UNIVERSITY OF
LEICESTER

Bioinformatics analyses of genetic variation in genomes of *Neisseria meningitidis* (the meningococcus)

Thesis submitted for the degree of

Doctor of Philosophy

At the University of Leicester

By

Mohammad Abdul Rahmman Mohammad Al-Maeni

Department of Genetics

University of Leicester

Leicester, UK

September 2017

Bioinformatics analyses of genetic variation in genomes of *Neisseria meningitidis* (the meningococcus)

Mohammad Abdul Rahmman Mohammad Al-Maeni

Abstract

Genetic variation is one of the key concepts underlying persistence of *Neisseria meningitidis* in its host and counteracting both innate and adaptive immune responses of the host. The mechanism of evolution involves the combined action of de novo mutation, recombination, and localised hypermutation. This study aimed to understand the contributions of these processes to within host evolution during host persistence of *N. meningitidis* for a period of months. A study of 40 isolates from one carrier and representing six months persistent carriage showed that de novo mutation resulting in single nucleotide polymorphisms (SNPs) was the major factor in structuring of the population. Allelic variants were subject to dynamic temporal fluctuations through persistence of meningococcal isolates over several months. Conversely, recombination was found to a powerful mechanism for generating SNPs and insertion/deletion within 25 paired isolates from 25 carriers and representing between 1 and 6 months host persistence. The processes of de novo mutation and recombination were infrequent but exhibited trends toward surface antigens especially pilin, porins, iron acquisition and capsule genes. Variation in intergenic regions was also examined in these isolates and a high level of variation was observed in conserved functional patterns of Corriea elements. Three carriers were examined for changes in expression of three phase variable genes (*opc*, *hpuAb*, and *nalP*); these were in the OFF state indicating that there may have been selection for low expression. A further investigation of microevolution within a clonal complex found a low rate of recombination within the 25 CC-174 disease and carriage isolates but it was many folds higher than recombination rates of species forming clonal population. Variable genes were distributed into several schemes including bacterial secretion systems, iron acquisition, capsule, surface antigen, bacterial mobility proteins, and antimicrobial resistance, and toxin genes. In conclusion, all the processes of evolution de novo mutation, recombination, and localized hypermutation facilitate asymptomatic carriage of meningococci and microevolution of CC-174.

ACKNOWLEDGMENTS

Praise to Allah, Lord of the Worlds for giving me opportunity to get knowledge and to finish my Ph.D. successfully. I introduce my special thanks to my supervisor Dr Christopher Bayliss for his scientific aid and valuable guidance towards awarding the PhD degree. I also deeply thank my second supervisor Dr Ralf Schmid for his scientific aid and critical guidance.

I am grateful to Dr Matthew Blades (Bioinformatics and Biostatistics Analysis Support Hub (B/BASH) - (University of Leicester), Prof. Martin Maiden (University of Oxford) and Dr Richard Haigh (University of Leicester) for their valuable advices.

My kindly respect and appreciation goes to my examinar Dr Alan McNally (Institute of Microbiology and Infection-University of Birmingham) and Dr Sandra Beleza (University of Leicester) who exhibited diligent care and effort in the critique of this thesis.

It is a pleasure to thank those who supported me financially especially the Higher Committee for Education Development in Iraq (HCED), Higher Education Ministry in Iraq and Baghdad University.

I indeed appreciate helping Prof. Haifa Hussani (Biology department-University of Baghdad) and my brother Omar as they aided me to complete requirements towards awarding the scholarship from HCED.

I am thankful to all the students of M.Sc in Bioinformatics field at Leicester University, staff members in lab. 121 at Leicester University, all the staff members in College of Science – Biology department-University of Baghdad.

I never forget helping my wife Shaymaa for the difficult time we spend together and for her great patience during the period of my study. I also never forget my mother, fathor, brothers, and my sisters for their support and their praying for me during all my Ph.D. study.

Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

°C	Degree Centigrade
A	Adenine
ABACAS	Algorithm based automatic contiguation of assembled sequences
AC	Allele comparator
ACT	Artemis comparison tool
BIGSdb	Bacterial Isolate Genome Sequence Database
BLAST	Basic local alignment search tool
bp	Base pairs
CC	Clonal complex
CDs	Coding sequences
CE	Correia element
CREE	Correia repeat enclosed element
DNA	Deoxyribonucleic acid
DUS	DNA uptake sequence
EDTA	Ethylenediaminetetraacetic acid

Fur	Ferric uptake regulator
g	Grams
GC	Genome comparator
H2O	Water
HGT	Horizontal gene transfer
HR	Homologous recombination
ID	Identifier
IGRs	Intergenic regions
IHF	Integration host factor
kbp	Kilobase pairs
KEGG	Kyoto encyclopedia of genes and genomes
l	Litres
M	Molar
MAFFT	Multiple alignment using fast fourier transform
MEGA	Molecular evolutionary genetics analysis

mg	Milligrams
MgCl2	Magnesium chloride
MLST	Multilocus sequence typing
ml	Millilitres
mM	Millimolar
Mw	Molecular weight
NaCl	Sodium chloride
NCBI	National Center for Biotechnology Information
Ng	Nanograms
NIME	Neisserial intergenic mosaic element
OMP	Outer membrane protein
ORF	Open reading frame
PC	Phylogenetic clade
PCR	Polymerase chain reaction
PhyML	Phylogenetic estimation using maximum likelihood
PNACL	Protein Nucleic Acid Chemistry Laboratory, University of Leicester

PV	Phase variation
REP	Repetitive extragenic palindromes
RMSs	Restriction modification systems
mRNA	Messenger ribonucleic acid
RAxML	Randomized axelerated maximum likelihood
SAP	Simple Allele discriminating PCR
SPAdes	St. Petersburg genome assembler
SNP	Single nucleotide polymorphism
SSR	Simple sequence repeats
ST	Sequence type
UniProt	Universal protein resource
μ	Upper bound of nucleotide divergence
μg	Micrograms
μl	Microliters
μM	Micromolar

1. Introduction

1.1. Overview of history, characterization and population structure of *Neisseria meningitidis*

Invasive Meningococcal disease (IMD) was discovered in 1805 after it claimed 33 lives in Geneva, Switzerland. The disease-causing agent for meningitis was discovered by Antone who called it "Diplococcus intracellular meningitidis" (Manchanda *et al.,* 2006; Yazdankhah and Caugant, 2004). However, following the discovery of *Neisseria gonorrhoeae* by Albert Neisser, the name of the bacterium was changed to *N. meningitidis. N. meningitidis* has a diplococcus shape and is a gram-negative β-proteobacterium bacterium (Stephens, 2009). In addition, this bacterium can be characterized by non-motility, non-sporulation, aerobic properties, palliation and encapsulation (Maiden *et al.*, 1998).The lack of apparent motility was countered when twitching motility was characterized in this organism though traversing of bacterial cells across the epithelium using type IV pili (Merz *et al.*, 1996; Wolfgang *et al.*, 1998).

This species colonises the mucosal surfaces of the upper respiratory tract especially the nasopharynx and has strategies to transfer from the mucosa into the blood stream and further into the brain. Therefore, during disease bacteria can be detected in cerebrospinal fluid (CSF) or blood of patients, saliva and throat and occasionally in the rectum and urogenital tract (Bidmos *et al.,* 2011; Doran *et al.,* 2016; Givan *et al.*, 1977).

Genotyping and phenotyping methods have been developed in order to elucidate the population structure and epidemiology of *N. meningitidis* (Harrison *et al.,* 2009). Serological grouping and serotyping are phenotyping methods that detect variation in the biochemical composition of capsular polysaccharides (CPS) and outer membrane proteins (OMPs) such as PorA and PorB (Russell *et al*., 2004). A variety of variable CPS structures surrounds the *N. meningitidis* with 13 different serogroups. Of note, only A, B, C, W and Y serogroups are responsible for most invasive infections in the human population (Oldfield *et al.*, 2016). Some serogroups are restricted to certain regions, for example, serogroup X is found only in the African population. The characteristic signature carried by serogroup A and X is N-acetyl-d-mannosamine-6-phosphate, while sialic acid is prevalent for B, C, W

and Y serogroups (Tzeng *et al.*, 2003; Litschko *et al.*, 2015; Schneider *et al.*, 2009). The worldwide distribution of various CPS of Neisseria is shown in (Figure 1.1). The serological typing scheme can lead to ambiguous typing as the CPS structures might be absent due to phase variation or other changes in the genes encoding such structures (Harrison *et al.*, 2009).

Genotyping techniques are very useful for differentiation between *N. meningitidis* strains for short-term epidemiological analysis. This includes a variety of techniques such as pulse-field electrophoresis (PFGE), RFLP (restriction fragments length polymorphism), and ribotyping. However, the major limitations associated with these techniques is that they cannot be used for studying variation over long time scales and they are not reproducible (Maiden *et al.*, 1998; Sabat *et al.*, 2013; Mechergui *et al.*, 2015). On the other hand, the variation that evolves slowly could be estimated by multilocus enzyme electrophoresis (MLEE). The variation that accumulates slowly is useful for understanding global epidemiology. The drawback of the MLEE method is that the results from different laboratories are difficult to compare therefore it was necessary to identify the allele sequences of seven housekeeping genes rather than trace the motilities of the enzymes that were coded by those allelic sequences (Maiden *et al.*, 1998; Bennett and Cafferkey, 2003). The principle of multilocus sequence typing (MLST) is that the sequences of seven housekeeping genes are different among different strains. The variation in the sequences of these genes can generate a multitude of different allele combinations, termed the sequence type, so that each strain has a novel type that can be differentiated from those of other strains (Maiden *et al.*, 1998). MLST is the most widely practiced technique for typing *N. meningitidis*, and is based on variations of seven house-keeping genes. These genes have been selected for discrimination of *N. meningitidis* strains due to their property of slow evolution, which means that they are not destabilized by selective pressures. Thus, it is easier to trace and analyze the variation in these genes in order to differentiate *N. meningitidis* strains (Mechergui *et al.*, 2015). A modern and reliable method to differentiate among different strains is eBURST, which can divide the strains into different groups of related isolates and clonal complexes (Feil *et al.*, 2004). The closely related strains with similar sequence type (ST) constitute a clonal complex (CC). In this way, each type of *N. meningitidis* is associated with a particular CC and specific ST (Budroni *et al.*, 2011). In

clonal models of populations, the structure of *N. meningitidis* populations is a phylogenetic clade (PC) structure in which acquisition of specific genes is influenced by restriction-modification enzymes (Budroni *et al.*, 2011). Recently, whole genome sequencing using next generation sequencing (NGS) has emerged as a gold tool for discriminating between different strains. It has demonstrated its advantages through tracking disease outbreaks in epidemiology studies (Fournier *et al.*, 2014; Salipante *et al.*, 2015).

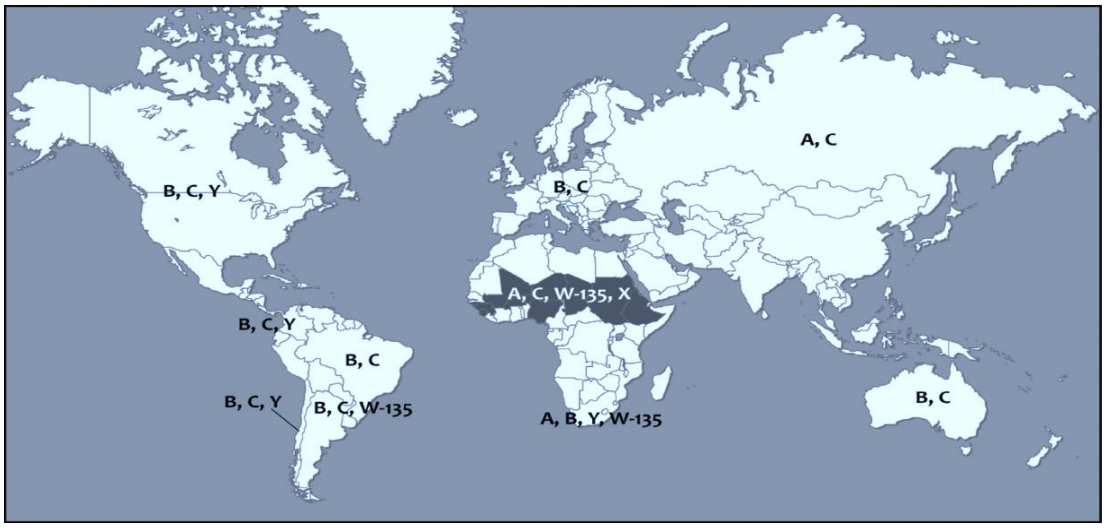


Figure1.1: Distribution of major meningococcal serogroups across the world. This figure was adapted from (Harrison *et al.*, 2009).

