 'Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins.', *Nucleic acids research*, p. gku1196--. doi: 10.1093/nar/gku1196.

Crum-Cianflone, N. and Sullivan, E. (2016) 'Meningococcal Vaccinations', *Infectious Diseases and Therapy*, pp. 89–112. doi: 10.1007/s40121-016-0107-0.

Cuthbertson, J. M., Doyle, D. A. and Sansom, M. S. P. (2005) 'Transmembrane helix prediction: a comparative evaluation and analysis.', *Protein engineering, design & selection : PEDS*, 18(6), pp. 295–308. doi: 10.1093/protein/gzi032.

Darmon, E. and Leach, D. R. F. (2014) 'Bacterial genome instability.', *Microbiology and molecular biology reviews : MMBR*, 78(1), pp. 1–39. doi: 10.1128/MMBR.00035-13.

Deasy, a. M. *et al.* (2015) 'Nasal inoculation of the commensal Neisseria lactamica inhibits carriage of Neisseria meningitidis by young adults: a controlled human infection study', *Clinical Infectious Diseases*, 60(10), pp. 1512–1520. doi: 10.1093/cid/civ098.

Deatherage, D. E. and Barrick, J. E. (2014) 'Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using *breseq*', *Methods in Molecular Biology*, 1151, pp. 165–188. doi: 10.1007/978-1-4939-0554-6.

DEFRA (2017) *Application for consent to release a GMO – organisms other than higher plants*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/682642/southampton-17r5001-parta1.pdf.

Van Deuren, M., Brandtzaeg, P. and Van Der Meer, J. W. M. (2000) 'Update on meningococcal disease with emphasis on pathogenesis and clinical management', *Clinical Microbiology Reviews*, 13(1), pp. 144–166. doi: 10.1128/CMR.13.1.144-166.2000.

Devoe, I. W. and Gilchrist, J. E. (1975) 'Pili on meningococci from primary cultures of nasopharyngeal carriers and cerebrospinal fluid of patients with acute disease', *The Journal of Experimental Medicine*, pp. 297–305.

References

Didelot, X., Dordel, J., *et al.* (2016) 'Genomic analysis and comparison of two gonorrhea outbreaks', *mBio*, 7(3). doi: 10.1128/mBio.00525-16.

Didelot, X., Walker, A. S., *et al.* (2016) 'Within-host evolution of bacterial pathogens', *Nature Reviews Microbiology*, 14(3), pp. 150–162. doi: 10.1038/nrmicro.2015.13.

Didelot, X. and Maiden, M. C. J. (2010) 'Impact of recombination on bacterial evolution', *Trends in Microbiology*, pp. 315–322. doi: 10.1016/j.tim.2010.04.002.

Didelot, X. and Wilson, D. J. (2015) 'ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes', *PLoS Computational Biology*, 11(2). doi: 10.1371/journal.pcbi.1004041.

Drake, S. L., Sandstedt, S. A. and Koomey, M. (1997) 'PilP, a pilus biogenesis lipoprotein in Neisseria gonorrhoeae, affects expression of PilQ as a high-molecular-mass multimer.', *Molecular microbiology*, 23(4), pp. 657–668. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9157238.

Durando, P., Faust, S. N. and Torres, A. (2015) 'Immunological features and clinical benefits of conjugate vaccines against bacteria', *Journal of Immunology Research*. doi: 10.1155/2015/934504.

Earl, D. *et al.* (2011) 'Assemblathon 1: A competitive assessment of de novo short read assembly methods', *Genome Research*, pp. 2224–2241. doi: 10.1101/gr.126599.111.

Ekblom, R. and Wolf, J. B. W. (2014) 'A field guide to whole-genome sequencing, assembly and annotation', *Evolutionary Applications*, pp. 1026–1042. doi: 10.1111/eva.12178.

Van Der Ende, A., Hopman, C. T. P. and Dankert, J. (1999) 'Deletion of porA by recombination between clusters of repetitive extragenic palindromic sequences in Neisseria meningitidis', *Infection and Immunity*, 67(6), pp. 2928–2934.

Enright, A. J., Van Dongen, S. and Ouzounis, C. A. (2002) 'An efficient algorithm for large-scale detection of protein families.', *Nucleic acids research*, 30(7), pp. 1575–1584. doi: doi: 10.1093/nar/30.7.1575.

Evans, C. M. *et al.* (2011) 'Nasopharyngeal colonization by Neisseria lactamica and induction of protective immunity against Neisseria meningitidis', *Clinical Infectious Diseases*, 52(1), pp. 70–77. doi: 10.1093/cid/ciq065.

Everitt, R. G. *et al.* (2014) 'Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus', *Nature Communications*, 5. doi: 10.1038/ncomms4956.

Ezewudo, M. N. *et al.* (2015) 'Population structure of Neisseria gonorrhoeae based on whole

genome data and its relationship with antibiotic resistance.', *PeerJ*, 3, p. e806. doi: 10.7717/peerj.806.

Fang, G. *et al.* (2012) 'Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing.', *Nature biotechnology*, 30(12), pp. 1232–1239. doi: 10.1038/nbt.2432.

Faridmoayer, A. *et al.* (2008) 'Extreme substrate promiscuity of the Neisseria oligosaccharyl transferase involved in protein O-glycosylation', *Journal of Biological Chemistry*, 283(50), pp. 34596–34604. doi: 10.1074/jbc.M807113200.

Findlow, J. *et al.* (2010) 'Multicenter, open-label, randomized phase II controlled trial of an investigational recombinant Meningococcal serogroup B vaccine with and without outer membrane vesicles, administered in infancy.', *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 51(10), pp. 1127–1137. doi: 10.1086/656741.

Finn, R. D. *et al.* (2014) 'Pfam: The protein families database', *Nucleic Acids Research*. doi: 10.1093/nar/gkt1223.

Finne, J., Leinonen, M. and Mäkelä, P. H. (1983) 'ANTIGENIC SIMILARITIES BETWEEN BRAIN COMPONENTS AND BACTERIA CAUSING MENINGITIS. Implications for Vaccine Development and Pathogenesis', *The Lancet*, 322(8346), pp. 355–357. doi: 10.1016/S0140-6736(83)90340-9.

Fiorito, T. *et al.* (2017) *Adverse Events Following Vaccination With Bivalent rLP2086 (Trumenba®): An Observational, Longitudinal Study During a College Outbreak and a Systematic Review*, *The Pediatric Infectious Disease Journal*. doi: 10.1097/INF.0000000000001742.

Fleischmann, R. *et al.* (1995) 'Whole-genome random sequencing and assembly of Haemophilus influenzae Rd', *Science*, 269(5223), pp. 496–512. doi: 10.1126/science.7542800.

Frosch, M. and Maiden, M. C. J. (2006) *Handbook of Meningococcal Disease: Infection Biology, Vaccination, Clinical Management*, *Handbook of Meningococcal Disease: Infection Biology, Vaccination, Clinical Management*. doi: 10.1002/3527608508.

Frye, S. A. *et al.* (2013) 'Dialects of the DNA Uptake Sequence in Neisseriaceae', *PLoS Genetics*, 9(4). doi: 10.1371/journal.pgen.1003458.

Fu, L. *et al.* (2012) 'CD-HIT: Accelerated for clustering the next-generation sequencing data', *Bioinformatics*, 28(23), pp. 3150–3152. doi: 10.1093/bioinformatics/bts565.

Garcia-Rodriguez, J. A. (2002) 'Dynamics of nasopharyngeal colonization by potential respiratory

pathogens', *Journal of Antimicrobial Chemotherapy*, 50(90003), pp. 59–74. doi: 10.1093/jac/dkf506.

Gasteiger, E. *et al.* (2003) 'ExPASy: The proteomics server for in-depth protein knowledge and analysis', *Nucleic Acids Research*, 31(13), pp. 3784–3788. doi: 10.1093/nar/gkg563.

Geer, L. Y. *et al.* (2002) 'CDART: Protein homology by domain architecture', *Genome Research*, 12(10), pp. 1619–1623. doi: 10.1101/gr.278202.

Ghalayini, M. *et al.* (2018) '"Evolution of a dominant natural isolate of *Escherichia coli* in the human gut over a year suggests a neutral evolution with reduced effective population size"', *Applied and Environmental Microbiology*, p. AEM.02377-17. doi: 10.1128/AEM.02377-17.

Gold, R. *et al.* (1975) 'Clinical evaluation of group A and group C meningococcal polysaccharide vaccines in infants', *Journal of Clinical Investigation*, 56(6), pp. 1536–1547. doi: 10.1172/JCI108235.

Gold, R. *et al.* (1978) 'Carriage of Neisseria meningitidis and Neisseria lactamica in infants and children.', *The Journal of infectious diseases*, 137(2), pp. 112–121. doi: 10.1093/infdis/137.2.112.

Goldberg, A. *et al.* (2014) 'Systematic identification and quantification of phase variation in commensal and pathogenic Escherichia coli', *Genome Medicine*, 6(11). doi: 10.1186/s13073-014-0112-4.

Golubchik, T. *et al.* (2013) 'Within-Host Evolution of Staphylococcus aureus during Asymptomatic Carriage', *PLoS ONE*, 8(5). doi: 10.1371/journal.pone.0061319.

Gorringe, A. R. (2005) 'Can Neisseria lactamica antigens provide an effective vaccine to prevent meningococcal disease?', *Expert Review of Vaccines*, pp. 373–379. doi: 10.1586/14760584.4.3.373.

Gorringe, A. R. *et al.* (2009) 'Phase I safety and immunogenicity study of a candidate meningococcal disease vaccine based on Neisseria lactamica outer membrane vesicles', *Clinical and Vaccine Immunology*, 16(8), pp. 1113–1120. doi: 10.1128/CVI.00118-09.

Gorringe, A. R. and Pajon, R. (2012) 'Bexsero: A multicomponent vaccine for prevention of meningococcal disease', *Human Vaccines and Immunotherapeutics*, pp. 174–183. doi: 10.4161/hv.18500.

Grad, Y. H. *et al.* (2014) 'Genomic epidemiology of Neisseria gonorrhoeae with reduced susceptibility to cefixime in the USA: A retrospective observational study', *The Lancet Infectious*

Diseases, 14(3), pp. 220–226. doi: 10.1016/S1473-3099(13)70693-5.

Granoff, D. M. *et al.* (1998) 'Induction of Immunologic Refractoriness in Adults by Meningococcal C Polysaccharide Vaccination', *The Journal of Infectious Diseases*. Oxford University Press, 178(3), pp. 870–874. Available at: http://www.jstor.org/stable/30114347.

Granoff, D. M. (2010) 'Review of Meningococcal Group B Vaccines', *Clinical Infectious Diseases*, 50(s2), pp. S54–S65. doi: 10.1086/648966.

Grant, J. R., Arantes, A. S. and Stothard, P. (2012) 'Comparing thousands of circular genomes using the CGView Comparison Tool', *BMC Genomics*, 13(1). doi: 10.1186/1471-2164-13-202.

Gray, S. J. *et al.* (2006) 'Epidemiology of meningococcal disease in England and Wales 1993/94 to 2003/04: Contribution and experiences of the Meningococcal Reference Unit', *Journal of Medical Microbiology*, 55(7), pp. 887–896. doi: 10.1099/jmm.0.46288-0.

Greenberg, D. *et al.* (2006) 'The contribution of smoking and exposure to tobacco smoke to Streptococcus pneumoniae and Haemophilus influenzae carriage in children and their mothers.', *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 42(7), pp. 897–903. doi: 10.1086/500935.

Grissa, I., Vergnaud, G. and Pourcel, C. (2007) 'CRISPRFinder: a web tool to identify clustered regularly interspace short palindromic repeats', *Nucleic Acids Research*, 35(Web Server issue), pp. 52–57. doi: 10.1093/nar/gkn228.

Guindon, S. *et al.* (2009) 'Estimating maximum likelihood phylogenies with PhyML', *Methods in Molecular Biology*, 537, pp. 113–137. doi: 10.1007/978-1-59745-251-9_6.

Hao, W. *et al.* (2011) 'Extensive genomic variation within clonal complexes of Neisseria meningitidis', *Genome Biology and Evolution*, 3(1), pp. 1406–1418. doi: 10.1093/gbe/evr119.

Harrison, O. B. *et al.* (2015) 'Genomic Analysis of the Evolution and Global Spread of Hyper-invasive Meningococcal Lineage 5.', *EBioMedicine*, 2(3), pp. 234–243. doi: 10.1016/j.ebiom.2015.01.004.

Harrison, O. B. *et al.* (2017) 'Genomic analysis of urogenital and rectal Neisseria meningitidis isolates reveals encapsulated hyperinvasive meningococci and coincident multidrug-resistant gonococci', *Sexually Transmitted Infections*, 93(6), pp. 445–451. doi: 10.1136/sextrans-2016-052781.

Hasman, H. *et al.* (2014) 'Rapid whole-genome sequencing for detection and characterization of

microorganisms directly from clinical samples', *Journal of Clinical Microbiology*, 52(1), pp. 139–146. doi: 10.1128/JCM.02452-13.

Heather, J. M. and Chain, B. (2016) 'The sequence of sequencers: The history of sequencing DNA', *Genomics*, pp. 1–8. doi: 10.1016/j.ygeno.2015.11.003.

Henson, J., Tischler, G. and Ning, Z. (2012) 'Next-generation sequencing and large genome assemblies', *Pharmacogenomics*, 13(8), pp. 901–915. doi: 10.2217/pgs.12.72.

Hill, D. J. *et al.* (2010) 'Cellular and molecular biology of Neisseria meningitidis colonization and invasive disease.', *Clinical science (London, England : 1979)*, 118(9), pp. 547–64. doi: 10.1042/CS20090513.

Hiller, K. *et al.* (2004) 'PrediSi: Prediction of signal peptides and their cleavage positions', *Nucleic Acids Research*, 32(WEB SERVER ISS.). doi: 10.1093/nar/gkh378.

Holt, K. E. *et al.* (2008) 'High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi', *Nat Genet*, 40. doi: 10.1038/ng.195.

Huang, Y. F. *et al.* (2012) 'Palindromic sequence impedes sequencing-by-ligation mechanism', *BMC Systems Biology*, 6(SUPPL.2). doi: 10.1186/1752-0509-6-S2-S10.

Hugerth, L. W. and Andersson, A. F. (2017) 'Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing', *Frontiers in Microbiology*. doi: 10.3389/fmicb.2017.01561.

Hughes, A. L. *et al.* (2008) 'Synonymous and nonsynonymous polymorphisms versus divergences in bacterial genomes', *Molecular Biology and Evolution*, 25(10), pp. 2199–2209. doi: 10.1093/molbev/msn166.

Huson, D. H. and Bryant, D. (2006) 'Application of phylogenetic networks in evolutionary studies', *Molecular Biology and Evolution*, pp. 254–267. doi: 10.1093/molbev/msj030.

Hyatt, D. *et al.* (2010) 'Prodigal: prokaryotic gene recognition and translation initiation site identification.', *BMC bioinformatics*, 11, p. 119. doi: 10.1186/1471-2105-11-119.

Illumina (2017) *An introduction to Next-Generation Sequencing Technology*. Available at: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf.

Jerome, J. P. *et al.* (2011) 'Standing Genetic Variation in Contingency Loci Drives the Rapid Adaptation of {<}italic{>}Campylobacter jejuni{<}/italic{>} to a Novel Host', *PLoS ONE*. Public

Library of Science, 6(1), p. e16399. Available at: http://dx.doi.org/10.1371%7B%25%7D2Fjournal.pone.0016399.

Johswich, K. O. *et al.* (2012) 'Invasive potential of nonencapsulated disease isolates of neisseria: Meningitidis', *Infection and Immunity*, 80(7), pp. 2346–2353. doi: 10.1128/IAI.00293-12.

Jolley, K. A. and Maiden, M. C. J. (2010) 'BIGSdb: Scalable analysis of bacterial genome variation at the population level', *BMC bioinformatics*, 11(1), p. 595. doi: 10.1186/1471-2105-11-595.

Joseph, B. *et al.* (2011) 'Virulence evolution of the human pathogen neisseria meningitidis by recombination in the core and accessory genome', *PLoS ONE*, 6(4). doi: 10.1371/journal.pone.0018441.

Katoh, K. and Standley, D. M. (2013) 'MAFFT multiple sequence alignment software version 7: Improvements in performance and usability', *Molecular Biology and Evolution*, 30(4), pp. 772–780. doi: 10.1093/molbev/mst010.

Kawai, M., Uchiyama, I. and Kobayashi, I. (2005) 'Genome comparison In Silico in Neisseria suggests integration of filamentous bacteriophages by their own transposase', *DNA Research*, 12(6), pp. 389–401. doi: 10.1093/dnares/dsi021.

Kchouk, M., Gibrat, J.-F. and Elloumi, M. (2017) 'Generations of Sequencing Technologies: From First to Next Generation', *Biology and Medicine*. OMICS International.,. doi: 10.4172/0974-8369.1000395.

Kellogg, D. S. J. *et al.* (1968) 'Neisseria gonorrhoeae. II. Colonial variation and pathogenicity during 35 months in vitro.', *Journal of bacteriology*. United States, 96(3), pp. 596–605.

Kent, W. J. (2002) 'BLAT — The BLAST -Like Alignment Tool', *Genome Research*, 12, pp. 656–664. doi: 10.1101/gr.229202.

De Kleijn, E. *et al.* (2001) 'Serum bactericidal activity and isotype distribution of antibodies in toddlers and schoolchildren after vaccination with RIVM hexavalent PorA vesicle vaccine', in *Vaccine*, pp. 352–358. doi: 10.1016/S0264-410X(01)00371-1.

Klockgether, J. *et al.* (2013) 'Intraclonal diversity of the Pseudomonas aeruginosa cystic fibrosis airway isolates TBCF10839 and TBCF121838: Distinct signatures of transcriptome, proteome, metabolome, adherence and pathogenicity despite an almost identical genome sequence', *Environmental Microbiology*, 15(1), pp. 191–210. doi: 10.1111/j.1462-2920.2012.02842.x.

Kong, Y. *et al.* (2013) 'Homologous recombination drives both sequence diversity and gene

content variation in Neisseria meningitidis', *Genome Biology and Evolution*, 5(9), pp. 1611–1627. doi: 10.1093/gbe/evt116.

Koren, S. *et al.* (2013) 'Reducing assembly complexity of microbial genomes with single-molecule sequencing.', *Genome biology*, 14(9), p. R101. doi: 10.1186/gb-2013-14-9-r101.

Koskiniemi, S. *et al.* (2012) 'Selection-driven gene loss in bacteria', *PLoS Genetics*, 8(6). doi: 10.1371/journal.pgen.1002787.

Krogh, a *et al.* (2001) 'Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.', *Journal of molecular biology*, 305(3), pp. 567–580. doi: 10.1006/jmbi.2000.4315.

Kroll, J. S. *et al.* (1998) 'Natural genetic exchange between Haemophilus and Neisseria: intergeneric transfer of chromosomal genes between major human pathogens.', *Proceedings of the National Academy of Sciences of the United States of America*, 95(21), pp. 12381–12385. doi: 10.1073/pnas.95.21.12381.

Kulski, J. K. (2016) 'Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications', in *Next Generation Sequencing - Advances, Applications and Challenges*. doi: 10.5772/61964.

Lamelas, A. *et al.* (2014) 'Emergence of a new epidemic neisseria meningitidis serogroup a clone in the African meningitis belt: High-resolution picture of genomic changes that mediate immune evasion', *mBio*, 5(5). doi: 10.1128/mBio.01974-14.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2.', *Nature methods*, 9(4), pp. 357–9. doi: 10.1038/nmeth.1923.

Laver, J. R., Hughes, S. E. and Read, R. C. (2015) 'Neisserial Molecular Adaptations to the Nasopharyngeal Niche', *Advances in Microbial Physiology*, 66, pp. 323–355. doi: 10.1016/bs.ampbs.2015.05.001.

Levade, I. *et al.* (2017) 'Vibrio cholerae genomic diversity within and between patients', *Microbial Genomics*. doi: 10.1099/mgen.0.000142.

Levene, M. J. *et al.* (2003) 'Zero-mode waveguides for single-molecule analysis at high concentrations.', *Science (New York, N.Y.)*, 299(5607), pp. 682–6. doi: 10.1126/science.1079700.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

References

Li, H. (2013) ‘Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM’, *arXiv preprint arXiv*, 0(0), p. 3. doi: arXiv:1303.3997 [q-bio.GN].

Li, H. and Durbin, R. (2009) ‘Fast and accurate short read alignment with Burrows-Wheeler transform’, *Bioinformatics*, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.

Li, Y. *et al.* (2006) ‘Immunization with live Neisseria lactamica protects mice against meningococcal challenge and can elicit serum bactericidal antibodies’, *Infection and Immunity*, 74(11), pp. 6348–6355. doi: 10.1128/IAI.01062-06.

Lin, Y. H., Ryan, C. S. and Davies, J. K. (2011) ‘Neisserial correia repeat-enclosed elements do not influence the transcription of pil genes in Neisseria gonorrhoeae and Neisseria meningitidis’, *Journal of Bacteriology*, 193(20), pp. 5728–5736. doi: 10.1128/JB.05526-11.

Ling, Z. *et al.* (2013) ‘Pyrosequencing analysis of the human microbiota of healthy Chinese undergraduates’, *BMC Genomics*, 14(1). doi: 10.1186/1471-2164-14-390.

Linz, B. *et al.* (2000) ‘Frequent interspecific genetic exchange between commensal neisseriae and Neisseria meningitidis’, *Molecular Microbiology*, pp. 1049–1058. doi: 10.1046/j.1365-2958.2000.01932.x.

Liu, S. V. *et al.* (2002) ‘Genome analysis and strain comparison of Correia repeats and Correia repeat-enclosed elements in pathogenic Neisseria’, *Journal of Bacteriology*, 184(22), pp. 6163–6173. doi: 10.1128/JB.184.22.6163-6173.2002.

Lo, H., Tang, C. M. and Exley, R. M. (2009) ‘Mechanisms of avoidance of host immunity by Neisseria meningitidis and its effect on vaccine development’, *The Lancet Infectious Diseases*, pp. 418–427. doi: 10.1016/S1473-3099(09)70132-X.