1.2 The genome features of *N. meningitidis*

Meningococcal genomes are highly dynamic, plastic, and flexible (Treangen *et al.*, 2008; Gasparini *et al.*, 2015; Schoen *et al.*, 2009). The average size of meningococcal genomes is 2.2 Mb with an average G+C content of 51.63 (± 0.25) % with coding regions constituting approximately 75-80% of the genome of which 82% are shared among all strains. A special characteristic of these genomes is the presence of many different repeat elements (e.g. simple sequence repeat) (Peng *et al.*, 2008; Schoen *et al.*, 2009; Parkhill *et al.*, 2000). Some of the repeat elements have a major role in genome diversity of *N. meningitidis* as they are considered hotspots for recombination and can vary by other mechanisms (Tettelin *et al.*, 2000; Bentley *et al.*, 2007). In addition, *N. meningitidis* has the largest repertoire of phase variable genes among prokaryotes which are an important mechanism for generation of reversible switching in the expression of different genes to evade the immune system and

adapt to other environmental changes (Tan *et al.*, 2016; Bentley *et al.*, 2007).

The most abundant repeat sequence in *N. meningitidis* is the DNA uptake sequence (DUS) which has a length of 12 base pairs (bp) and a consensus pattern of 5'-GCCGTCTGAA-3'. The main role of DUS is in the transformation process (Duffin and Seifert, 2010; Ambur *et al.*, 2007; Muzzi *et al.*, 2013). The neisserial intergenic mosaic elements (NIMEs) are the next most abundant repeat tract in *N. meningitidis*, and consist of multiple reverse sequence (RS) elements. These elements have 20-bp inverted repeats (ATTCCCNNNNNNNNGGGAAT) flanking 50-150bp spacers and they are a big source of variation as NIME arrays may act as substrates for recombination (Bentley *et al.*, 2007; Schoen *et al.*, 2009). Similarly, the Correia element (CE) is another abundant type of repeat and is also called Correia repeat enclosed elements (CREEs). These elements move by transposition and recombination. CEs are found as either complete or partial elements, which are 153-157 bp or 104-108 bp, respectively. They can cause changes in the expression of genes or inactivate particular genes (Liu *et al.*, 2002; Siddique *et al.*, 2011; Lin *et al.*, 2011; Roberts *et al.*, 2016).

Transposases and insertional elements are mobile elements that can also shift from one position into another position within the genome and can generate variation in the genomic sequence (Bentley *et al.*, 2007). Moreover, the genomes of *N. meningitidis* are intrinsically polygenic as they contain genetic islands of horizontal transfer including prophages, mobile elements and island-related hypothetical proteins, gonococcal genetic islands (GGIs) and meningococcal disease-associated island (MDA) (Snyder *et al.*, 2005; Marri *et al.*, 2010; Joseph *et al.*, 2011; Gasparini *et al.*, 2015). The multiple genetic islands are an important feature of *N. meningitidis* as there are differences among isolates and some of them relate with pathogenicity (Parkhill *et al.*, 2000; Hotopp *et al.*, 2006; Tettelin *et al.*, 2000). In addition to these various repetitive elements and genetic islands, genetic variation is generated by horizontal or lateral gene transfer, allelic exchange/rearrangement, multiple copy genes and homologous intragenic recombination leading to frequent surface structural variation (Gasparini *et al.*, 2015; Cahoon and Seifert, 2011).

1. 3 Carriage isolates

The upper respiratory tract especially the naso-pharynx is a reservoir for meningococci (Tryfinopoulou *et al.*, 2016). A commensal, asymptomatic relationship is established between human hosts and *N. meningitidis* located in the respiratory tract, and pathological symptoms are not observed at this stage. This phenomenon is termed as carriage (Caugant and Maiden, 2009). It has been reported that *N. meningitidis* is also found as part of the normal flora in buccal mucosa, anus, urethra, urogenital mucosa, and dental plaque (Stephens, 2009).

Most carriage studies were conducted to determine the carriage rate in specific populations and the persistence of carriage with their hosts as carriage is a reservoir of virulence determinants and for disease isolates. The prevalence of carriage and disease is correlated with age. Carriage is high (>23.7%) in the adolescent and young-adult age groups and significantly lower (<4.5%) in newborns and in school-age children (Gabutti *et al.*, 2015). In the African meningitis belt, the carriage rate is higher in:- males than females; in individuals aged 5–14 years rather than those aged 15–29 years; in rural areas rather than in urban areas (Ali *et al.*, 2015). The rates of incidence of meningococcal carriage ranged between 10–35% for healthy people. Carriage rates are significantly increased in closed populations such as university students or military recruits (Hill *et al.*, 2010). It has been reported that life style can affect on carriage rates of *N. meningitidis* with higher carriage rates being associated with active and passive smoking, concomitant viral or bacterial respiratory infections, low socioeconomic status and crowding (MacLennan *et al.*, 2006; Tryfinopoulou *et al.*, 2016).

Several meningococcal carriage studies have investigated the duration of carriage and its correlation with host immune responses during carriage. A study of persistent carriage was carried out on a cohort of 190 first-year students from six residential halls at Nottingham University between November 2008 and May 2009. The results showed that since the start of the academic year, rapid clonal expansion was proposed for three serogroups Y clones and one serogroup 29E clone due to high prevalence in particular halls in November 2008. In addition, 45% of carriers exhibited persistence for the same meningococcal strains. However, clonal replacement strains were observed in 36% of carriers with frequent

differences in the capsules and antigenic structures of PorA, FetA, NadA and HmbR. (Bidmos *et al.*, 2011). Many studies have shown that carriage isolates serve as reservoirs of virulence alleles for their pathogenic counterparts and are frequently involved in genetic exchange through horizontal gene transfer (HGT) (Marri *et al.*, 2010; Schoen *et al.*, 2008; Perrin *et al.*, 2002).

1. 4 Disease isolates

Following a successful colonization event, *N. meningitidis* sometimes manages to enter the bloodstream by passing through the mucosal barrier.This can lead to the development of various disease conditions in a very short period, usually under 24 hours with symptoms often appearing within 1 to 14 days after colonization of a host. The prominent diseases associated with neisserial infection are meningitis and septicaemia, but more rarely it can also cause arthritis, pneumonia, otitis and urethritis (Doran *et al.*, 2016; Nassif, 2010; Caugant, 2008; Thompson *et al.*, 2006; Tzeng and Stephens, 2000).

The carrier strains of *N. meningitidis* are known to interact with the host in a complex manner; and the outcomes of these interactions are regulated mainly by the titre of meningococcal-specific antibodies. The probability of risk of disease increases with generation of low titers of bactericidal antibodies, innate deficiencies in the host immune system and acquisition of pathogenic *N. meningitidis* strains (Orihuela *et al.*, 2009; Ram and Vogel, 2006; Picard *et al.*, 2011). Furthermore, the availability of nutrients such as transferrin, lactoferrin and haemoglobin as sources of iron and evasion from the host's immune system such as destruction of immunoglobulin by IgA protease is one of key importance for survival of *N. meningitidis* within the host (Caugant *et al.*, 1994; Jordan and Saunders, 2009). The carriage isolates maintain a large reservoir of virulence genes and presence of silent virulence factors in the carrier strains, environmental factors and host susceptibility in immune compromised individuals are the major determinants which can lead to the conversion of carrier strains into virulent strains (Pace and Pollard, 2012; Muzzi *et al.*, 2013). Nevertheless, it has been shown by the MLST method that diseases isolates of *N. meningitidis* are restricted to a very few CCs which are called the hyper virulent lineages (Claus *et al.*, 2005; Schoen *et al.*, 2008). According to Stephens (2009), carriage isolates rarely cause any disease symptoms in their hosts. Furthermore, the duration of persistence

of carrier strains in the host also varies, for example, they can hold a commensal relationship with the host transiently or chronically for several months before they are cleared by the host defense system (Bentley *et al.*, 2007; Caugant and Maiden, 2009). Both genotypic and phenotypic studies have demonstrated a great deal of diversity associated with carriage isolates as compared to disease isolates, for example, both capsulated and acapsulated forms are observed in carriage isolates but only capsulated forms are observed in disease isolates (Claus *et al.*, 2005).

The incidence of disease, meningococcal serogroup distribution and epidemics exhibits large fluctuations due to multiple factors such as population immunity and environmental stresses. In the sub-Saharan Africa, China and Russia, meningococcal serogroup A is highly distributed while serogroups X, W and C are additional major causes of epidemics in African countries, such as Niger, Togo, and Western Kenya. The serogroup Y strains predominately affect people in the USA and other countries in the Americas while serogroup B is dominant in the UK causing 90% of disease cases especially CC-41/44, CC-269, CC213 and CC32, *N. meningitidis* serogroup C was also recorded with high rates during 2015 in Italy (Harrison, 2010; Brooks *et al.*, 2006; Molesworth *et al.*, 2002; Sinclair *et al.*, 2010; Boisier *et al.*, 2007; Ladhani *et al.*, 2012; Chow *et al.*, 2016; Stefanelli *et al.*, 2016) but is not a major cause of disease in the UK due to effective use of a MenC vaccine.

The incidence of meningococcal disease is unstable and involves sporadic outbreaks with an average of approximately 1 disease case per 100000 in Europe, whereas Sub-Saharan countries experience epidemics with 1000 disease cases per 100000. The determinants of the mortality rate are the type and severity of invasive disease and it is recorded as being up to 55% in fulminant septicemia, while it is reported to be up to 25% and 5% in meningitis associated with septicemia and meningitis without sepsis respectively. In general, mortality occurs in 10 % of patients with invasive meningococcal disease (Brandtzaeg, 2006; Hill *et al.*, 2010).

1. 5 Tackling disease: antibiotics and vaccines

Early diagnosis of IMD is a vital to reduce the fatality rate. The polymerase chain reaction (PCR) approach is important for reducing diagnosis time but treatment is usually initiated on the observation of symptoms before presence of organisms is confirmed (Bosis *et al.*,

2015). Antibiotic therapy is the first line of treatment against acute bacterial meningitis and is given as early as possible for reducing fatality rate, as suggested by a UK study (Strelow and Vidal, 2013). Antibiotic therapy shows a significant killing action in three to four hours in the CSF, inhibiting proliferation of meningococci and significant reductions in the endotoxin in plasma by 50% within two hours of full treatment over a duration of 7 days (Strelow and Vidal, 2013; Ahmed-Abakur, 2014).

In developed countries, such as the USA and UK, intravenous penicillin or a third-generation cephalosporin is used as the initial treatment (Bosis *et al.*, 2015) with follow-up by ceftriaxone or ciprofloxacin for curing nasopharyngeal carriage. In cases of penicillin allergy, chloramphenicol and meropenem can be used; however, an extended-spectrum third generation cephalosporin is recommended (Nadel, 2016). Intravenous ceftriaxone followed by penicillin or ampicillin/amoxicillin are used in countries with no resistance to penicillin, otherwise intravenous ceftriaxone followed by vancomycin is recommended (McGill *et al.*, 2016). In cases of anaphylaxis for penicillin or cephalosporin, chloramphenicol is recommended (McGill *et al.*, 2016).