Loman, N. J. *et al.* (2012) ‘Performance comparison of benchtop high-throughput sequencing platforms’, *Nature Biotechnology*, 30(5), pp. 434–439. doi: 10.1038/nbt.2198.

Löytynoja, A. (2014) ‘Phylogeny-aware alignment with PRANK’, *Methods in Molecular Biology*, 1079, pp. 155–170. doi: 10.1007/978-1-62703-646-7_10.

Lucidarme, J. *et al.* (2013) ‘Genetic distribution of noncapsular meningococcal group B vaccine antigens in Neisseria lactamica’, *Clinical and Vaccine Immunology*, 20(9), pp. 1360–1369. doi: 10.1128/CVI.00090-13.

Lucidarme, J. *et al.* (2015) ‘Genomic resolution of an aggressive, widespread, diverse and expanding meningococcal serogroup B, C and W lineage’, *Journal of Infection*, 71(5), pp. 544–552.

doi: 10.1016/j.jinf.2015.07.007.

Łyskowski, A., Leo, J. C. and Goldman, A. (2011) 'Structure and biology of trimeric autotransporter adhesins', *Advances in Experimental Medicine and Biology*, 715, pp. 143–158. doi: 10.1007/978-94-007-0940-9_9.

MacLennan, J. *et al.* (2006) 'Social behavior and meningococcal carriage in British teenagers', *Emerging Infectious Diseases*, 12(6), pp. 950–957. doi: 10.3201/eid1206.051297.

Maiden, M. C. J. *et al.* (1998) 'Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms', *Proceedings of the National Academy of Sciences*, 95(6), pp. 3140–3145. doi: 10.1073/pnas.95.6.3140.

Maiden, M. C. J. and Harrison, O. B. (2016) 'Population and functional genomics of neisseria revealed with gene-by-gene approaches', *Journal of Clinical Microbiology*, pp. 1949–1955. doi: 10.1128/JCM.00301-16.

Maiden, M. C. J. and Stuart, J. M. (2002) 'Carriage of serogroup C meningococci 1 year after meningococcal C conjugate polysaccharide vaccination', *Lancet*, 359(9320), pp. 1829–1830. doi: 10.1016/S0140-6736(02)08679-8.

Mardis, E. R. (2008) 'Next-generation DNA sequencing methods.', *Annual review of genomics and human genetics*, 9, pp. 387–402. doi: 10.1146/annurev.genom.9.081307.164359.

Margulies, M. *et al.* (2005) 'Genome sequencing in microfabricated high-density picolitre reactors', *Nature*, 437(7057), pp. 376–380. doi: 10.1038/nature03959.

Marks, L. R., Reddinger, R. M. and Anders, P. (2012) 'High Levels of Genetic Recombination during Nasopharyngeal Carriage and Biofilm Formation in Streptococcus pneumoniae', *mBio*, 3(5), pp. e00200-12. doi: 10.1128/mBio.00200-12.Editor.

Marri, P. R. *et al.* (2010) 'Genome sequencing reveals widespread virulence gene exchange among human Neisseria species', *PLoS ONE*, 5(7). doi: 10.1371/journal.pone.0011835.

Martinon-Torres, F. *et al.* (2014) 'A randomized, phase 1/2 trial of the safety, tolerability, and immunogenicity of bivalent rLP2086 meningococcal B vaccine in healthy infants', *Vaccine*, 32(40), pp. 5206–5211. doi: 10.1016/j.vaccine.2014.07.049.

Marvig, R. L. *et al.* (2015) 'Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis.', *Nature genetics*, 47(1), pp. 57–64. doi: 10.1038/ng.3148.

Masignani, V. *et al.* (2001) 'Mu-like prophage in serogroup B Neisseria meningitidis coding for

surface-exposed antigens', *Infection and Immunity*, 69(4), pp. 2580–2588. doi: 10.1128/IAI.69.4.2580-2588.2001.

Mathee, K. *et al.* (2008) 'Dynamics of Pseudomonas aeruginosa genome evolution', *Proceedings of the National Academy of Sciences*, 105(8), pp. 3100–3105. doi: 10.1073/pnas.0711982105.

McGee, Z. A. *et al.* (1979) 'Pili of Neisseria meningitidis: Effect of media on maintenance of piliation, characteristics of pili, and colonial morphology', *Infection and Immunity*, 24(1), pp. 194–201.

McKernan, K. J. *et al.* (2009) 'Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding', *Genome Research*, 19(9), pp. 1527–1541. doi: 10.1101/gr.091868.109.

McNamara, L. A. *et al.* (2017) 'Meningococcal Carriage Following a Vaccination Campaign With MenB-4C and MenB-FHbp in Response to a University Serogroup B Meningococcal Disease Outbreak-Oregon, 2015-2016', *The Journal of infectious diseases*, 216(9), pp. 1130–1140. doi: 10.1093/infdis/jix446.

Medini, D. *et al.* (2005) 'The microbial pan-genome', *Current Opinion in Genetics and Development*, pp. 589–594. doi: 10.1016/j.gde.2005.09.006.

Méric, G. *et al.* (2014) 'A reference pan-genome approach to comparative bacterial genomics: Identification of novel epidemiological markers in pathogenic Campylobacter', *PLoS ONE*, 9(3). doi: 10.1371/journal.pone.0092798.

Miller, E., Salisbury, D. and Ramsay, M. (2001) 'Planning, registration, and implementation of an immunisation campaign against meningococcal serogroup C disease in the UK: A success story', *Vaccine*, 20(SUPPL. 1). doi: 10.1016/S0264-410X(01)00299-7.

Milne, I. *et al.* (2009) 'Tablet-next generation sequence assembly visualization', *Bioinformatics*, pp. 401–402. doi: 10.1093/bioinformatics/btp666.

Moran, P. (Coastal Z. and E. S. D. N. F. S. C. (1994) *OVERVIEW OF COMMONLY USED DNA TECHNIQUES*, *NOAA-NMFS-NWFSC TM-17: Application of DNA Technology to the Management of Pacific Salmon*. Available at: https://www.nwfsc.noaa.gov/publications/scipubs/techmemos/tm17/papers/moran.htm (Accessed: 15 March 2018).

Moxon, R., Bayliss, C. and Hood, D. (2006) 'Bacterial Contingency Loci: The Role of Simple Sequence DNA Repeats in Bacterial Adaptation', *Annual Review of Genetics*, 40(1), pp. 307–333.

doi: 10.1146/annurev.genet.40.110405.090442.

Muir, P. *et al.* (2016) 'The real cost of sequencing: Scaling computation to keep pace with data generation', *Genome Biology*, 17(1). doi: 10.1186/s13059-016-0917-0.

Mulhall, R. M. *et al.* (2016) 'Resolution of a protracted Serogroup B meningococcal outbreak with whole-genome sequencing shows interspecies genetic transfer', *Journal of Clinical Microbiology*, 54(12), pp. 2891–2899. doi: 10.1128/JCM.00881-16.

Murphy, E. *et al.* (2009) 'Sequence diversity of the factor H binding protein vaccine candidate in epidemiologically relevant strains of serogroup B Neisseria meningitidis.', *The Journal of infectious diseases*, 200(3), pp. 379–389. doi: 10.1086/600141.

Mustapha, M. M., Marsh, J. W. and Harrison, L. H. (2016) 'Global epidemiology of capsular group W meningococcal disease (1970-2015): Multifocal emergence and persistence of hypervirulent sequence type (ST)-11 clonal complex', *Vaccine*, pp. 1515–1523. doi: 10.1016/j.vaccine.2016.02.014.

Namouchi, A. *et al.* (2012) 'After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection', *Genome Research*, 22(4), pp. 721–734. doi: 10.1101/gr.129544.111.

NHS (2018) *Meningitis Complcations*. Available at: https://www.nhs.uk/conditions/meningitis/complications/.

Nielsen, R. (2005) 'Molecular Signatures of Natural Selection', *Annual Review of Genetics*, 39(1), pp. 197–218. doi: 10.1146/annurev.genet.39.073003.112420.

Nyrén, P. (1987) 'Enzymatic method for continuous monitoring of DNA polymerase activity.', *Analytical biochemistry*, 167(2), pp. 235–238. doi: 10.1016/0003-2697(87)90158-8.

Ogata, H. *et al.* (1999) 'KEGG: Kyoto encyclopedia of genes and genomes', *Nucleic Acids Research*, pp. 29–34. doi: 10.1093/nar/27.1.29.

Oldfield, N. J. *et al.* (2013) 'Prevalence and Phase Variable Expression Status of Two Autotransporters, NalP and MspA, in Carriage and Disease Isolates of Neisseria meningitidis', *PLoS ONE*, 8(7). doi: 10.1371/journal.pone.0069746.

Oliver, K. J. *et al.* (2002) 'Neissetia lactamica protects against experimental meningococcal infection', *Infection and Immunity*, 70(7), pp. 3621–3626. doi: 10.1128/IAI.70.7.3621-3626.2002.

Page, A. J. *et al.* (2015) 'Roary: Rapid large-scale prokaryote pan genome analysis', *Bioinformatics*,

31(22), pp. 3691–3693. doi: 10.1093/bioinformatics/btv421.

Palmer, M. E. *et al.* (2013) 'Broad conditions favor the evolution of phase-variable loci.', *mBio*, 4(1). doi: 10.1128/mBio.00430-12.

Pandey, A. K. *et al.* (2017) 'Neisseria lactamica Y92-1009 complete genome sequence', *Standards in Genomic Sciences*, 12, p. 41. doi: 10.1186/s40793-017-0250-6.

Parikh, S. R. *et al.* (2016) 'Effectiveness and impact of a reduced infant schedule of 4CMenB vaccine against group B meningococcal disease in England: a national observational cohort study', *The Lancet*, 388(10061), pp. 2775–2782. doi: 10.1016/S0140-6736(16)31921-3.

Parkhill, J. *et al.* (2000) 'Complete DNA sequence of a serogroup A strain of Neisseria meningitidis 79491', *Nature*, 404(6777), pp. 502–506. doi: 10.1038/35006655.

Pearson, T. a and Manolio, T. a (2008) 'How to interpret a genome-wide association study.', *JAMA : the journal of the American Medical Association*, 299(11), pp. 1335–44. doi: 10.1001/jama.299.11.1335.

Peltola, H. *et al.* (1992) 'Rapid disappearance of Haemophilus influenzae type b meningitis after routine childhood immunisation with conjugate vaccines', *The Lancet*, 340(8819), pp. 592–594. doi: 10.1016/0140-6736(92)92117-X.

Petersen, T. N. *et al.* (2011) 'SignalP 4.0: discriminating signal peptides from transmembrane regions', *Nature methods*, 8(10), pp. 785–786. doi: 10.1038/nmeth.1701.

Pfeifer, B. *et al.* (2014) 'PopGenome: An efficient swiss army knife for population genomic analyses in R', *Molecular Biology and Evolution*, 31(7), pp. 1929–1936. doi: 10.1093/molbev/msu136.

Piekarowicz, A. *et al.* (2007) 'Characterization of the dsDNA prophage sequences in the genome of Neisseria gonorrhoeae and visualization of productive bacteriophage', *BMC Microbiology*, 7. doi: 10.1186/1471-2180-7-66.

Pizza, M. *et al.* (2000) 'Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing', *Science*, 287(5459), pp. 1816–1820. doi: 10.1126/science.287.5459.1816.