The drawbacks of using antibiotic therapy are concerns about resistant meningococcal strains. Penicillin, chloramphenicol and ciprofloxacin show reductions in the susceptibility to these antibiotics due to the ability of some meningococcal strains to alter penicillin-binding protein 2 (*PBP2*), *catP* gene encoding chloramphenicol acetyl transferase and *gyrA*, a gene associated with ciprofloxacin resistance (Ahmed-Abakur, 2014; Harcourt *et al.*, 2015). (Ahmed-Abakur, 2014; Harcourt *et al.*, 2015). The *gyrA* gene encodes the A subunit of the DNA gyrase enzyme that plays an important role in introduction of negative supercoils into DNA. The supercoiling triggers chromosome condensation due to the DNA helix winding around its axis in the opposite direction to the turns of right-handed helix, this process leads to an increase in the free energy that is stored in the DNA molecule. The supercoiling also enhances separation of the DNA strands and promotes the local melting which is essential for transcription and replication processes. The local melting is more favorable for the strands in negatively supercoiled than in relaxed DNA and triggers by the action of RNA polymerase. The mechanism of ciprofloxacin inhibition occurs via mutations in the quinolone resistance-determining region (QRDR) of the gyrA gene

(encodes subunit A of DNA gyrase). These mutations prevent the binding of ciprofloxacin with the DNA gyrase - DNA complex and inhibit the action of this antibiotic in blocking DNA replication and inhibiting DNA synthesis (Fàbrega *et al*., 2009; Witz and Stasiak, 2010 ; Hong *et al*., 2013; Enrquez *et al*., 2007).

High reductions in the susceptibility to rifampicin have also been reported and are comparable with ciprofloxacin resistance levels (Telisinghe *et al.,* 2014). Fortunately, reduction in the susceptibility to penicillin G is very low in the UK and African meningitis belt (Harcourt *et al.*, 2015; Hedberg *et al.*, 2009).

The first types of vaccines were primarily designed to induce immune responses against the polysaccharide capsules of serogroups A, C, Y and W and included Menomune (Sanofi Pasteur) and Mencevax (GlaxoSmithKline). These vaccines showed significant protection in army recruits, however the immune response did not elicit a long-term protective effect for the host (the duration was around 3 years) and no immune response was detected for children with ages of less than 24 months (Crum-Cianflone and Sullivan, 2016; Gasparini *et al.*, 2015; Gasparini and Panatto, 2011). The polysaccharide vaccines (MPSV) introduced since 1970 were initially against serogroups A or C in USA but were then combined as trivalent (serogroups A, C and W) or tetravalent (serogroups A, C, Y and W) vaccines (Cook, 1992; Vipond *et al.*, 2012). In the 1990s, new forms of capsular vaccines were generated that involved conjugating proteins and polysaccharide and are known as conjugate vaccines. These vaccines include Menactra (Sanofi Pasteur), Menveo (Novartis) and MenAfrivacTM. These vaccines have been utilized since 1999 and elicit immune responses with long periods of protection and powerful responses against capsule polysaccharide in adults and in children with ages of less than 24 months (Crum-Cianflone and Sullivan, 2016; Gasparini *et al.*, 2015; Cohn and Harrison, 2013; Gasparini and Panatto, 2011; Yogev and Tan, 2011). Conjugate MenACWY vaccines are currently in use in the USA and Europe (Gasparini *et al.*, 2015). The conjugate MenAfriVac vaccine is in use in the African meningitis belt and has produced a significant reduction from 0.7% to 0.02% in carriage of serogroup A strains (Ali *et al.*, 2015).

In the developed world, vaccines have been introduced against all serogroups causing meningococcal meningitis and septicaemia except MenB due to the variability of their

surface proteins (Shea, 2013). In addition, autoimmunity is thought to be a possibility for a polysaccharide vaccine against serogroup B strains due to the similarity between the homopolymer of sialic acid that comprises the MenB capsular polysaccharide and neural cell adhesion molecules. For this reason, new forms of vaccine were designed that enhanced immune responses against outer membrane proteins. One type of vaccine consists of outer membrane vesicles (OMVs) such as MeNZBÒ, which showed significant reductions in cases of serogroup B disease due to a specific strain in New Zealand (Crum-Cianflone and Sullivan, 2016; Caugant *et al.*, 1994; Caugant and Maiden, 2009). However, this vaccine showed limited protection against other serogroup B strains due to high variably in proteins contained in this vaccine (e.g. PorA) (Gasparini *et al.*, 2015). A new vaccine was designed to elicit immune responses against a wider range of serogroup B strains that is 4CMenB (Bexsero). This vaccine was licensed in 2013 and has been shown to produce significant reduction in MenB cases of disease in USA and Europe (O'Ryan *et al.*, 2014; Crum-Cianflone and Sullivan, 2016). Bexsero contains neisserial heparin binding antigen (NHBA) (GNA2132, fused with the GNA1030 protein), factor H binding protein (fHBP) (fused with the GNA2091 protein), Neisseria adhesion A (NadA) and OMVs of a New Zealand serogroup B strain (Gasparini *et al.*, 2015; Snape *et al.*, 2010). In the UK, the Bexsero vaccine was introduced into the infant immunisation schedule in September 2015 and involves three doses of vaccine at ages 2, 4 and 12 months. This vaccine elicits bactericidal antibodies against the three recombinant antigens plus PorA antigens and showed immunization against other meningococcal groups such as MenW CC-11 strain (Ladhani *et al.*, 2016).

1.6 Tracking infectivity of *N. meningitidis*

During transmission, the capsule is important for protection of *N. meningitidis* from environmental factors (Romero and Outschoorn, 1994). After gaining entry into the human body, *N. meningitidis* first colonizes the mucosal surfaces of the upper respiratory tract, and uses type IV pili to attach to nasopharyngeal epithelial cells and to avoid clearance by physical defense mechanisms such as mucus clearance (Stephens, 2009; Vernikos and Medini, 2014). After establishing its presence at the site of first attachment, the bacteria start to proliferate on the non-ciliated cells of humans by forming microcolonies (Stephens, 2009). Adhesion and invasion of *N. meningitidis* is facilitated by interactions between some

OMPs with their cognate host receptors such as opacity proteins (Opa) and the Opc protein that interact with carcinoembryonic antigen cell adhesion molecules (CEACAM) and heparan sulfate proteoglycan (HSPG). This interaction triggers host cell signaling mechanisms. During this proliferation stage, there may be immune system stimulation and release of cytokines such as IL (interleukin)-6 and IL-8 leading to release or recruitment of immune effectors such as defensins, serum, and phagocytic cells. Serum bactericidal activity (SBA) is a major protective effector. These effectors are countered by phase variation and antigenic variation of OMPs that helps in escaping these immune responses (Virji *et al.*, 1992; Sim *et al.*, 2000).

In a small fraction of people, *N. meningitidis* can enter into the blood stream. In this compartment, lipo-oligosaccharide (LOS), the main component of the outer membrane, can enhance the severity of disease (meningococcal sepsis). LOS induces the immune system to produce different types of interleukins such as IL-1, IL-6, IL-8, tumor necrosis factor (TNF) (Brandtzaeg *et al.*, 1995; van Deuren *et al.*, 1995). However, there are 12 immunotypes of LOS (L1– L12) and hence the LOS can diversify its antigenic structure. The OMPs of *N. meningitidis* play a significant role in the virulence and resistance to the immune systems in the blood stream of the host. The key proteins are fHBP, two porins (PorA and PorB), and Opc that all promote serum resistance and resistance to complement-mediated killing (Hill *et al.*, 2010). It has been reported that several bacterial factors facilitate entry of *N. meningitidis* into the meninges such as pili and Opa proteins by playing a major role in the interaction between the bacteria and the meningioma cells (Nassif, 2010; Hardy *et al.*, 2000).

1. 7 Virulence factors

N. meningitidis encodes several virulence factors that enable this organism to cope with host immune responses such as capsule, pili, lipopolysaccharide (LPS) and mechanisms of antigenic variation and iron acquisition.

1.7.1 Capsule

The capsule polysaccharide (CPS) is considered as crucial in the virulence of meningococci, helping the bacterium to evade the host's innate and adaptive immune systems especially for capsules of serogroups A, B, C, W, X, and Y (Jones *et al.*, 2016).

The meningococcal capsular operon comprises genes encoding for proteins involved in capsular synthesis and transport. One of the gene that is involved in the transport of CPS is the *ctrA* gene, which encodes a conserved meningococcal OMP (Gioia *et al.*, 2015). Capsule is involved in many important functions including resistance against phagocytosis and complement activation by limiting the deposition of C3 molecules. The α2,8-linked sialic acid homopolymers play a significant role in limiting the deposition of C3 molecules through its binding with the host regulatory protein factor H and preventing conversion C3 to C3b (Schneider *et al.,* 2007). The antiphagocytic and antibactericidal properties of CPS are partly due to their ability to act as a shield to reduce the binding of antimicrobial proteins (AMPs) to the bacterial surface (Tzeng and Stephens, 2015). Capsule switching refers to a change in capsule type due to transformation and homologous recombination of novel biosynthetic enzymes into the CPS locus. A capsule switch can assist the bacterium in immune evasion and an example is the switch in C: 2a:P1.7-2, 4 strain to W: 2a:P1.7-2, 4 strain from serogroup C to W. The evidence of capsule switching arises from sequence analysis through recombination of 45 kb of DNA that led into the exchange of the entire capsule locus (Tzeng and Stephens, 2015; Beddek *et al.*, 2009).

1.7.2 Pili/Fimbriae

Pili are outer membrane filamentous structures. In *N. meningitidis,* this extracellular filamentous organelle is a Type IV pili (Tfp) and is composed of a single structural protein called PilE. The Tfp are classified into two classes (class I and class II) based on their antibody reactivity. The assembly and disassembly of the pili is mediated by around 15 proteins. Tfp have several functions such as adhesion, intracellular signaling, competence, and twitching motility (Giltner *et al.*, 2012; Gault *et al.*, 2015; Carbonnelle *et al.*, 2005). DNA transformation is also ascribed to pili (Merz *et al.*, 1999). Pili have a major role in serum resistance through auto agglutination (Hubert *et al.*, 2012). The proliferation of *N. meningitidis* in the blood stream and during carriage relies on evasion of the immune system of pili specific antibodies. Recently, it has been reported that the Tfp (PilE and PilV) binds CD147 on the brain endothelial cells. This binding triggers signaling through a receptor and then formation of microcolonies on the meningeal tissues (Doran *et al.*, 2016).

Antigenic variation occurs in class I pilin, Sequence variation in class I pilin occurs mainly

by intragenic recombination with silent pilin genes (pilS) (Hill *et al*., 2010).

There is no evidence of antigenic variation in class II pilin (Davies *et al.*, 2014) but sequence variation could occur by normal horizontal gene transfer and mutation as occurs in other meningococcal genes. However, structural variation of a class II pilin can be mediated by variation in glycosylation through changes in the sugar modifications rather than the pilin primary structure. Several of the pilin glycosylation genes are subject to phase variation. Modification of PilE involves up to five glycosylation sites per monomer on the pilus surface of class II pili but only one in class I pili (Gault *et al.*, 2015). The pili modification through glycosylation is thought to induce protection of bacteria against immune responses (Virji, 2009).

1.7.3 Lipopolysaccharide

The *N. meningitidis* lipooligosaccharide (LOS) is composed of heptose, 3-deoxy-D-manno-2-octulosonic acid (KDO) and a lipid A moiety. The important part of the LOS for pathogenicity and immunogenicity is the lipid A moiety that is an acyl chain (C12 and C14) with attachments on the di-galactosamine backbone. The modification of lipid A moiety helps *N. meningitidis* in AMP resistance through inhibition of the interaction between AMPs and phosphorylated head group of lipid A. This modification comprises removal of the phosphate, addition of positively charged moieties and alteration in the degree of lipid A acylation (Tzeng and Stephens, 2015). The lipid A component induces proinflammatory cytokines during meningococcal sepsis and meningitis and this process is a key contributor to disease. The composition of phosphorylation and phosphoethanolaminylation of lipid A varies among disease and carriage isolates, phosphoethanolamine and sialic acid substitutions on the oligosaccharide were higher in invasive than carriage. These differences can enable *N. meningitidis* to induce immune tolerance (John *et al.*, 2016; Fransen *et al.*, 2009).

It has been reported that pili help *N. meningitidis* to attach to nasopharyngeal epithelial cells, while the LPS intervenes to strengthen this adhesion (Gasparini *et al.*, 2015). LPS also has a role in serum resistance. In a mutagenesis study, it was shown that mutations in four genes *NMB0065, NMB0352, NMB0638* and *NMB2076* encoding hypothetical protein,

sugar isomerase, UTP-glucose-1-phsphate uridylyl transferase and transferase respectively (encoding LPS biosynthetic genes) showed lower serum resistance compared with wild-type cells (Geoffroy *et al.*, 2003).

Meningococcal clones can display different LPS structures simultaneously through phase variation of various LPS biosynthesis genes that alter the saccharides chains on the surface-exposed face of the LPS. This leads to alterations in the antigenic properties of LPS (Hill *et al.*, 2010; Bayliss *et al.*, 2008). There are 12 (L1 to L12*)* different LPS immunotypes in *N. meningitidis* (Scholten *et al.*, 1994). Changes between some of these immunotypes can occur by phase variation. Differences in the immunotypes are mainly due to variations in the sugar moiety of LPS (terminal LPS structure) (Geoffroy *et al.*, 2003).

The serogroup B or C strains are often associated with L3, L7, and L9 immunotypes, which may be sialylated by endogenous sialyl transferases. Sialylation increases resistance to complement and induction of anti-LOS antibodies (Braun *et al.*, 2002; Jones *et al.*, 1992).

1.7.4 Opa proteins

The Opa and Opc proteins are important for virulence and are expressed from up to four loci (*opaA*, *opaB*, *opaD*, and *opaJ)* dispersed throughout the genome (Callaghan *et al.*, 2006). These proteins are composed of a conserved β-barrel domain with eight transmembrane loops and four variable surface-exposed loops (Hill *et al.*, 2010). The opacity-associated outer membrane protein (Opa) has a major role in adherence mediating binding to mucosal epithelia and endothelia. Opa proteins can also cause immunomodulation through interactions with different receptors of carcinoembryonic antigen cell adhesion molecule (CEACAMs) family on neutrophils, and T and B-lymphocytes and macrophages. There is some evidence that this leads to arresting of T-lymphocyte activation rather than infection induced cytotoxicity through activation of tyrosine phosphatases SHP-1 and SHP-2. In addition, these Kinase activities play a role in suppression of TLR2-dependent innate responses in epithelial cells. Therefore, they affect both adaptive and innate immune response. Moreover, interaction of Opa proteins with CEACAM1 of B-lymphocytes leads to induction of cell death in B-cell antibody production through Bruton's tyrosine kinase rather than inhibitory signals or SHP-1 and SHP-2 (Callaghan *et al.*, 2006; Sadarangani *et al.*, 2011). The Opc protein mediates

adhesion through either binding with serum vitronectin on epithelial cells to form a trimolecular complex or binding directly to HSPGs (Hill *et al.*, 2010).There is also some limited information on a role for Opa/CEACAM receptor interactions with the brain endothelial vessel cells (Doran *et al.,* 2016) and for Opc mediated direct or indirect (through vitronectin or fibronectin) binding to brain endothelial cells, which processes may enhance the bacterial ability to cause meningitis. Some signaling pathways have been recorded for this binding (Doran *et al.*, 2016).

A study carried out by Liu *et al*. (2016) to investigate bacterial adaptation in a mouse model of infection found that the expression states of 51 proteins changed during infections caused by *N. meningitidis*. These 51 proteins were mainly found in the bacteria collected from the blood and CSF as compared to bacteria in the nasal mucosa. The glutamate dehydrogenase (gdhA) and Opa proteins exhibited high abundance in all the isolates under the study among the 51 proteins that showed changed during infection. This indicates that these genes have an important role in bacterial survival in vivo.

Recombination is a major source of variation in the *opa* genes of four closely related ET-37 complexes of *N. meningitidis* through Horizontal gene transfer (HGT) (Hobbs *et al.,* 1998).

1.7.5 Autotransporters proteins

The autotransporter proteins include different types and are used by *N. meningitidis* to help adhesion and other aspects of infections such as IgA1 protease; meningococcal serine protease A(MspA); adhesion and penetration protein (App) ; autotransporter serine protease (AusI), NadA and NalP (Khairalla *et al.*, 2015; Del Tordello *et al.*, 2014). Autotransporter proteins consist from C-terminal translocator domain, a central passenger domain and N-terminal signal peptide. The C-terminal translocator domain forms the main domain with a beta barrel structure while the passenger domains of autotransporters proteins encode the functions of the protein and are highly diverse among *N. meningitidis* strains (Benz and Schmidt, 2011).

The IgA1 protease, App and MspA are part of the S6-peptidase family of autotransporters that induce toxicity and/or immune modulation through degradation of extracellular or intracellular host proteins (Ruiz-Perez and Nataro, 2014). At the adhesion stage, App

proteins help in the attachment of *N. meningitidis* to epithelial cells through bacterial aggregation and microcolony formation on the epithelial cells, while at the invasion stage the protein is degraded to enhance meningococcal deattachment and dispersal (Khairalla *et al.*, 2015; Serruto *et al.*, 2003).

The *nadA* gene encodes a protein with three domains, which are an N-terminal globular domain, intermediate a-helix and conserved C-terminal membrane domain (Bambini *et al.*, 2014; Malito *et al.*, 2014; Capecchi *et al.*, 2005). The *nadA* gene is present in approximately 30% of clinical isolates and can be separated into three variants NadA-1, NadA-2, and NadA-3 whereas this gene is only present in approximately 16% of carrier isolates with an extra variant, NadA-4. Recently, a new protein variant the NadA-5 was identified in the ST-213 clonal complex (Bambini *et al.*, 2014). The NadA protein is a strongly immunogenic antigen that can elicit bactericidal antibodies and a protective response in the infant rat models, these observations resulted in its inclusion as one of three recombinant antigens of the Bexsero vaccine (Vogel *et al.*, 2013; Comanducci *et al.*, 2002). The NadA-1 is carried in the ST-32 isolates and NadA-2 and NadA-3 in the ST-8 and ST-11 lineages while NadA-3 is associated with ST-174 strains, mostly serogroup Y (Bambini *et al.*, 2014). The *nadA* gene is subject to phase variation due to a repeat tract (TAAA) located upstream of the -35 element of its promoter. Changing the repeat tract induces changes in the level of expression of NadA, resulting in reductions in expression (Martin *et al.*, 2005; Hill *et al.*, 2010). It has been reported that the *nadA* gene is down regulated by NadR therefore, during infections *N. meningitidis* isolates can respond to niche-specific signals resulting in repressing NadR activity and elevated *nadA* expression (Fagnocchi *et al.*, 2013; Cloward and Shafer, 2013). For example, in human saliva, 4-hydroxyphenylacetic acid (4-HPA) can also down regulate NadR activity and elevate *nadA* expression (Metruccio *et al.*, 2009).

1.7.6 Porin proteins

The mechanism of phase variation of PorA occurs through a tract of guanidine residues within the promoter regions. The spacer of the core promoter varies in length between 14 and 24 bp, this variation generates different levels of expression. In addition, it has been reported that the deletion of *porA* has been detected in some isolates due to recombination

(Peak *et al*., 2014, Van der Ende *et al.* 1999). The PorA protein is a porin which is located within the outer membrane and has two major variable regions (VR1 and VR2). Strong serum bactericidal antibody (SBA) responses are elicited against PorA during infections which is antigenically highly variable and the responses are sub type specific so not great for a vaccine against multiple strains (Boan *et al.,* 2014; Behrouzi *et al.,* 2014). The PorA and PorB proteins are part of the serological typing schemes for meningococcal isolates with PorB giving the types and the PorA OMP being the subtypes (Russell *et al*., 2004). PorA and PorB serves as channels for diffusion of hydrophilic molecules across the outer membrane (Peak *et al*., 2014; Rouphael and Stephens, 2012). It has been reported that loops located on surface of PorB protein espcially loop regions VR1, VR2, VR3, and VR4 are antigenically variable among different isolates. Some of these loops induce immune responses through Toll-like receptor 2 (TLR2)-mediated signaling (Kattner *et al.,* 2014). It has also been reported that variation in the *porA* and *porB* porins can be acquired through horizontal genetic transfer and recombination to form mosaic alleles (Dyet and Martin, 2005).

1.7.7 Mechanisms of Iron Acquisition as virulence factors of the meningococcus

Iron is necessary for multiple metabolic processes of both pathogenic bacteria and their hosts. Hosts use several different strategies to restrict iron availability for pathogens. As a result, many bacterial species, including *N. meningitidis* use different strategies for iron acquisition under iron-replete or iron-restricted conditions (Parrow *et al.,* 2013). Iron acquisition systems in *N. meningitidis* include four loci, which encode outer-membrane proteins; these are *hmbr, hpuAB, tbpAB* and *lbpAB* (Bidmos *et al.,* 2015). Availability of sources of iron in the body is different for each place of colonization in a host, in mucosal secretions iron is bound to lactoferrin, in CSF fluid iron is bound to transferrin and in the blood stream iron is present in transferrin, haemoglobin and haemoglobin-haptoglobin complexes (Jordan and Saunders, 2009). LbpAB is the receptor for lactoferrin and it provides a source of iron during adhesion in the nasopharynx (Perkins-Balding *et al.,* 2004). During the early stages of an infection, TbpAB is a receptor for iron-loaded transferrin and acts as a source of iron, in the later stages of infections Hmbr and HpuAB are potential receptors for haemoglobin and haemoglobin-haptoglobin complexes (Bidmos *et al.,* 2015). A study was carried out by Bidmos *et al.* (2015) on the *hmbr* iron acquisition

gene in MC58, a strain that lacks the HpuAB receptor. Mutants with deletions of *hmbr* and/or *tbpBA* were constructed and grown in healthy human blood, growth of an MC58 *hmbr* mutant was not affected while the growth of a MC58 *tbpAB* mutant was disrupted. This indicated that *N. meningitidis tbpAB* genes might be critical in blood for acquisition of iron from transferrin rather than *hmbr* during the early stages of disease. The iron acquisition genes are regulated by the Ferric uptake regulator (Fur) which forms a complex with free iron molecules and then binds to specific DNA sequences in the promoters of iron-regulated genes and represses gene expression (Lee and Helmann, 2007). During iron starvation, Fur protein is not complexed with free iron molecules and hence is unable to bind to its DNA sequence that results in activation of gene expression (Delany *et al.*, 2004). In addition to Fur regulation, HpuAB and Hmbr are also controlled at the level of translation by phase-variable polyG tracts in the reading frame (Bidmos *et al.*, 2015).

1.7.7.1 Transferrin and lactoferrin transportation systems of the meningococcus

TbpAB and LbpAB are outer membrane complexes formed of TbpA and LbpA, TonB-dependent transmembrane proteins, and TbpB and LbpB, surface-exposed lipoprotein. These complexes can efficiently strip iron that is sequestered by transferrin and lactoferrin (Wong *et al.*, 2015).

The initial capture of iron-loaded transferrin is achieved by fatty acyl chains of an N-terminal peptide region of TbpB that extend from the outer-membrane, then the iron passes through a pore in the TbpA component. Contrastingly, the major role of LbpB may be in protection against microbial peptides rather than to increase the affinity of acquisition of iron (Morgenthau *et al.*, 2014). LbpA is however integrated in the outer-membrane and forms a channel for passage of iron stripped from lactoferrin. There are two distinct antigenic variants of TbpB, which are isotype I and isotype II (Adamiak *et al.*, 2015). It has been reported that TbpA and TbpB are strong inducers of bactericidal antibody response, TbpB is however highly variable and so TbpA is preferred as a target for vaccination (West *et al.*, 2001). Critically, *tbpA* and *tbpB* are found in most disease and carriage isolates. It has been reported that the repeat arrays termed NIME are targets for intergenic recombination and exhibit variation around *tbpAB* in the FAM18, MC58 and Z2491 strains, which indicates that *tbpA* and *tbpB* may be subjected to recombination on a frequent basis

(Bentley *et al.*, 2007).

1.7.7.2 Hmbr and hpuAB transportation systems of the meningococcus

HpuA/HpuB can extract haem from haemoglobin or haemoglobin–haptoglobin complexes (Harrison *et al.*, 2013). HpuA is a surface-exposed lipoprotein while HpuB forms a TonB-dependent pore for entry of iron or haem. Hmbr is a transmembrane protein which can bind directly to haemoglobin (Harrison *et al.*, 2013; Tauseef *et al.*, 2011; Rohde *et al.*, 2002).