Pompanon, F. *et al.* (2005) 'Genotyping errors: Causes, consequences and solutions', *Nature Reviews Genetics*, pp. 847–859. doi: 10.1038/nrg1707.

Popescu, A. A., Huber, K. T. and Paradis, E. (2012) 'Ape 3.0: New tools for distance-based

phylogenetics and evolutionary analysis in R', *Bioinformatics*, 28(11), pp. 1536–1537. doi: 10.1093/bioinformatics/bts184.

Pouwels, K. B. *et al.* (2013) 'Cost-effectiveness of vaccination against meningococcal B among Dutch infants: Crucial impact of changes in incidence', *Human Vaccines and Immunotherapeutics*, 9(5), pp. 1129–1138. doi: 10.4161/hv.23888.

Power, P. M. and Moxon, E. R. (2006) 'Phase Variation and Adaptive Strategies of N. meningitidis: Insights into the Biology of a Commensal and Pathogen', in *Handbook of Meningococcal Disease: Infection Biology, Vaccination, Clinical Management*, pp. 99–118. doi: 10.1002/3527608508.ch6.

Pujol, C. *et al.* (1999) 'The meningococcal PilT protein is required for induction of intimate attachment to epithelial cells following pilus-mediated adhesion.', *Proceedings of the National Academy of Sciences of the United States of America*, 96(7), pp. 4017–4022. doi: 10.1073/pnas.96.7.4017.

Quackenbush, J. (2003) 'Open-source software accelerates bioinformatics', in *Genome Biology*. doi: 10.1186/gb-2003-4-9-336.

Quandt, E. M. *et al.* (2015) 'Fine-tuning citrate synthase flux potentiates and refines metabolic innovation in the lenski evolution experiment', *eLife*, 4(OCTOBER2015). doi: 10.7554/eLife.09696.

Quick, J. *et al.* (2016) 'Real-time, portable genome sequencing for Ebola surveillance', *Nature*, 530(7589), pp. 228–232. doi: 10.1038/nature16996.

Rambaut, A. (2009) 'FigTree, a graphical viewer of phylogenetic trees.', *Institute of Evolutionary Biology University of Edinburgh*. Available at: http://tree.bio.ed.ac.uk/software/figtree/.

Read, R. C. *et al.* (2014) 'Effect of a quadrivalent meningococcal ACWY glycoconjugate or a serogroup B meningococcal vaccine on meningococcal carriage: An observer-blind, phase 3 randomised clinical trial', *The Lancet*, 384(9960), pp. 2123–2131. doi: 10.1016/S0140-6736(14)60842-4.

Read, R. C. (2014) 'Neisseria meningitidis; clones, carriage, and disease', *Clinical Microbiology and Infection*, 20(5), pp. 391–395. doi: 10.1111/1469-0691.12647.

Read, R. C. *et al.* (2017) 'A phase III observer-blind randomized, controlled study to evaluate the immune response and the correlation with nasopharyngeal carriage after immunization of university students with a quadrivalent meningococcal ACWY glycoconjugate or serogroup B meningo', *Vaccine*, 35(3), pp. 427–434. doi: 10.1016/j.vaccine.2016.11.071.

Rhoads, A. and Au, K. F. (2015) 'PacBio Sequencing and Its Applications', *Genomics, Proteomics and Bioinformatics*, pp. 278–289. doi: 10.1016/j.gpb.2015.08.002.

Ribeiro, A. *et al.* (2015) 'An investigation of causes of false positive single nucleotide polymorphisms using simulated reads from a small eukaryote genome', *BMC Bioinformatics*, 16(1). doi: 10.1186/s12859-015-0801-z.

Rice, P., Longden, I. and Bleasby, A. (2000) 'EMBOSS: the European Molecular Biology Open Software Suite', *Trends Genet*, 16. doi: 10.1016/S0168-9525(00)02024-2.

Rice, P., Longden, L. and Bleasby, A. (2000) 'EMBOSS: The European Molecular Biology Open Software Suite', *Trends in Genetics*, 16(6), pp. 276–277. doi: 10.1016/S0168-9525(00)02024-2.

Richmond, P. C. *et al.* (2012) 'A bivalent Neisseria meningitidis recombinant lipidated factor H binding protein vaccine in young adults: Results of a randomised, controlled, dose-escalation phase 1 trial', *Vaccine*, 30(43), pp. 6163–6174. doi: 10.1016/j.vaccine.2012.07.065.

Roberts, S. B. *et al.* (2016) 'Correia Repeat Enclosed Elements and Non-Coding RNAs in the Neisseria Species', *Microorganisms*. Edited by D. W. Ussery. MDPI, 4(3), p. 31. doi: 10.3390/microorganisms4030031.

Ronaghi, M. *et al.* (1996) 'Real-time DNA sequencing using detection of pyrophosphate release', *Analytical Biochemistry*, 242(1), pp. 84–89. doi: 10.1006/abio.1996.0432.

Rosenstein, N. E. *et al.* (2001) 'Meningococcal disease.', *The New England journal of medicine*, 344(18), pp. 1378–88. doi: 10.1056/NEJM200105033441807.

Rothberg, J. M. *et al.* (2011) 'An integrated semiconductor device enabling non-optical genome sequencing', *Nature*, 475(7356), pp. 348–352. doi: 10.1038/nature10242.

Rotman, E. and Seifert, H. S. (2014) 'The Genetics of Neisseria Species.', *Annual review of genetics*, 48(September), pp. 405–431. doi: 10.1146/annurev-genet-120213-092007.

Rouli, L. *et al.* (2015) 'The bacterial pangenome as a new tool for analysing pathogenic bacteria', *New Microbes and New Infections*, 7, pp. 72–85. doi: 10.1016/j.nmni.2015.06.005.

Rouquette-Loughlin, C. E. *et al.* (2004) 'Modulation of the mtrCDE-encoded efflux pump gene complex of Neisseria meningitidis due to a Correia element insertion sequence', *Molecular Microbiology*, 54(3), pp. 731–741. doi: 10.1111/j.1365-2958.2004.04299.x.

Rutherford, K. *et al.* (2000) 'Artemis: sequence visualization and annotation.', *Bioinformatics (Oxford, England)*, 16(10), pp. 944–945. doi: 10.1093/bioinformatics/16.10.944.

Sahl, J. W. *et al.* (2016) 'The effects of signal erosion and core genome reduction on the identification of diagnostic markers', *mBio*, 7(5). doi: 10.1128/mBio.00846-16.

Salzberg, S. L. *et al.* (2012) 'GAGE: A critical evaluation of genome assemblies and assembly algorithms', *Genome Research*, 22(3), pp. 557–567. doi: 10.1101/gr.131383.111.

Sanger, F. *et al.* (1978) 'The nucleotide sequence of bacteriophage ϕX174', *Journal of Molecular Biology*, 125(2), pp. 225–246. doi: 10.1016/0022-2836(78)90346-7.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences*, 74(12), pp. 5463–5467. doi: 10.1073/pnas.74.12.5463.

Santolaya, M. E. *et al.* (2013) 'Persistence of antibodies in adolescents 18-24 months after immunization with one, two, or three doses of 4CMenB meningococcal serogroup B vaccine', *Human Vaccines and Immunotherapeutics*, 9(11), pp. 2304–2310. doi: 10.4161/hv.25505.

Sater, M. R. A. *et al.* (2015) 'DNA Methylation Assessed by SMRT Sequencing Is Linked to Mutations in Neisseria meningitidis Isolates.', *PloS one*, 10(12), p. e0144612. doi: 10.1371/journal.pone.0144612.

Saunders, N. J. *et al.* (2000) 'Repeat-associated phase variable genes in the complete genome sequence of Neisseria meningitidis strain MC58', *Molecular Microbiology*, pp. 207–215. doi: 10.1046/j.1365-2958.2000.02000.x.

Schliep, K. P. (2011) 'phangorn: phylogenetic analysis in R.', *Bioinformatics (Oxford, England)*, 27(4), pp. 592–3. doi: 10.1093/bioinformatics/btq706.

Schoen, C. *et al.* (2008) 'Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in Neisseria meningitidis.', *Proceedings of the National Academy of Sciences of the United States of America*, 105(9), pp. 3473–3478. doi: 10.1073/pnas.0800151105.

Schoen, C. *et al.* (2009) 'Genome flexibility in Neisseria meningitidis', *Vaccine*, 27(SUPPL. 2). doi: 10.1016/j.vaccine.2009.04.064.

Schoen, C. *et al.* (2014) 'Metabolism and virulence in Neisseria meningitidis', *Frontiers in Cellular and Infection Microbiology*, 4. doi: 10.3389/fcimb.2014.00114.

Seemann, T. (2014) 'Prokka: Rapid prokaryotic genome annotation', *Bioinformatics*, 30(14), pp. 2068–2069. doi: 10.1093/bioinformatics/btu153.

Segerman, B. (2012) 'The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories', *Frontiers in Cellular and Infection Microbiology*, 2. doi: 10.3389/fcimb.2012.00116.

Seib, K. L., Zhao, X. and Rappuoli, R. (2012) 'Developing vaccines in the era of genomics: A decade of reverse vaccinology', *Clinical Microbiology and Infection*, pp. 109–116. doi: 10.1111/j.1469-0691.2012.03939.x.

Shea, M. W. (2013) 'The Long Road to an Effective Vaccine for Meningococcus Group B (MenB)', *Annals of Medicine and Surgery*, pp. 53–56. doi: 10.1016/S2049-0801(13)70037-2.

Shirley, M. and Taha, M. K. (2018) 'MenB-FHbp Meningococcal Group B Vaccine (Trumenba®): A Review in Active Immunization in Individuals Aged ≥ 10 Years', *Drugs*, 78(2), pp. 257–268. doi: 10.1007/s40265-018-0869-7.

Siddique, A., Buisine, N. and Chalmers, R. (2011) 'The transposon-like correia elements encode numerous strong promoters and provide a potential new mechanism for phase variation in the meningococcus', *PLoS Genetics*, 7(1). doi: 10.1371/journal.pgen.1001277.

Siguier, P. *et al.* (2012) 'Exploring bacterial insertion sequences with ISfinder: Objectives, uses, and future developments', *Methods in Molecular Biology*, 859, pp. 91–103. doi: 10.1007/978-1-61779-603-6_5.

Sigurlásdóttir, S. *et al.* (2017) 'Host cell-derived lactate functions as an effector molecule in Neisseria meningitidis microcolony dispersal', *PLoS Pathogens*, 13(4). doi: 10.1371/journal.ppat.1006251.

Snape, M. D. *et al.* (2016) 'Persistence of bactericidal antibodies after infant serogroup B meningococcal immunization and booster dose response at 12, 18 or 24 months of age', in *Pediatric Infectious Disease Journal*, pp. e113–e123. doi: 10.1097/INF.0000000000001056.

Snyder, L. A. S., Shafer, W. M. and Saunders, N. J. (2003) 'Divergence and transcriptional analysis of the division cell wall (dcw) gene cluster in Neisseria spp.', *Molecular Microbiology*, 47(2), pp. 431–441. doi: 10.1046/j.1365-2958.2003.03204.x.