A study carried out by Tauseef *et al*. (2011) on the distribution of polyG tract lengths and ON/OFF status showed that both receptors Hmbr and HpuAB were highly prevalent in invasive isolates and were in >90% of isolates from CC5, CC8 and CC11 while HpuAB alone was underrepresented in disease isolates and was completely deleted or replaced by an insertion element in many isolates. The one or both receptors were found in 91% of disease isolates in ON state. This result may indicate that these genes are important for meningococcal virulence during invasion, however Lucidarme *et al.* (2013) showed that only 76.3% of disease isolates had one or the both receptors in an ON state. The high prevalence of *hmbr* in disease isolates may be due to haemoglobin being important for blood infections. It has been reported that antigenic variation of Hmbr is due to a combination of intraspecies horizontal genetic exchange and de novo mutation. The variation was localized in three variable regions (VR1–VR3) which encode outer membrane loops 2, 3 and 4 and are subject to diversifying selection presumably due to immune attack (Harrison *et al.*, 2013).

1.7.7.3 Mechanism of indirect iron-acquisition from human of the meningococcus

In general, most gram-negative bacteria produce siderophores for acquisition of iron; *N. meningitidis* can utilize siderophores that have been generated by other microorganism through the ferric enterobactin transport (FetA) protein (Noinaj *et al.*, 2013; Morgenthau *et al.*, 2014).

FetA is a TonB-dependent outer-membrane receptor consisting of a 22-stranded β-barrel, 11 surface exposed loops and an N-terminal domain (Harrison *et al.*, 2013). A variable region encoded by *fetA* defines nine families, F1 to F9 (Boan *et al.*, 2014; Norheim *et al.*, 2015; Sanders *et al.*, 2015; Marsay *et al.*, 2015).

Localized hypermutation in the poly C tract between the -35 and -10 boxes of *fetA* is associated with persistent meningococcal carriage and a reduction in expression, which may be due to selection mediated by specific immune responses (Alamro *et al.,* 2014). In the general carriage population, point mutations and horizontal genetic exchange are the main drivers of antigenic diversity of FetA (Thompson *et al.,* 2003). It has, however, been reported that deletion of *fetA* has occurred in some isolates possibly mediated by repeat arrays flanking the gene - such as Correia elements, duplicated RS3 and repeat sequence 13 (RS13) that are potential targets for genetic recombination (Claus *et al.*, 2007; Marsh *et al.*, 2007).

1.8 Mechanisms of genetic variation

Genome variation in bacteria arises by different mechanisms through acquiring DNA from other strains/isolates or through changing the genome content in what is known as substitution or insertion/deletion (indel) events. In general, transference of DNA between bacterial cells is carried out by three mechanisms that are conjugation, transformation and transduction (Soucy *et al.*, 2015). Transduction is the process of transferring DNA between two bacterial cells by phage, with two different approaches, which are generalized transduction and specialized transduction (Arber, 2014). However, transformation is the main process of acquiring the DNA through HGT in *N. meningitidis*. Genome diversity in bacteria is strongly driven by homologous recombination (HR), non- homologous recombination and site specific recombination (Niehus *et al.*, 2015). HR and site specific recombination are the main processes occurring in *N. meningitidis* but intragenic recombination also occurs at high rates in the pilus locus and at lower rates in other multi copy genes (Niehus *et al.*, 2015). Nonrandom mutation occurs through several mechanisms such as deamination, alkylation or depurination of specific bases where as random mutations occur through errors in DNA replication (Wright, 2000). The following sections provide more details on the mechanisms of genome variation occurring in *N. meningitidis*.

1.8.1 Natural transformation

Transformation is the process of receiving naked DNA from the environment; the naked DNA becomes available through release from dead bacteria or excretion by living bacterial cells (Johnston *et al.*, 2014). It depends entirely on the ability of bacteria to form natural

competence, which is tightly controlled by factors such as quorum sensing and nutritional signals (Macfadyen *et al.*, 2001; Hamoen *et al.*, 2003). The process of natural transformation in gram-negative bacteria compromises three steps, which are foreign DNA binding, entrance of the DNA through components of the cell membrane and DNA recombination (Duffin and Seifert, 2010). In Neisseria sp., transformation occurs through DNA binding mediated by specific or non-specific receptors, DNA up-take through the outer membrane is mediated by type IV pili (Tfp) (Cehovin *et al.*, 2013). The Tfp comprises the pilin structural subunit PilE and minor pilin proteins such as ComP (Wolfgang *et al*., 1998; Assalkhou *et al*., 2007). ComP enhances natural competence in *N. meningitidis* (Cehovin *et al.*, 2013). This is because ComP has DUS-binding activity and so the pilus can bind DNA with higher affinity when it contains DUS (Cehovin *et al.,* 2013; Berry *et al.,* 2016). On the other hand, PilV functions antagonistically to ComP, and reduces the DUS-binding activity (Berry *et al.,* 2013).

DNA transformation is facilitated by PilT (a protein involved in pilus retraction), which supplies the energy for the Tfp/pseudopili to pull DNA through the outer membrane (Cehovin *et al.*, 2013). DNA passes through the PilQ complex that forms an outer-membrane channel (Assalkhou *et al.*, 2007). ComE and ComL facilitate transfer of dsDNA through the periplasm and the peptidoglycan layer and delivery of DNA to the translocase machinery (Draskovic and Dubnau, 2005; Benam *et al.*, 2011; Duffin and Seifert, 2010). The *comE* gene enhances the efficiency of transformation though nonspecific binding associated with non DUS-specific receptors (Chen and Gotschlich, 2001). ComA forms a channel for transfer of DNA across the cytoplasmic membrane whereby one strand is degraded by the nuclease, and the other stand is protected by the cytoplasmic proteins (Draskovic and Dubnau, 2005). Finally, RecA mediates integration of DNA into the chromosome of the recipient cell (Cehovin *et al.*, 2013; Draskovic and Dubnau, 2005) (Figure 1.2).

Neisseria sp. contain thousands of DUS sequences in their genomes (~1% of the entire chromosomes). The DUS pattern has a sequence with 12 bp that can increase the efficiency of transformation over non-DUS containing DNA whilst a single mismatch in the DUS core 5′-CTG-3′ reduces the efficiency by more than two orders of magnitude (Frye *et al*., 2013).

Ambur *et al.* (2012) investigated other factors that affect transformation in *N. meningitidis*; the results showed that the NlaIV restriction system decreased the efficiency of transformation when the transforming DNA segment was associated with a heterogeneous region whereas the efficiency increased with homologous regions.

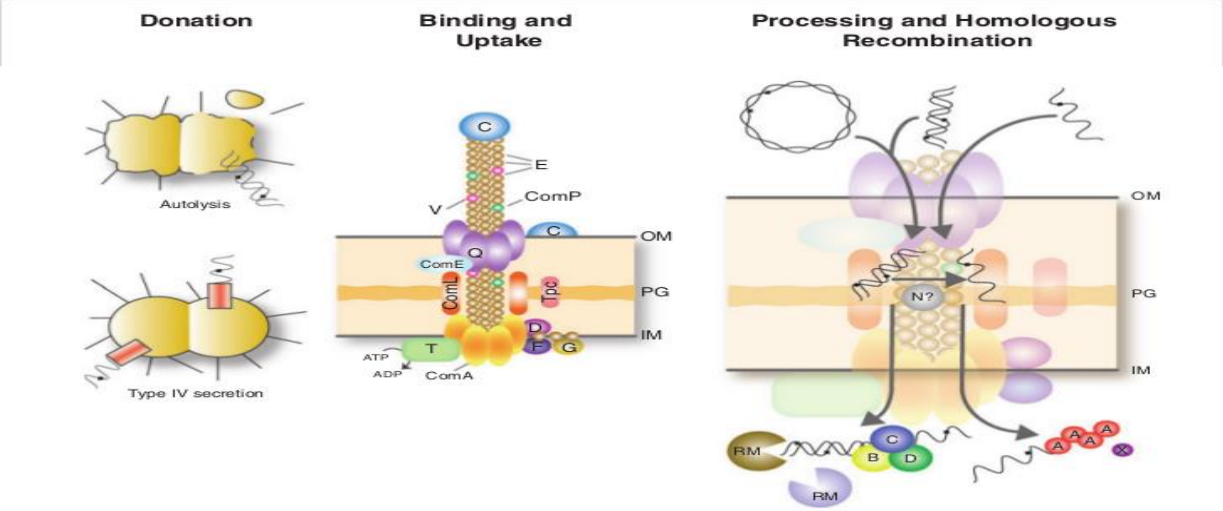


Figure 1.2: Transformation events associated with different processes in Neisseria. Type IV secretion of DNA may act to donate the DNA in a process known as DNA donation. Pilus-related proteins such as ComP, ComL, ComE, ComA and Tpc act to bind and mediate uptake of the DNA into the periplasm. Finally, processing of dsDNA into ssDNA occurs and homologous recombination mediated by RecA. This figure was adapted from (Hamilton and Dillard, 2006).

1.8.2 Recombination

Microorganisms are able to interchange and alter DNA through recombinatorial processes including HR, non-homologous recombination, intragenic recombination (resulting in antigenic variation, phase variation, or changes in repeat arrays), and site-specific recombination (Schoen *et al.*, 2009; Feavers *et al.*, 1992). HR is a powerful tool for inducing genetic exchanges between closely related isolates and species that have high sequence homology. Non-homologous recombination occurs when there is no homology or

low homology between the foreign and recipient DNA. Recombination can also mediate integration of phage genomes or conjugative elements, where homology is between short sequences in the integrating agents and the bacterial chromosomal DNA (Bryant *et al.*, 2012).

1.8.2.1 Homologous recombination (HR)

There are two mechanisms of HR controlled by RecA, one is specific to double-strand ends and is mediated by the RecBCD pathway, and the other is specific to replication gaps and is mediated by the RecF-pathway. In mutants or strains lacking the RecBCD pathway, the RecF-pathway can compensate and initiate recombination on double-strand ends (Spies and Kowalczykowski, 2005). The helicase/nuclease activity of the RecBCD enzyme is involved in the exonucleolytic processing at double-strand breaks, while RecJ exonuclease and RecQ helicase accomplish these activities in the RecF- pathway (Persky and Lovett, 2008). The assembly of the RecA protein on ssDNA is mediated by either RecBCD or the RecF, RecO, and RecR proteins. Invasion of single ssDNA into homologous dsDNA is initiated and driven by RecA (Spies and Kowalczykowski, 2005). The resolution of the intersection between foreign and recipient DNA is conducted by RecG helicase and RuvABC, which mediate migration and cleavage of Holliday junctions. Additionally, the PriA replication restart protein may be involved in processing of recombination intermediates, such as from D-loops, into replication forks (Persky and Lovett, 2008) (Figure 1.3).

In Neisseria, all the enzymes required for HR are present and so the RecBCD and RecF pathways probably drive HR. However, Experimental work suggests a third pathway mediates HR as a RecBCD-/RecF- mutant had a reduced efficiency of transformation of 40 fold but still exhibited incorporation of foreign DNA. RecX controls the activity of RecA as strains lacking RecX showed fivefold reductions in efficiency of transformation (Hamilton and Dillard, 2006; Beyene *et al.*, 2016). A *recJ* mutant did not change the efficiency of transformation whereas *recN* reduced the efficiency (Skaar *et al.*, 2002).

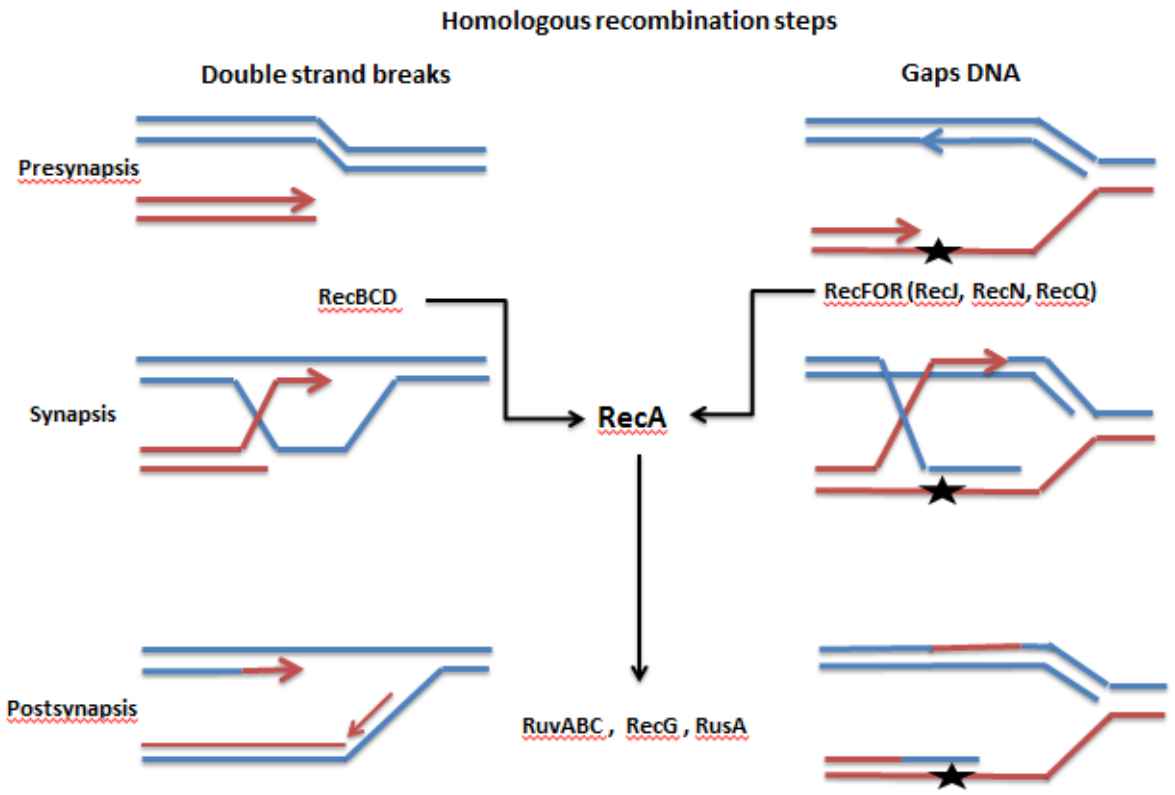


Figure 1.3: Mechanism of HR in bacteria. HR occurs either on double strand break DNA or gapped DNA with three different stages which are presynapsis, synapsis and postsynapsis. Presynapsis involves converting dsDNA into ssDNA for RecA loading. Synapsis involves the formation of D-loops mediated by RecA through strand exchange of homologous sequences. Postsynapsis involves migration of Holliday junctions by RuvAB and resolution of this structure by RuvC or RusA. This figure was adapted from (Benghezal *et al.*, 2014).

1.8.2.2 Site-specific recombination

Site-specific recombination involves breaking, exchanging, and rejoining four DNA stands at two different recombination sites. Site-specific recombination mediate integration, excision, and inversion of DNA stretches by the action of site-specific recombinases of the tyrosine or serine recombinase families (Grindley *et al.*, 2006).

The disposition of the recombination sites is a crucial component of this type of recombination resulting in different outcomes such as inversion as mediated by inverted recombination sites (Hallet and Sherratt, 1997; Grindley *et al.*, 2006).

The recombinases recognise and bind to recombination sites in association with other proteins to form synaptic complexes (Echols, 1990; Hallet and Sherratt, 1997). Recombinases use the energy released by breaking DNA strands to rebuild the new recombinant patterns (Grindley *et al.*, 2006; Sadowski, 1986). Recombination by tyrosine recombinases involves formation of a Holliday junction intermediate though breaking and rejoining single strands in pairs. In contrast, serine recombinases initiate cutting of all the strands followed by exchange and re-ligation (See Figure 1.4; Grindley *et al.*, 1997; Grindley *et al.*, 2006).

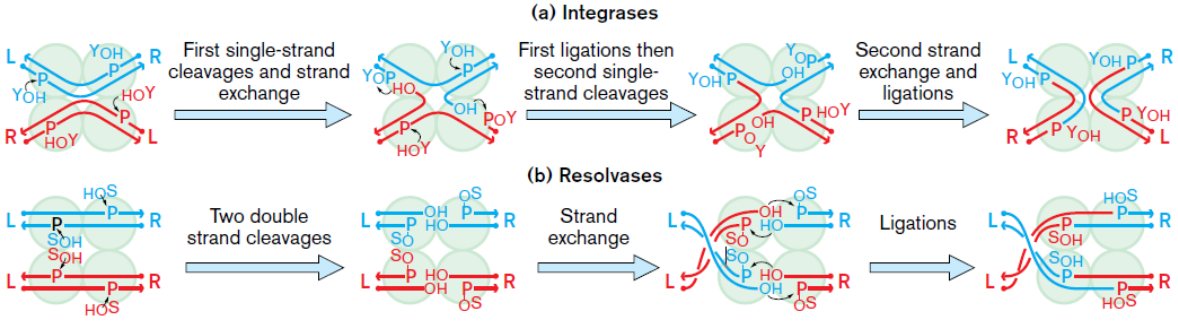


Figure 1.4: Mechanism of site specific recombination mediated by repeat arrays. A: Recombination mediated by integrases occurs through a single strand break, exchange and ligation in the first recombination pattern then another single strand break, exchange and ligation in the second recombination pattern. **B:** Recombination by resolvases occurs through double strand break, exchange, and ligation in both recombination partners. This figure was adapted from (Grindley *et al.*, 1997).

Conservative site-specific recombination plays a major role in adaptation by enhancing genetic diversity of specific loci or by changes in expression of specific genes by flipping the orientation of promoters (Nash, 1996).

Contrastingly, transposable elements are able to move from one position into another and hence can affect any gene with an insertion site. There are two mechanisms, which are non-replicative and replicative transposition. The non-replicative transposition involves excision from the original sequence and integration into a new location, leaving a gap in the donor that can be filled by the action of HR. Replicative transposition involves creation of a copy of the transposon element and then movement into the same DNA or another DNA molecule. In replicative transposition, transposons move into another DNA molecule, cointegrate structures are formed by joining repeated copies of the transposon at each junction followed by recombination between the two molecules to generate the original transposon and a new copy in the target molecule (Hallet and Sherratt, 1997).

N. meningitidis uses a number of different repeat elements as a target for recombination. Some examples are reported as follows:-

Neisseria genomes harbour an AT-rich of 183 bp in length termed the Miniature Inverted-repeat Transposable Elements (MITEs). These elements can change gene expression when they move into promoter regions or terminate transcription when they move into coding regions (Croucher *et al.*, 2011).

Van der Ende *et al.* (1999) examined the deletion of *porA* in three different Neisseria strains and found that recombination between RS3 repeats was responsible for these deletions.

Site-specific recombination is an alternate mechanism of phase variation between ON/OFF states. In this case, the recombinase changes the orientation of promoter sequences by an inversion event mediated by inverted repeats located around the promoter. This mechanism is found in fim gene in *E. coli* (Gally *et al.*, 1994; Wisniewski-Dy and Vial, 2008) (Figure 1.5).

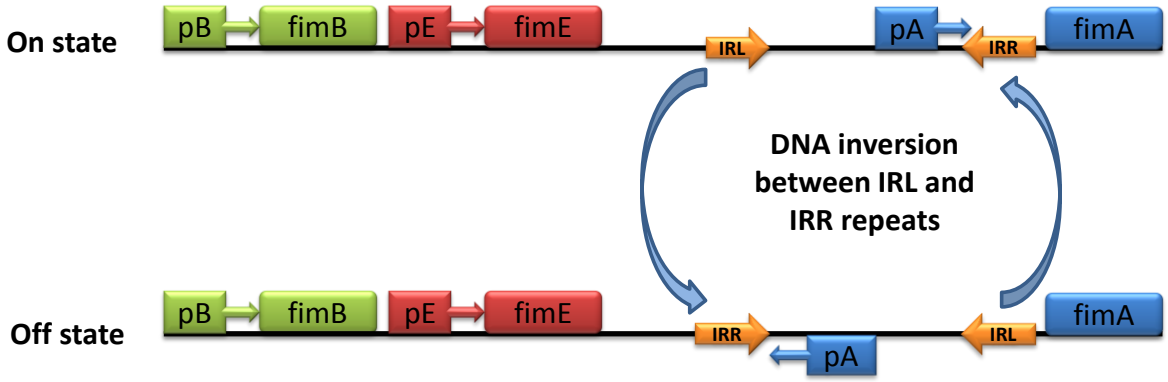


Figure 1.5: Example on mechanism of site specific recombination mediated by phase variation. The inversion event mediated by inversion repeat left (IRL) and inversion repeat right (IRR) leads to inversion of the promoter pA belonging to gene fimA and results in a switch in the direction of transcription from ON (top) to OFF (bottom).

1.8.2.3 Intragenic recombination mediated antigenic variation

The mechanism of antigenic variation is well documented in *N. gonorrhoeae* and assumed to be happening similarly in *N. meningitidis* (Aho and Cannon, 1988). The *pilS* genes, which are silent genes located downstream of the *pilE* gene, consist of semivariable and hypervariable regions without a promoter (Urwin *et al.*, 2002).

Gene conversion between *pilE* and *pilS* genes (insertion of "mini cassettes" from *pilS* into *pilE*) require three important elements, conserved and variable regions within the *pilE* gene, repetitive elements within *pilS* and a recombinase binding site near to *pilE* (Urwin *et al.*, 2002; Haas and Meyer, 1986) (Figure 1.6).

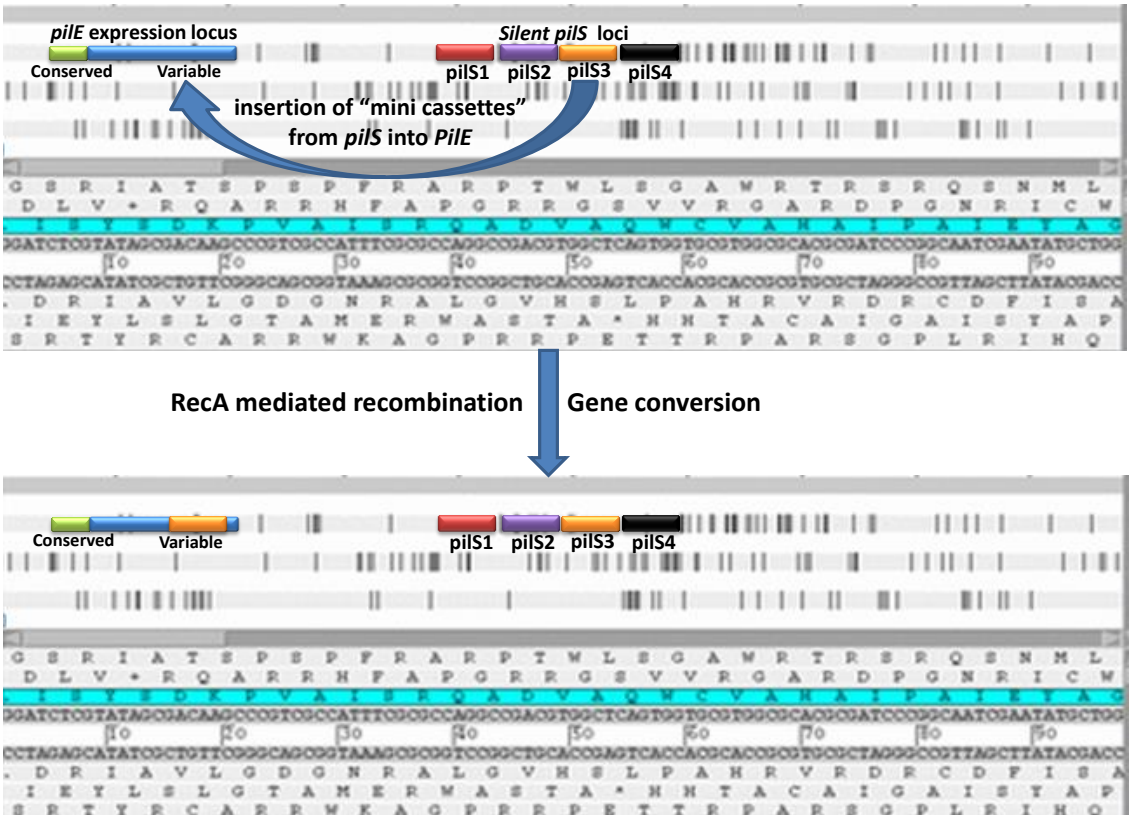


Figure 1.6: Schematic representation of an antigenic variation in *N. gonorrhoeae*. Gene conversion occurs through HR mediated by RecA between *pilS3*, a silent *pilS* gene, and the variable region of the expressed *pilE*. This event leads to antigenic variation but occasionally can result in a stop to the expression of *pilE*.