Snyder, L. A. and Saunders, N. J. (2006) 'The majority of genes in the pathogenic Neisseria species are present in non-pathogenic Neisseria lactamica , including those designated as "virulence genes"', *BMC Genomics*, 7(1), p. 128. doi: 10.1186/1471-2164-7-128.

Soeters, H. M. *et al.* (2017) 'Meningococcal carriage evaluation in response to a Serogroup B meningococcal disease outbreak and mass vaccination campaign at a College-Rhode Island, 2015-

2016', *Clinical Infectious Diseases*, 64(8), pp. 1115–1122. doi: 10.1093/cid/cix091.

Springer, M. (2006) 'Applied biosystems: Celebrating 25 years of advancing science', *American Laboratory*, pp. 4–8.

Staden, R. (1979) 'A strategy of DNA sequencing employing computer programs', *Nucleic Acids Research*, 6(7), pp. 2601–2610. doi: 10.1093/nar/6.7.2601.

Stearns, J. C. *et al.* (2015) 'Culture and molecular-based profiles show shifts in bacterial communities of the upper respiratory tract that occur with age', *ISME Journal*, 9(5), pp. 1246–1259. doi: 10.1038/ismej.2014.250.

de Steenhuijsen Piters, W. A. A., Sanders, E. A. M. and Bogaert, D. (2015) 'The role of the local microbial ecosystem in respiratory health and disease', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1675), p. 20140294. doi: 10.1098/rstb.2014.0294.

de Steenwinkel, J. E. M. *et al.* (2012) 'Drug susceptibility of mycobacterium tuberculosis Beijing genotype and association with MDR TB', *Emerging Infectious Diseases*, 18(4), pp. 660–663. doi: 10.3201/eid1804.110912.

Stein, L. D. (2010) 'The case for cloud computing in genome informatics', *Genome Biology*. doi: 10.1186/gb-2010-11-5-207.

Stephens, D. S. (2007) 'Conquering the meningococcus', *FEMS Microbiology Reviews*, pp. 3–14. doi: 10.1111/j.1574-6976.2006.00051.x.

Stephens, D. S. (2009) 'Biology and pathogenesis of the evolutionarily successful, obligate human bacterium Neisseria meningitidis', *Vaccine*, 27(SUPPL. 2). doi: 10.1016/j.vaccine.2009.04.070.

Stephens, D. S., Krebs, J. W. and McGee, Z. A. (1984) 'Loss of pili and decreased attachment to human cells by Neisseria meningitidis and Neisseria gonorrhoeae exposed to subinhibitory concentrations of antibiotics', *Infection and Immunity*, 46(2), pp. 507–513.

Strum, N. (2015) *DNA Mutation & Repair*, *nbs*.

Tan, A. *et al.* (2016) 'Distribution of the type III DNA methyltransferases modA, modB and modD among Neisseria meningitidis genotypes: implications for gene regulation and virulence', *Scientific Reports*. The Author(s), 6, p. 21015. Available at: http://dx.doi.org/10.1038/srep21015.

Tappero, J. W. *et al.* (1999) 'Immunogenicity of 2 serogroup B outer-membrane protein meningococcal vaccines: A randomized controlled trial in Chile', *Journal of the American Medical Association*, 281(16), pp. 1520–1527. doi: 10.1001/jama.281.16.1520.

References

Tatusov, R. L. (2001) 'The COG database: new developments in phylogenetic classification of proteins from complete genomes', *Nucleic Acids Research*, 29(1), pp. 22–28. doi: 10.1093/nar/29.1.22.

Tatusova, T. *et al.* (2016) 'NCBI prokaryotic genome annotation pipeline', *Nucleic Acids Research*, 44(14), pp. 6614–6624. doi: 10.1093/nar/gkw569.

Tauseef, I., Ali, Y. M. and Bayliss, C. D. (2013) 'Phase variation of PorA, a major outer membrane protein, mediates escape of bactericidal antibodies by Neisseria meningitidis', *Infection and Immunity*, 81(4), pp. 1374–1380. doi: 10.1128/IAI.01358-12.

Tettelin, H. *et al.* (2005) 'Genome analysis of multiple pathogenic isolates of {<}i{>}Streptococcus agalactiae{<}/i{>}: implications for the microbial "pan-genome".', *PNAS*, 102(39), pp. 13950–13955. doi: 10.1073/pnas.0506758102.

The Lancet Infectious Diseases (2014) 'The case for vaccinating against meningitis B', p. 359. doi: 10.1016/S1473-3099(14)70756-X.

Tobiason, D. M. and Seifert, H. S. (2010) 'Genomic content of neisseria species', *Journal of Bacteriology*, 192(8), pp. 2160–2168. doi: 10.1128/JB.01593-09.

van Tonder, A. J. *et al.* (2014) 'Defining the Estimated Core Genome of Bacterial Populations Using a Bayesian Decision Model', *PLoS Computational Biology*, 10(8). doi: 10.1371/journal.pcbi.1003788.

Treangen, T. J. *et al.* (2008) 'The impact of the neisserial DNA uptake sequences on genome evolution and stability.', *Genome biology*, 9(3), p. R60. doi: 10.1186/gb-2008-9-3-r60.

Trivedi, K., Tang, C. M. and Exley, R. M. (2011) 'Mechanisms of meningococcal colonisation', *Trends in Microbiology*, pp. 456–463. doi: 10.1016/j.tim.2011.06.006.

Tsai, C. M. and Civin, C. I. (1991) 'Eight lipooligosaccharides of Neisseria meningitidis react with a monoclonal antibody which binds lacto-N-neotetraose (Gal??1-4GlcNAc??1-3Gal??1-4Glc)', *Infection and Immunity*, 59(10), pp. 3604–3609.

Turnbaugh, P. J. *et al.* (2007) 'The Human Microbiome Project', *Nature*, pp. 804–810. doi: 10.1038/nature06244.

Turner, T. (2013) 'Plot protein: Visualization of mutations', *Journal of Clinical Bioinformatics*, 3(1). doi: 10.1186/2043-9113-3-14.

Vaughan, T. E. *et al.* (2006) 'Proteomic analysis of Neisseria lactamica and N eisseria meningitidis

outer membrane vesicle vaccine antigens', *Vaccine*, 24(25), pp. 5277–5293. doi: 10.1016/j.vaccine.2006.03.013.

Vélez Acevedo, R. N. *et al.* (2014) 'Identification of regulatory elements that control expression of the tbpBA operon in Neisseria gonorrhoeae', *Journal of Bacteriology*, 196(15), pp. 2762–2774. doi: 10.1128/JB.01693-14.

Vernikos, G. and Medini, D. (2014) 'Bexsero® chronicle', *Pathogens and Global Health*, 108(7), pp. 305–316. doi: 10.1179/2047773214Y.0000000162.

Vipond, C., Care, R. and Feavers, I. M. (2012) 'History of meningococcal vaccines and their serological correlates of protection', *Vaccine*, 30(SUPPL. 2). doi: 10.1016/j.vaccine.2011.12.060.

Voelkerding, K. V., Dames, S. A. and Durtschi, J. D. (2009) 'Next-generation sequencing:from basic research to diagnostics', *Clinical Chemistry*, pp. 641–658. doi: 10.1373/clinchem.2008.112789.

Vos, M. and Didelot, X. (2008) 'A comparison of homologous recombination rates in bacteria and archaea', *The ISME Journal*, 3, pp. 199–208. doi: 10.1038/ismej.2008.93.

Wahdan, M. H. *et al.* (1973) 'A controlled field trial of a serogroup A meningococcal polysaccharide vaccine', *Bulletin of the World Health Organization*, 48(6), pp. 667–673.

Walker, B. J. *et al.* (2014) 'Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement', *PLoS ONE*, 9(11). doi: 10.1371/journal.pone.0112963.

Wetzel, J., Kingsford, C. and Pop, M. (2011) 'Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies', *BMC Bioinformatics*, 12. doi: 10.1186/1471-2105-12-95.

Wheeler, D. A. *et al.* (2008) 'The complete genome of an individual by massively parallel DNA sequencing', *Nature*, 452(7189), pp. 872–876. doi: 10.1038/nature06884.

Wickham, H. (2009) *Ggplot2*, *Elegant Graphics for Data Analysis*. doi: 10.1007/978-0-387-98141-3.

Wilson, H. D. and Overman, T. L. (1976) 'Septicemia due to Neisseria lactamica', *Journal of Clinical Microbiology*, 4(3), pp. 214–215.

Wong, S. *et al.* (2007) 'New Zealand epidemic strain meningococcal B outer membrane vesicle vaccine in children aged 16-24 months', *Pediatric Infectious Disease Journal*, 26(4), pp. 345–350. doi: 10.1097/01.inf.0000258697.05341.2c.

Wood, D. E. and Salzberg, S. L. (2014) 'Kraken: Ultrafast metagenomic sequence classification

using exact alignments', *Genome Biology*, 15(3). doi: 10.1186/gb-2014-15-3-r46.

Van Der Woude, M. W. and Bäumler, A. J. (2004) 'Phase and antigenic variation in bacteria', *Clinical Microbiology Reviews*, pp. 581–611. doi: 10.1128/CMR.17.3.581-611.2004.

Yan, M. *et al.* (2013) 'Nasal microenvironments and interspecific interactions influence nasal microbiota complexity and S. aureus carriage', *Cell Host and Microbe*, 14(6), pp. 631–640. doi: 10.1016/j.chom.2013.11.005.

Yang, L. *et al.* (2011) 'Evolutionary dynamics of bacteria in a human host environment.', *Proceedings of the National Academy of Sciences of the United States of America*, 108(18), pp. 7481–6. doi: 10.1073/pnas.1018249108.

Yates, A. *et al.* (2016) 'Ensembl 2016', *Nucleic Acids Research*, 44(D1), pp. D710–D716. doi: 10.1093/nar/gkv1157.

Zhao, S. *et al.* (2017) 'Adaptive evolution within the gut microbiome of individual people', *bioRxiv*, p. 208009. doi: 10.1101/208009.

Zheng, W. *et al.* (2016) 'NeisseriaBase: a specialised Neisseria genomic resource and analysis platform.', *PeerJ*, 4, p. e1698. doi: 10.7717/peerj.1698.

Zhou, Y. *et al.* (2011) 'PHAST: A Fast Phage Search Tool', *Nucleic Acids Research*, 39(SUPPL. 2). doi: 10.1093/nar/gkr485.

Appendix A Supplementary Material

A.1 Chapter 2 supplementary data

A.1.1 Breseq script

Figure 0-1 Script for running breseq on Iridis.

run_trimmomatic.pbs breseq_job

```
#!/bin/bash

# Set requirements for the job.
# - these can be overridden on the qsub command line (this is for a 30 minute job)

#PBS -l walltime=6:00:00
#PBS -l nodes=1:ppn=16

# Change to directory from which job was submitted.
# (The actual name is held in the PBS environment variable $PBS_O_WORKDIR)

cd $PBS_O_WORKDIR

# Load the R module so that the R command can be found.

module load breseq

# Run Breseq non interactively.
# The output from this command will be written to "r_script.Rout".

breseq -r 115_breseq_control.gbk filename-paired_1.fastq filename-paired_2.fastq -j 16
exit
fi
```

The reference in this example is named (**115_breseq_control.gbk**" this means it is a genbank format file used by Breseq as the reference assembly. Paired end reads **(.fastq files**) are accepted. The **[-j]** command asks the pipeline to use 16 processors in multithreaded steps. This was taken to reduce program running time.

A.1.2 Trimmomatic script

```
n#!/bin/bash

# Job script to run pbs job in job array

# to submit; qsub -t 1-20 run_array.sh

# or

# qsub -t 2,5,6,20-35 run_array.sh

#=============================================

# set default resource requirements for job

# - these can be overridden on the qsub command line

#PBS -l nodes=1:ppn=1

#PBS -l walltime=01:00:00

#Change to directory from which job was submitted

cd $PBS_O_WORKDIR

# load software module

module load jdk/1.7.0

adapter_file="nextera.fa"

fastq_file="filefastq.gz"

R2fastq_file="file.fastq.gz"

echo "processing files $fastq_file and  $R2fastq_file "

if [ -f $fastq_file ]

then

# if file exist - run script

  java -jar /local/software/trimmomatic/0.32/trimmomatic-0.32.jar PE $fastq_file  $R2fastq_file
GSK-$PBS_ARRAYID-paired_1.fastq.gz GSK-$PBS_ARRAYID-unpaired_1.fastq.gz GSK-
$PBS_ARRAYID-paired_2.fastq.gz GSK-$PBS_ARRAYID-unpaired_2.fastq.gz
SLIDINGWINDOW:10:20 MINLEN:50 ILLUMINACLIP:nextera.fa:1:40:15

# if file no found - print error message and exit

echo 'File '$fastq_file' not found, quit'

exit

fi
```

A.2 Chapter 3 supplementary data

Genes or gene variants identified as unique to the previously sequenced PHE *N.lac* Y92-1009 assembly by kraken. Fifteen of these genes were found to contain quality control issues.

This section is referenced in section 3.3.2 of the main body of the thesis.

Figure 0-2 Genes or gene variants **(part 1/2, 1-32 genes)** identified as unique to the PHE genome assembly by Kraken.

Gene	Annotation	QC
group_1349	30S ribosomal protein S1 [Neisseria lactamica]	
group_3944	ADP-heptose:LPS heptosyltransferase II [Neisseria lactamica 020-06]	
neo	Aminoglycoside 3'-phosphotransferase	
group_601	apolipoprotein acyltransferase [Neisseria lactamica]	
group_864	ATP-dependent DNA helicase RecQ [Neisseria lactamica]	
bla	Beta-lactamase TEM precursor	Investigate
group_1396	DNA cytosine methyltransferase [Neisseria lactamica]	Investigate
group_336	DNA polymerase III subunit epsilon [Neisseria lactamica]	
group_3935	DnaA regulatory inactivator Hda [Neisseria lactamica]	
group_1143	DNA-directed RNA polymerase subunit beta [Neisseria lactamica]	
group_848	enterobactin ABC transporter permease [Neisseria lactamica]	Investigate
group_1307	glutamate dehydrogenase [Neisseria lactamica]	
group_289	glycosyl transferase 2 family protein [Neisseria lactamica]	
group_3958	homoserine dehydrogenase [Neisseria lactamica]	Investigate
group_35	hypothetical membrane protein [Neisseria lactamica 020-06]	Investigate
group_1296	hypothetical protein	Investigate
group_1546	hypothetical protein	Investigate
group_3936	hypothetical protein	Investigate
group_3950	hypothetical protein	Investigate
group_3951	hypothetical protein	Investigate
group_3955	hypothetical protein	Investigate
group_1564	hypothetical protein	
group_163	hypothetical protein	
group_2704	hypothetical protein	
group_3953	hypothetical protein	
group_3957	hypothetical protein	
group_638	hypothetical protein [Neisseria lactamica 020-06]	
group_1138	hypothetical protein [Neisseria lactamica]	Investigate
group_1418	hypothetical protein [Neisseria lactamica]	Investigate
group_2735	hypothetical protein [Neisseria lactamica]	Investigate
group_3942	hypothetical protein [Neisseria lactamica]	Investigate

Cells highlighted in red have flagged up quality control errors. The **gene** column denotes what protein clustering group the given gene (describes in the **description** section) has been categorised in. **QC** is an acronym for quality control.

Figure 0-3 Genes or gene variants **(part 2/2, 33-65 genes)** identified as unique to PHE genome assembly by Kraken

Gene	Annotation	QC
group_2096	hypothetical protein [Neisseria lactamica]	
group_25	hypothetical protein [Neisseria lactamica]	
group_3943	hypothetical protein [Neisseria lactamica]	
group_789	hypothetical protein [Neisseria lactamica]	
group_685	insertase [Neisseria lactamica]	
group_2329	integral membrane protein NnrS [Neisseria lactamica]	
group_324	ligand-gated channel [Neisseria lactamica]	Investigate
group_588	L-serine dehydratase [Neisseria lactamica 020-06]	Investigate
group_3949	membrane protein [Neisseria lactamica]	Investigate
group_1028	membrane protein [Neisseria lactamica]	
group_3948	membrane protein [Neisseria lactamica]	
group_890	membrane protein [Neisseria lactamica]	
group_3945	methylated-DNA--protein-cysteine methyltransferase [Neisseria lactamica]	
group_3956	MULTISPECIES: bacterioferritin [Neisseria]	Investigate
group_3937	MULTISPECIES: yqey-like family protein [Neisseria]	
group_201	oligopeptidase A [Neisseria lactamica]	Investigate
group_323	oxidoreductase [Neisseria lactamica]	
group_3939	preprotein translocase subunit SecE [Neisseria lactamica]	
group_1514	putative ATP-dependent protease [Neisseria lactamica 020-06]	Investigate
group_489	putative inner membrane protease [Neisseria lactamica 020-06]	
group_1060	putative integrase/recombinase [Neisseria lactamica 020-06]	Investigate
group_964	putative polyamine permease substrate-binding protein [Neisseria lactamica 020-06]	Investigate
group_309	putative restriction modification system DNA specificity domain [Neisseria lactamica 020-06]	Investigate
group_2099	putative transposase [Neisseria lactamica 020-06]	
group_3954	putative TspB protein [Neisseria lactamica 020-06]	
group_137	restriction endonuclease subunit M [Neisseria lactamica]	Investigate
group_3941	shikimate dehydrogenase [Neisseria lactamica]	
group_768	short-chain dehydrogenase [Neisseria lactamica]	
group_124	TonB-dependent receptor [Neisseria lactamica]	
group_3938	transcription antitermination protein [Neisseria lactamica 020-06]	
group_344	type III restriction/modification system enzyme [Neisseria lactamica]	

Cells highlighted in red have flagged up quality control errors. The **gene** column denotes what protein clustering group the given gene (describes in the **description** section) has been categorised in. **QC** is an acronym for quality control.

Appendix A

Figure 0-4 Table of genes or gene variants unique to PacBio RSII/Illumina Hiseq 2000 hybrid assembly (1-50)

Gene	Annotation
group_604	23S rRNA methyltransferase [Neisseria lactamica]
group_606	3-oxoacyl-(acyl-carrier-protein) synthase ii (ec 2.3.1.41) [Neisseria lactamica 020-06]
group_240	4-hydroxybenzoyl-CoA thioesterase [Neisseria lactamica]
group_2279	50S ribosomal protein L30 [Neisseria lactamica]
group_1898	5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase [Neisseria lactamica]
group_1275	alpha 1,2 N-acetylglucosamine transferase [Neisseria lactamica]
group_2333	alpha-glucan phosphorylase [Neisseria lactamica]
group_1742	aminopeptidase N [Neisseria lactamica]
group_196	AMP-binding protein [Neisseria lactamica]
group_2332	aspartyl-tRNA synthetase [Neisseria lactamica 020-06]
group_1491	beta-1,4-glucosyltransferase [Neisseria lactamica 020-06]
group_2348	biopolymer transport protein [Neisseria lactamica 020-06]
group_2347	biopolymer transporter ExbB [Neisseria lactamica]
group_1272	cell envelope protein TonB [Neisseria lactamica]
group_739	chloride channel protein [Neisseria lactamica]
group_1207	competence protein ComE [Neisseria lactamica]
group_2342	crossover junction endodeoxyribonuclease [Neisseria lactamica 020-06]
group_1899	cysteine synthase A [Neisseria lactamica]
group_2349	cytochrome c oxidase subunit [Neisseria lactamica 020-06]
group_1058	diaminopimelate epimerase [Neisseria lactamica]
group_1897	DNA polymerase III subunit delta' [Neisseria lactamica]
group_2408	DnaA regulatory inactivator Hda [Neisseria lactamica]
group_1142	DNA-directed RNA polymerase subunit beta [Neisseria lactamica]
group_2453	elongation factor 4 [Neisseria lactamica]
group_260	exodeoxyribonuclease V subunit gamma [Neisseria lactamica]
group_1270	FAD-dependent oxidoreductase [Neisseria lactamica]
group_2354	glutamate dehydrogenase [Neisseria lactamica]
group_2346	glutaredoxin 2 [Neisseria lactamica 020-06]
group_607	glycosyl transferase [Neisseria lactamica]
group_1276	glycosyl transferase family 2 [Neisseria lactamica]
group_2353	GntR family transcriptional regulator [Neisseria lactamica]
group_1743	GTP pyrophosphokinase [Neisseria lactamica]
group_34	hypothetical membrane protein [Neisseria lactamica 020-06]
group_1123	hypothetical protein
group_1297	hypothetical protein
group_1548	hypothetical protein
group_165	hypothetical protein
group_1976	hypothetical protein
group_2011	hypothetical protein
group_2037	hypothetical protein
group_2079	hypothetical protein
group_2084	hypothetical protein
group_2090	hypothetical protein
group_2091	hypothetical protein
group_2648	hypothetical protein
group_2660	hypothetical protein
group_2665	hypothetical protein
group_2705	hypothetical protein
group_2719	hypothetical protein
group_2731	hypothetical protein
group_7	hypothetical protein
group_885	hypothetical protein
group_951	hypothetical protein
group_1057	hypothetical protein [Neisseria lactamica 020-06]

Appendix A

Figure 0-5 List of genes or gene variants unique to PacBio RSII/Illumina Hiseq 2000 hybrid assembly (51-100)