RecBCD recombination is not necessary for antigenic variation, however the RecF-like recombination and Holliday Junction pathways are required for antigenic variation (Obergfell and Seifert, 2015; Cahoon and Seifert, 2011; Hill and Davies, 2009). The Sma/Cla repeat carrying SmaI and ClaI restriction endonuclease sites may also participate in the antigenic variation as its absence produces a reduction in the efficiency of antigenic variation (Obergfell and Seifert, 2015).

The initiation of antigenic variation starts with the separation of dsDNA strands through transcription of a noncoding sRNA located near to the 5' end of *pilE*; this triggers formation of a G4 quadraduplex in one strand of the DNA (Obergfell and Seifert, 2015). This G4 quadraduplex structure stalls the replication fork, which enhances nicking of the strand opposite the G4. The resolution of this structure is mediated by RecQ and RecJ which degrade one DNA strand from the nick generating ssDNA thereby allowing RecA to initiate recombination (Figure 1.7) (Cahoon and Seifert, 2011).

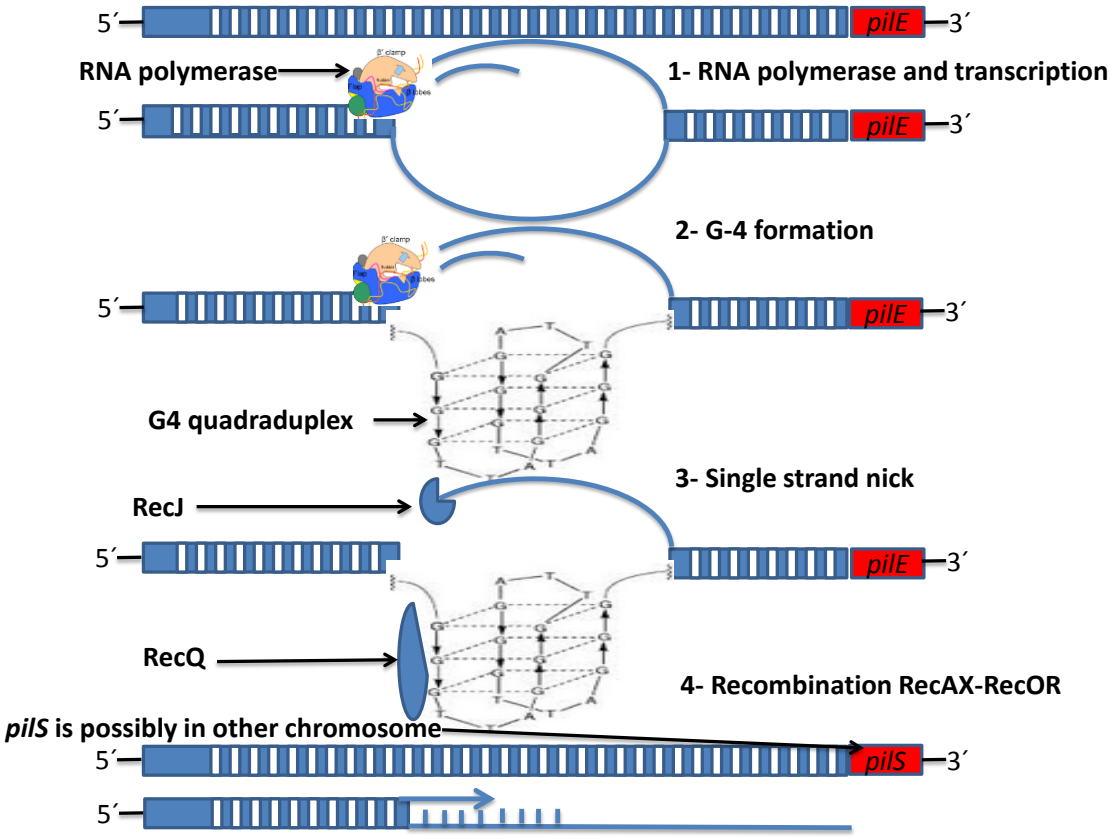


Figure 1.7: Schematic representation of the initiation of antigenic variation in *pilE*. Initiation of antigenic variation occurs through formation of a G4 quadraduplex structure mediated by transcription of an sRNA upstream of *pilE*. Stalling of the replication fork leads to cleavage of the strand opposite the G4 and invitation of recombination by RecQ and RecJ. These events trigger RecA to mediate recombination between *pilE* and *pilS* genes.

There are three models for how RecA mediates antigenic variation: - unequal crossing-over, successive half crossing-over and a hybrid intermediate model (Obergfell and Seifert, 2015). In the unequal crossing-over mode, RecBCD cleaves the 5′ ends after breaking the dsDNA at the *pilE* locus, leaving 3′ ends free to be associated with RecA, and then a D-loop is constructed and invades the *pilS* locus (Hill and Davies, 2009). Then, the *pilS* gene is used by DNA polymerase as a template to amplify the 3′ end. Finally, the double Holliday junction is resolved leading to re-forming of the altered *pilE* sequence with no change in the donor *pilS* sequence (Obergfell and Seifert, 2015).

In the successive half crossing-over model, half crossing-over events initiate recombination between the *pilE* and a *pilS* locus on a sister chromosome using RecA and RecOR after breaking dsDNA at the region of homology in the *pilE* (Hill and Davies, 2009). A second crossing-over event initiates recombination between the *pilE: pilS* hybrid and the original *pilE* locus in a homologous region located downstream of the first event. Finally altered *pilE* and degraded *pilS* locus are formed by this type of recombination (Obergfell and Seifert, 2015; Hill and Davies, 2009).

In hybrid intermediate model, as in the successive half crossing-over model, half crossing-over events initiate recombination between the *pilE* and a *pilS* locus on a sister chromosome resulting in a *pilE: pilS* hybrid intermediate (Obergfell and Seifert, 2015). Then, two steps of recombination occur between the *pilE: pilS* hybrid intermediate and recipient *pilE* on different chromosomes. Recombination junctions form in homologous regions upstream of the gene and in microhomology regions in the variable regions. Finally, double Holliday junctions are resolved thereby forming altered *pilE* sequences on the recipient chromosome (Obergfell and Seifert, 2015).

HR in the Neisseria pilin locus can trigger phase variation into an OFF state (no expression of pili) by introducing a premature stop codon from a silent *pilS* copy into the *pilE* locus. A second HR event can remove the stop codon from *pilE* and restore pilin expression (Obergfell and Seifert, 2015; Segal *et al.*, 1985).

1.8.3 Mutation

Point mutations, commonly referred to as single nucleotide polymorphisms (SNPs), are changes in the DNA sequence due to a substitution of one nucleotide without or with a change in the encoded amino acid (i.e. synonymous or non-synonymous respectively) or insertion/deletion of a single nucleotide resulting in a frameshift which alters the encoded proteins (Bryant *et al.*, 2012). Frameshift mutations occur frequently in simple repeat tracts due to the formation of misalignments between the nascent and template strands during DNA synthesis (Lovett, 2004). Duplications occur when there is forward slippage of the nascent strand whereas deletions result from backward slippage (Lovett, 2004).

These mutations can arise as a result of DNA replication errors or due to damage from exogenous agents (e.g. UV light, ionizing radiation) or endogenous agents (e.g. reactive oxygen species) (Norbury and Hickson, 2001; Wright, 2000). The contributions of different factors to mutation rates and genome wide hypermutation are discussed in the next sections.

1.8.3.1 Contributions of DNA replication to mutation rates

Replication error rate is controlled by three important factors (Fijalkowska *et al.*, 2012). These are inserting the correct nucleotide by the action of DNA polymerase III, correcting the misinserted nucleotide by the action of the proof-reading sub unit of DNA polymerase III (Lovett, 2011) and correcting mismatched nucleotides through the action of DNA mismatch repair system (MMR) (Criss *et al.*, 2010; Denamur and Matic, 2006). In *E. coli*, low error rates are accomplished by the action of the three previously mentioned factors as follows; inserting the correct base contributes approximately 10^{-5} , proofreading is around 10^{-2} and DNA mismatch repair system associates with around 10^{-3}. Therefore, the overall replication error rate was estimated to be around 10^{-10} (Fijalkowska *et al.*, 2012). The transcription error rate is higher than the replication error rate in bacteria with rates of 10^{-4}-10^{-5} in *E. coli* while the translation error rate was the highest (Traverse and Ochman, 2016).

1.8.3.2 Genome-Wide Hypermutation

Hypermutation is associated with defects in the DNA replication processes such as the accuracy of base selection, proof reading of errors and MMR (Jayaraman, 2009). The core of DNA polymerase III is encoded by three genes, *dnaE* (α subunit), *dnaQ* (ε) and *holE* (θ);

mutations in these genes can cause hypermutability (Jayaraman, 2009). Mutations in *mutS*, *mutY*, *mutM*, *mutT*, and DNA adenine methyltransferase (*Dam*) lead to high transversions and can elevate mutation rates up to 200-fold (Jolivet-Gougeon *et al.*, 2011). Hypermutation is also caused by defects in recombination genes such as *recB, recC, recD* and *recA*. Stressful environmental conditions such as reactive oxygen species (ROS), antibiotics, carbon starvation and anaerobic conditions can trigger hypermutation through inactivation of DNA repair systems (Jolivet-Gougeon *et al.*, 2011; Jayaraman, 2009).

In *N. meningitidis*, the mutator phenotype was mainly generated through mutations in the mismatch repair system (Criss *et al.*, 2010). These mutations lead to increases in the phase variation rate of more than 100 fold (Richardson *et al.*, 2002), resistance to antibiotics, and higher antigenic variation (Criss *et al.*, 2010).

1.8.3.3 Contributions of deamination, alkylation, and depurination to mutation rates

Apurinic-apyrimidinic (AP) sites occur at a relatively high frequency and can cause mutations, blocks to replication forks or DNA double strand breaks (Freudenthal *et al.*, 2015). AP sites form due to exposure of the N-glycosyl bond of deoxyribonucleotides to hydrolytic attack. An AP site causes the random insertion of bases by DNA polymerases leading to transversions and transitions (Loeb and Preston, 1986).

Chemical mutagens induce DNA damage through three different mechanisms which are modifying a base on the DNA such as conversion of cytosine into uracil through oxidative deamination of the amino group, frameshift formation by intercalating agents and base analogues such as 5-bromouracil (Najafi and Pezeshki, 2014).

Environmental factors can also damage DNA. Thus, pyrimidine dimers are an effect of UV-light, free hydroxyl radicals can cause formation of 8-oxoguanosine (8-oxo-dG) while cis-thymine glycol induces deprotonation and DNA base damage (Figure 1.8). AT to GC transitions are mainly caused by cis-thymine glycol while GC to TA transversions are the main effect of 8-oxo-dG (Cheng *et al.*, 1992).