Gene	Annotation
group_1468	hypothetical protein [Neisseria lactamica 020-06]
group_2732	hypothetical protein [Neisseria lactamica 020-06]
group_637	hypothetical protein [Neisseria lactamica 020-06]
group_1309	hypothetical protein [Neisseria lactamica]
group_1417	hypothetical protein [Neisseria lactamica]
group_1800	hypothetical protein [Neisseria lactamica]
group_19	hypothetical protein [Neisseria lactamica]
group_1905	hypothetical protein [Neisseria lactamica]
group_2287	hypothetical protein [Neisseria lactamica]
group_2288	hypothetical protein [Neisseria lactamica]
group_23	hypothetical protein [Neisseria lactamica]
group_2344	hypothetical protein [Neisseria lactamica]
group_2439	hypothetical protein [Neisseria lactamica]
group_2446	hypothetical protein [Neisseria lactamica]
group_261	hypothetical protein [Neisseria lactamica]
group_2729	hypothetical protein [Neisseria lactamica]
group_376	hypothetical protein [Neisseria lactamica]
group_953	hypothetical protein [Neisseria lactamica]
group_979	integrase [Neisseria lactamica]
group_982	integrase [Neisseria lactamica]
group_1741	lipid A biosynthesis lauroyl acyltransferase [Neisseria lactamica]
group_2352	L-lactate permease [Neisseria lactamica]
group_589	L-serine dehydratase [Neisseria lactamica 020-06]
group_2340	lysyl-tRNA synthetase [Neisseria lactamica 020-06]
group_47	mafB silent cassette [Neisseria lactamica 020-06]
group_1273	membrane protein [Neisseria lactamica]
group_1402	membrane protein [Neisseria lactamica]
group_224	membrane protein [Neisseria lactamica]
group_2339	membrane protein [Neisseria lactamica]
group_297	membrane protein [Neisseria lactamica]
group_377	membrane protein [Neisseria lactamica]
group_603	membrane protein [Neisseria lactamica]
group_950	membrane protein [Neisseria lactamica]
group_1746	multidrug transporter [Neisseria lactamica]
group_2331	MULTISPECIES: 30S ribosomal protein S20 [Proteobacteria]
group_2343	MULTISPECIES: 3'-5' exonuclease [Neisseria]
group_2457	MULTISPECIES: bacterioferritin [Neisseria]
group_2341	MULTISPECIES: Fis family transcriptional regulator [Neisseria]
group_366	MULTISPECIES: glycosyl transferase group 1 family protein [Neisseria]
group_2334	MULTISPECIES: GTP-binding protein [Neisseria]
group_1274	MULTISPECIES: hypothetical protein [Neisseria]
group_2336	MULTISPECIES: hypothetical protein [Neisseria]
group_2730	MULTISPECIES: hypothetical protein [Neisseria]
group_428	MULTISPECIES: hypothetical protein [Neisseria]
group_2335	MULTISPECIES: peptidase M23 [Neisseria]

Appendix A

Figure 0-6 List of genes or gene variants unique to PacBio RSII/Illumina Hiseq 2000 hybrid assembly (101-154)

Gene	Annotation
group_2350	MULTISPECIES: peptidase S41 [Neisseria]
group_1268	MULTISPECIES: polyamine ABC transporter substrate-binding protein [Ne
group_2355	MULTISPECIES: thymidylate synthase [Neisseria]
group_2449	MULTISPECIES: uracil phosphoribosyltransferase [Neisseria]
group_378	outer membrane ferripyoverdine receptor [Neisseria lactamica]
group_259	peptidase [Neisseria lactamica]
group_740	peptidase M24 [Neisseria lactamica]
group_2728	peptidase M50 [Neisseria lactamica]
group_986	phosphoesterase [Neisseria lactamica]
group_738	phospholipase A(1) [Neisseria lactamica]
group_375	phosphoserine phosphatase SerB [Neisseria lactamica]
group_2469	pilus assembly protein [Neisseria lactamica]
group_2468	pilus assembly protein PilO [Neisseria lactamica]
group_2467	pilus assembly protein PilP [Neisseria lactamica]
group_2273	preprotein translocase subunit SecE [Neisseria lactamica]
group_2093	putative adhesin [Neisseria lactamica 020-06]
group_506	putative ApbE family lipoprotein [Neisseria lactamica 020-06]
group_1512	putative ATP-dependent protease [Neisseria lactamica 020-06]
group_1269	putative bifunctional purine biosynthesis protein [Neisseria lactamica 02(
group_1744	putative cytochrome C oxidase subunit [Neisseria lactamica 020-06]
group_2351	putative cytochrome C oxidase, subunit III (ec 1.9.3.1) [Neisseria lactami
group_1745	putative drug efflux protein [Neisseria lactamica 020-06]
group_2450	putative glutaredoxin [Neisseria lactamica 020-06]
group_2447	putative HemY protein [Neisseria lactamica 020-06]
group_460	putative integral membrane protein [Neisseria lactamica 020-06]
group_889	putative large surface adhesin [Neisseria lactamica 020-06]
group_2338	putative NAD(P) transhydrogenase alpha subunit [Neisseria lactamica 02(
group_2337	putative NAD(P) transhydrogenase beta subunit [Neisseria lactamica 020
group_2466	putative pilus assembly protein [Neisseria lactamica 020-06]
group_2451	putative pilus biogenesis protein [Neisseria lactamica 020-06]
group_2458	putative regulator [Neisseria lactamica 020-06]
group_308	putative restriction modification system DNA specificity domain [Neisser
group_539	putative secreted protein [Neisseria lactamica 020-06]
group_2454	putative signal peptidase I [Neisseria lactamica 020-06]
group_1896	putative TatD-related deoxyribonuclease [Neisseria lactamica 020-06]
group_505	putative Ton-B dependent receptor [Neisseria lactamica 020-06]
group_238	putative tRNA methyltransferase [Neisseria lactamica 020-06]
group_1401	putative uroporphyrin-III c-methyltransferase HemX [Neisseria lactamica
group_2070	putative uroporphyrinogen-III synthase HemD [Neisseria lactamica 020-(
group_743	pyridine nucleotide-disulfide oxidoreductase [Neisseria lactamica]
group_139	restriction endonuclease subunit M [Neisseria lactamica]
group_1740	RNA helicase [Neisseria lactamica]
group_1492	Suppressor of fused protein (SUFU)
group_682	Suppressor of fused protein (SUFU)
group_835	thioredoxin [Neisseria lactamica]
group_125	TonB-dependent receptor [Neisseria lactamica]
group_2274	transcription antitermination protein [Neisseria lactamica 020-06]
group_328	transcriptional regulator [Neisseria lactamica]
group_429	transporter [Neisseria lactamica]
group_186	transposase [Neisseria lactamica]
group_65	transposase [Neisseria lactamica]
group_952	tRNA-dihydrouridine synthase B [Neisseria lactamica]
group_2452	twitching motility protein PilT [Neisseria lactamica]
group_1087	type IV pilus biogenesis and competence protein PilQ [Neisseria lactamic

A.3 Chapter 5 supplemental data

Figure 0-7 List of meningococcal genomes with PubMLST ID's found to be carried by volunteers artificially inoculated with *N. lactamica* Y92-1009

id	Week	Volunteer	strain_designation	ST (MLST)	clonal_complex (MLST)	Contigs
26490	16	64	B: P1.21,16: F1-15: ST-839 (cc41/44)	839	ST-41/44 complex	218
26353	26	64	B: P1.21,16: F1-15: ST-839 (cc41/44)	839	ST-41/44 complex	305
26298	16	77	B: P1.17-1,23: F1-5: ST-1423 (cc41/44)	1423	ST-41/44 complex	196
26268	26	77	B: P1.17-1,23: F1-5: ST-1423 (cc41/44)	1423	ST-41/44 complex	232
26398	0	104	ND: P1.5-1,10-10: F4-1: ST-1655 (cc23)	1655	ST-23 complex	240
26309	26	128	ND: P1.18-1,1: F1-6: ST-865 (cc865)	865	ST-865 complex	144
26360	0	221	ND: P1.5-2,10-2: F1-7: ST-466 (cc60)	466	ST-60 complex	255
26464	16	221	ND: P1.5-2,10-2: F1-7: ST-466 (cc60)	466	ST-60 complex	250
26494	0	255	ND: P1.22,14: F5-5: ST-ND (-)			315
26509	16	255	ND: P1.22,14: F5-5: ST-213 (cc213)	213	ST-213 complex	256
26377	26	255	ND: P1.22,14: F5-5: ST-213 (cc213)	213	ST-213 complex	362
26349	26	279	ND: P1.5,2: F3-1: ST-60 (cc60)	60	ST-60 complex	304
26386	8	309	ND: P1.5-1,10-1: F4-1: ST-1655 (cc23)	1655	ST-23 complex	230

Appendix A

Figure 0-8 Artificially inoculated *N. lactamica* Y92-1009 genomes and PubMLST ID's part 1/2

BIGSDB isolate id	volunteer	time point	Contigs
36913	38	week 08	65
36953	38	week 16	86
36884	63	week 02	90
36926	63	week 08	74
36976	63	week 26	95
36883	64	week 02	78
36921	64	week 08	74
36955	64	week 16	78
36982	64	week 26	86
36898	72	week 08	98
36945	72	week 16	76
36965	72	week 26	79
36923	77	week 08	86
36963	77	week 16	80
36876	82	week 02	70
36916	82	week 08	85
36971	82	week 26	73
36902	88	week 08	75
36944	88	week 16	84
36935	93	week 08	80
36987	93	week 26	72
36859	104	week 02	69
36942	104	week 16	74
36885	113	week 02	78
36924	113	week 08	75
36964	113	week 16	72
36995	113	week 26	81
36972	118	week 26	67
36899	158	week 02	69
36928	158	week 08	82
36986	158	week 26	67
37016	160	week 02	85
36912	160	week 04	76
36985	160	week 26	74
36877	166	week 02	70
36994	166	week 26	78
36895	170	week 02	73
36919	170	week 04	71
36929	170	week 08	78
36962	170	week 16	67
36997	170	week 26	73
36914	172	week 08	74
36952	172	week 16	74
36967	172	week 26	72
36878	181	week 02	70
36915	181	week 08	74
36984	181	week 26	79

Appendix A

Figure 0-9 Artificially inoculated *N. lactamica* Y92-1009 genomes and PubMLST ID's part 2/2

BIGSDB isolate id	volunteer	time point	Contigs
36870	190	week 02	75
36907	190	week 08	72
36950	190	week 16	65
36863	201	week 02	154
36905	201	week 08	85
36949	201	week 16	75
36970	201	week 26	70
36867	207	week 02	77
36890	207	week 08	84
37017	213	week 02	74
36882	221	week 02	67
36954	221	week 16	75
36981	221	week 26	87
36910	222	week 02	89
36936	222	week 08	67
36909	227	week 02	74
36920	227	week 04	71
36930	227	week 08	75
36958	227	week 16	55
36993	227	week 26	84
36903	253	week 02	75
36943	253	week 08	84
36960	253	week 16	70
36906	255	week 02	68
36918	255	week 04	81
36887	260	week 02	81
36925	260	week 08	75
37018	260	week 16	83
36991	260	week 26	100
36881	264	week 02	69
36922	264	week 08	79
36951	264	week 16	75
36873	279	week 02	79
36911	279	week 08	90
36879	284	week 02	86
36940	284	week 08	66
36865	290	week 02	65
36904	290	week 08	78
36948	290	week 16	72
36966	290	week 26	66
36931	309	week 08	63
36932	312	week 08	71
36961	312	week 16	80
36896	313	week 02	68
36959	313	week 16	82
36998	313	week 26	73
36893	315	week 02	75
36937	315	week 16	72

Appendix A

Legend for Figure 0-8 and Figure 0-9

All isolates possessed the following allele numbers, MLST scores & strain designations; **abcZ**: 80, **adk:** 45, **aroE**: 98, **fumC**: 100, **gdh**: 94, **pdhC**: 158, **pgm:** 56, **Sequence type**: 3493, **clonal complex:** ST-613, **species**: *Neisseria lactamica*

Figure 0-10 Putative homopolymeric tracts identified within coding sequences of *N.lac* y92-1009 part 1/2

start	end	polymeric_base	gene_position	gene/proteome_number	Mutated during study	gene_product
27483	27490	T	< coding (280-287/1326 nt)	ftsW	no	cell division protein FtsW
58187	58194	A	> coding (809-816/1113 nt)	protein_00057	no	quinolinate synthase
130011	130018	A	< coding (133-140/1173 nt)	protein_00134	no	hemolysin D
135275	135283	C	< coding (747-755/1434 nt)	protein_00139	no	membrane protein
159428	159435	T	> coding (82-89/1347 nt)	protein_00166	no	tryptophan permease
161330	161341	C	> coding (485-496/573 nt)	lgtG	**yes**	L glycosyltransferase G
166958	166965	T	< coding (12-19/1902 nt)	Thic	no	phosphomethylpyrimidine synthase ThiC
218308	218315	A	> coding (307-314/639 nt)	protein_00229	no	hypothetical protein
229996	230003	T	> coding (135-142/345 nt)	protein_00243	no	hypothetical protein
230024	230031	T	> coding (163-170/345 nt)	protein_00243	no	hypothetical protein
230142	230149	T	> coding (281-288/345 nt)	protein_00243	no	hypothetical protein
237589	237596	T	> coding (158-165/411 nt)	protein_00254	no	hypothetical protein
238793	238800	A	> coding (4-11/318 nt)	protein_00257	no	hypothetical protein
241051	241058	T	> coding (192-199/360 nt)	protein_00260	no	hypothetical protein
256701	256710	G	> coding (736-745/798 nt)	pglA	**yes**	[protein-PII] uridylyltransferase
261695	261702	C	> coding (171-178/261 nt)	protein_00277	no	hypothetical protein
261703	261711	A	> coding (179-187/261 nt)	protein_00277	no	hypothetical protein
261715	261722	A	> coding (191-198/261 nt)	protein_00277	no	hypothetical protein
302369	302376	A	< coding (458-465/2586 nt)	protein_00315	no	hypothetical protein
303061	303068	T	< coding (66-73/195 nt)	protein_00316	no	multidrug transporter
305046	305054	T	< coding (352-360/1320 nt)	protein_00318	no	hypothetical protein
386514	386523	G	< coding (9-18/123 nt)	protein_00411	no	hypothetical protein
395561	395568	A	> coding (787-794/867 nt)	protein_00419	no	glucose-1-phosphate thymidylyltransferase
410064	410071	A	> coding (56-63/1227 nt)	pilT	no	twitching motility protein PilT
426357	426364	A	< coding (404-411/840 nt)	protein_00448	no	MULTISPECIES: iron permease [Neisseria]
441348	441355	T	> coding (44-51/984 nt)	protein_00461	no	membrane protein
442259	442267	A	> coding (955-963/984 nt)	protein_00461	no	membrane protein
443850	443857	A	> coding (1205-1212/1254 nt)	protein_00463	no	UDP-N-acetylglucosamine 1-carboxyvinyltransferase
482698	482705	C	> coding (890-897/1482 nt)	protein_00499	no	membrane protein
503105	503112	A	> coding (1705-1712/1716 nt)	protein_00521	no	hypothetical protein
517585	517592	T	> coding (370-377/426 nt)	protein_00538	no	hypothetical protein [Neisseria]
539455	539462	T	< coding (44-51/615 nt)	protein_00552	no	3-hydroxyisobutyrate dehydrogenase [Neisseria]
539834	539841	T	< coding (921-928/1260 nt)	protein_00553	no	DNA modification methylase
540484	540491	T	< coding (271-278/1260 nt)	protein_00553	no	DNA modification methylase
546832	546839	A	> coding (1480-1487/2589 nt)	protein_00558	no	DNA cytosine methyltransferase
547434	547441	T	> coding (2082-2089/2589 nt)	protein_00558	no	DNA cytosine methyltransferase
547915	547922	T	> coding (2563-2570/2589 nt)	protein_00558	no	DNA cytosine methyltransferase
592273	592280	T	< coding (1138-1145/2157 nt)	fetA	**yes**	iron-regulated outer membrane protein FetA
617853	617862	C	< coding (58-67/1026 nt)	hpuA	**yes**	hemoglobin-haptoglobin-utilization protein
659607	659619	G	> coding (158-170/936 nt)	lgtC	**yes**	L glycosyltransferase C
731323	731330	A	< coding (562-569/2991 nt)	pilC	no	pilus assembly protein PilC
731323	731330	A	< coding (562-569/2991 nt)	pilC	no	pilus assembly protein PilC
742740	742750	G	> coding (403-413/426 nt)	pglIL	**yes**	lipopolysaccharide modification acyltransferase
754832	754839	T	> coding (211-218/495 nt)	protein_00751	no	putative lipoprotein signal peptidase

Figure 0-11 Putative homopolymeric tracts identified within coding sequences of N.lac y92-1009 part 2/2

start	end	polymeric_base	gene_position	gene/proteome_number	Mutated during study	gene_product
776725	776737	G	< coding (617-629/1170 nt)	pglH	**yes**	lipopolysaccharide modification acyltransferase
807424	807434	C	> coding (17-27/330 nt)	protein_00797	no	hypothetical protein
834691	834698	T	< coding (55-62/1587 nt)	protein_00826	no	lactate permease
865096	865103	T	< coding (4-11/819 nt)	protein_00851	no	serine acetyltransferase
873636	873643	T	< coding (181-188/1344 nt)	protein_00861	no	Na(+)-translocating NADH-ubiquinone reductase
893426	893433	T	< coding (166-173/1209 nt)	protein_00878	no	putative adhesin
921120	921127	T	< coding (184-191/1305 nt)	protein_00907	no	ammonia channel protein
969505	969512	A	> coding (5-12/183 nt)	protein_00956	no	MULTISPECIES: hypothetical protein [Neisseria]
976062	976069	A	> coding (410-417/483 nt)	protein_00965	no	MULTISPECIES: hypothetical protein [Neisseria]
985496	985503	G	> coding (81-88/2835 nt)	protein_00974	no	ligand-gated channel
1014474	1014481	G	> coding (354-361/594 nt)	protein_01002	no	hypothetical protein
1063647	1063654	T	> coding (220-227/606 nt)	protein_01061	no	membrane protein
1096225	1096232	T	< coding (131-138/579 nt)	protein_01094	no	putative peptidyl-tRNA hydrolase
1115279	1115286	A	> coding (147-154/648 nt)	protein_01114	no	adenylate kinase
1124229	1124237	G	> coding (643-651/1245 nt)	hsdS	**yes**	NgoAV type I restriction modification system specificity subunit
1143902	1143909	T	< coding (422-429/630 nt)	protein_01140	no	hypothetical protein
1183268	1183275	T	< coding (469-476/2226 nt)	protein_01183	no	isocitrate dehydrogenase
1224806	1224813	T	< coding (740-747/1134 nt)	protein_01224	no	multidrug DMT transporter
1224833	1224840	T	< coding (713-720/1134 nt)	protein_01224	no	multidrug DMT transporter
1240306	1240313	A	> coding (83-90/519 nt)	protein_01239	no	hypothetical protein
1240722	1240729	A	> coding (499-506/519 nt)	protein_01239	no	hypothetical protein
1251454	1251461	T	< coding (1403-1410/1413 nt)	protein_01249	no	Adenosine monophosphate-protein transferase SoFic
1263830	1263837	A	> coding (6994-7001/7104 nt)	protein_01255	no	hemagglutinin/hemolysin-related protein
1264260	1264267	T	> coding (324-331/360 nt)	protein_01256	no	MULTISPECIES: hypothetical protein [Neisseria]
1275890	1275897	A	> coding (78-85/498 nt)	protein_01271	no	hypothetical protein, partial
1276089	1276096	T	> coding (277-284/498 nt)	protein_01271	no	hypothetical protein, partial
1289886	1289893	A	> coding (4-11/1203 nt)	protein_01285	no	hypothetical protein
1293680	1293687	T	< coding (338-345/1056 nt)	protein_01288	no	sulfate ABC transporter substrate-binding protein
1293901	1293908	T	< coding (117-124/1056 nt)	protein_01288	no	sulfate ABC transporter substrate-binding protein
1294757	1294764	T	< coding (1033-1040/1188 nt)	protein_01290	no	phosphoribosylaminoimidazole carboxylase ATPase subunit
1305117	1305124	T	< coding (145-152/603 nt)	protein_01299	no	glycerol-3-phosphate acyltransferase
1336633	1336642	A	> coding (4-13/1254 nt)	protein_01331	no	tRNA(Ile)-lysidine synthetase
1347183	1347190	T	< coding (528-535/1770 nt)	protein_01340	no	sulfite reductase subunit beta
1385878	1385885	A	> coding (24-31/330 nt)	protein_01380	no	hypothetical protein
1425439	1425447	T	< coding (201-209/357 nt)	protein_01437	no	hypothetical protein
1447539	1447546	A	> coding (253-260/300 nt)	protein_01456	no	hypothetical protein
1483207	1483216	G	> coding (306-315/1014 nt)	c-hp/NEIS1156	**yes**	glycosyl transferase family 2
1674192	1674201	C	> coding (13-22/2073 nt)	protein_01653	**yes**	TonB dependent, outer membrane protein similar to *fetA*
1698893	1698900	A	> coding (591-598/1002 nt)	protein_01663	no	tRNA-dihydrouridine synthase C
1732506	1732513	A	> coding (304-311/2013 nt)	protein_01687	no	ATP-dependent DNA helicase Rep
1734191	1734198	A	> coding (1989-1996/2013 nt)	protein_01687	no	ATP-dependent DNA helicase Rep
1750082	1750092	C	> coding (36-46/1542 nt)	protein_01701	**yes**	Hp; Phosphoglycerol transferase (blastP), sulfatase (pfam)
1779558	1779567	G	> coding (845-854/987 nt)	protein_01723	no	putative large surface adhesin
1779780	1779787	A	> coding (75-82/219 nt)	protein_01724	no	hypothetical protein
1791058	1791065	A	> coding (46-53/282 nt)	protein_01739	no	hypothetical protein
1875587	1875594	T	< coding (26-33/384 nt)	protein_01817	no	diacylglycerol kinase
1917069	1917076	T	< coding (251-258/1170 nt)	protein_01861	no	DNA cytosine methyltransferase [Neisseria]
1926870	1926877	T	< coding (141-148/1170 nt)	protein_01868	no	O-succinylhomoserine sulfhydrolase
1960665	1960672	A	< coding (172-179/1635 nt)	protein_01891	no	putative sulfatase
2062821	2062829	C	> coding (8-16/2163 nt)	protein_01983	no	outer membrane ferripyoverdine receptor
2093971	2093978	T	< coding (232-239/390 nt)	protein_02009	no	hypothetical protein
2102455	2102463	A	< coding (281-289/291 nt)	protein_02017	no	hypothetical protein
2102601	2102608	A	< coding (136-143/291 nt)	protein_02017	no	hypothetical protein
2114892	2114900	T	< coding (245-253/357 nt)	protein_02027	no	hypothetical protein

Caption for Figure 0-10 and Figure 0-11

This two-part table lists the 8-14 bp homopolymeric tracts detected in the *N. lactamica Y92-1009* genome. The **polymeric base** indicates the type of homopolymeric tract. The **gene position** column gives the location of the tract within the coding sequence of a given gene which is described in the **gene product** column.