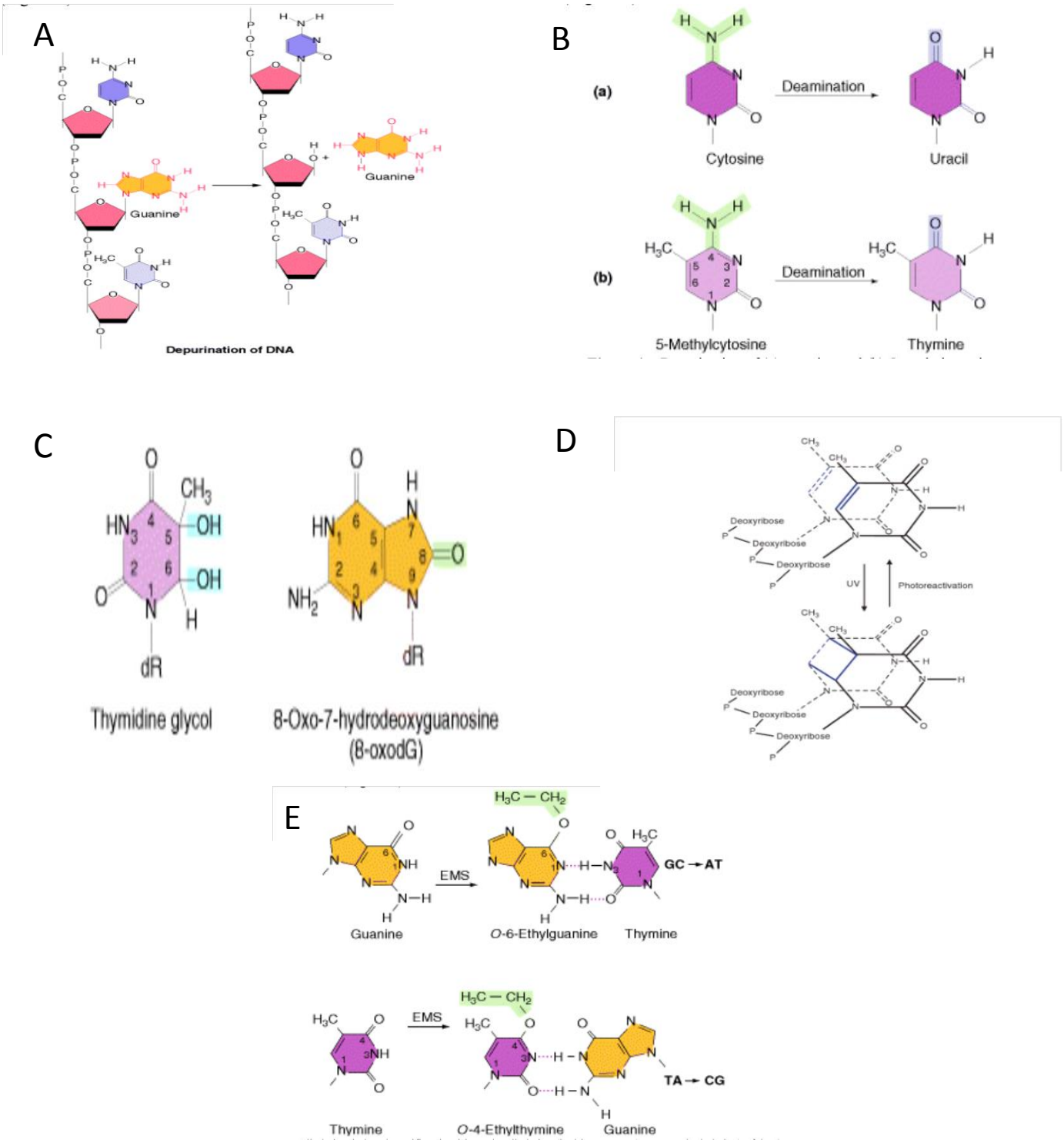


Figure 1.8: Mechanisms DNA damage. A: Depurination mediated by removing the purine residue. **B:** Deamination of cytosine into uracil and 5-methylcytosine into thymine. **C:** Free radicals mediate conversion of thymidine into hydrodeoxyguanosine. **D:** Thymine dimer mediated by UV-light. **E:** Mispairing mediate by chemical agents such as alkylation. This figure was adapted from (Najafi and Pezeshki, 2014).

N. meningitidis lacks an SOS response and so efficient base excision repair (BER) has a significant role in counteracting oxidative DNA damage (Beyene *et al.*, 2016). The BER enzymes in *N. meningitidis* consist of AP endonuclease (NApe), NExo, 3′-RPase, bi-functional glycosylase, MutM and MutY. These enzyme play a significant role in removal of damaged nucleotides and incision of the DNA backbone. The mutation of these enzymes can lead to higher replication error rates and the accumulation of heritable deleterious mutations affacting other cellular functionality (Beyene *et al.*, 2016; Nagorska *et al.*, 2012).

1.8.4 Mechanisms of phase variation

Phase variation is one of the adaptive strategies of pathogenic and commensal bacteria. This process is characterized by hyper-mutation of DNA sequences or hypervariable methylation in particular regions of the genome sequence and in a reversible manner. Phase variation mechanisms include slipped strand mispairing, site-specific recombination, homologous recombination and epigenetic modification (Bayliss *et al.*, 2008). The major mechanism utilised in Neisseria is slipped strand mispairing and so this will be the focus of this section.

Slipped strand mispairing of simple sequence repeats (SSR) during DNA replication leads to reversible changes in repeat number for tracts located in the ORF or the promoter region of a gene. Thus, the location of an SSR in an open reading frame (ORF) may lead to a frame shift mutation, whereas alterations in an SSR in a promoter may cause a change in the distance between different components of a promoter. Therefore, phase variation can lead to an abnormal or missing product or a level of expression that is higher or lower than the normal one (Metruccio *et al.*, 2009) (Figure 1.9). Studies of the mutation rates and patterns of mutation of SSR provide an insight into the amount of genetic variation generated by these repetitive sequences (Bayliss *et al.*, 2008). Mutability of the SSR is influenced by cis-acting factors such as repeat length and trans-acting factors such as DNA replication and repair factors (Moxon *et al.*, 2006; Bayliss, 2009; Bayliss *et al.*, 2001).

Many outer membrane proteins of *N. meningitidis* undergo phase variation and this may facilitate escape of the immune system. Bayliss *et al.* (2008) have shown that alterations in repeat tracts of *lgtG* were observed when *N. meningitidis* was subject to selection by the bactericidal activity of a monoclonal antibody specific for a phosphoethanolamine-

containing epitope of the LPS. The LgtG product controlled addition of a glucose blocking addition of phosphoethanolamine to this specific position in LPS. These results indicated that phase variation mediates changes in the expression of *lgtG* resulting in variants that could escape an immune response and hence adaptation to a stress condition.

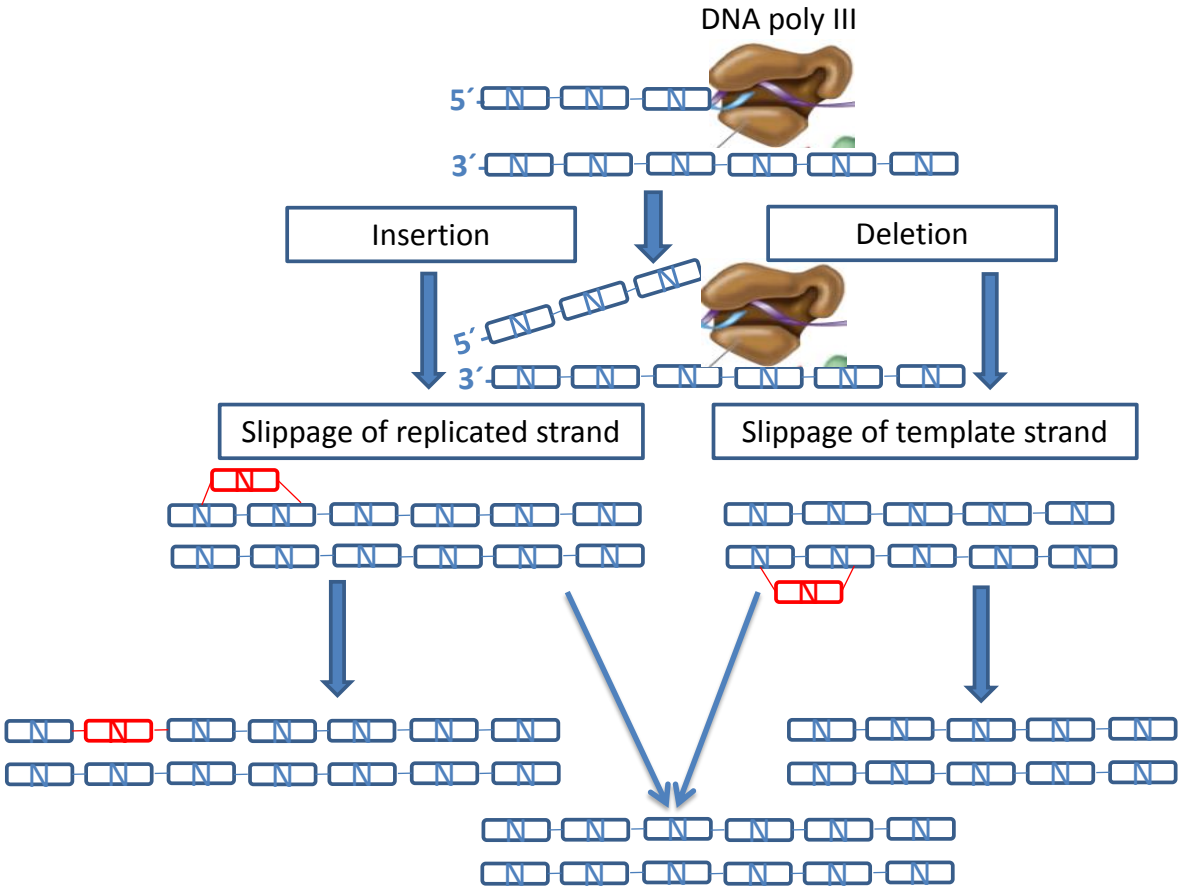


Figure 1.9: Mechanism of phase variation mediated by SSR. A gain or loss of a repeat unit is achieved during DNA replication through the process of separation and reannealing of the DNA strands, Insertion happens when the misalignment occurs on newly replicated strands and deletions happen in cases where the misalignment occurs on the template strand.

1.9 HGT as a driving force for evolution

HGT dramatically changes the shape of evolution of genes from tree-like relationships, as proposed for vertical transmission, into network-like relationship as proposed for horizontal transmission (Koonin, 2015). The mosaic gene structure of Neisseria sp. arises due to the ability of these organisms to take up the DNA by natural transformation and to recombine these molecules into the chromosome. The mosaic gene structure is highly prevalent in the house-keeping genes and the outer-membrane proteins (Stabler *et al.*, 2005).

Niehus *et al.* (2015) suggest that the rate of observed HGT is high for the traits that are influenced by the action of positive selection. Moreover, mobile genetic elements, such as plasmids and integrative conjugative elements (ICEs), enhance the rate of HGT especially for genes that introduce selective advantages for the recipient such as genes coding for virulence factors, toxin resistance genes and metabolic genes (Soucy *et al.*, 2015). It has been reported that the rate of HGT is elevated in the accessory genes as compared with core genes in Neisseria populations (Bennett *et al.,* 2010). In addition, the rate of HGT dramatically increases among closely related organisms because the frequency of HR is higher when there is high homology between the incoming DNA and the recipient genome (Soucy *et al.*, 2015).

In *N. meningitidis*, high rates of HGT are found in Minimal Mobile Elements (MMEs), islands of horizontally transferred DNA (IHTs), canonical genomic islands (Gis) and prophages (Schoen *et al.*, 2009). MMEs are defined as conserved flanking regions between which different whole gene cassettes are present. MME are distributed around the genome of *N. meningitidis* and are associated with whole genes cassettes such as the capsule biosynthetic and transport genes, *opa* genes and restriction modification systems. These cassettes are found between conserved genes such as *rfaE* and *rfaD*, *fpr* and *dinP*, *pheS* and *pheT,* and *tex* and *galE*. These cassettes exhibit high variability between lineages of *N. meningitidis* and between *N. meningitidis* and *N. gonorrhoeae,* suggesting that these cassettes are subject to high frequencies of HGT and HR (Zhu *et al.*, 1999; Bart *et al.*, 2001; Klee *et al.*, 2000; Snyder *et al.*, 2007).

Gis have associations with tRNA loci and frequently have repeats and genes or pseudogenes encoding genetic mobility functions in their flanking regions (Schoen *et al.*,

2009). The importance of Gis arises as they distinguish disease and carriage strains through pathogenicity islands (PAIs). These islands have special features such as coding for virulence proteins and only being present in pathogenic strains (Che *et al.*, 2014). Gis are rarely found in *N. meningitidis.* IHTs lack most of the properties of Gis but are similar in their G + C content and codon usage (Schoen *et al.*, 2009). Several IHTs have been found (e.g. MuMenB/PNM2, PNM1, IHT-B, IHT-C, IHT-D, IHT-E) and these islands exhibit high structural variation within the neisserial species and meningococcal strains. Many genes are associated with these islands including restriction/modification systems, a fimbrial protein precursor, putative TonB-dependent receptors, numerous uncharacterized hypothetical genes and outer-membrane proteins (Hotopp *et al.*, 2006). Most islands show exchange between *Neisseria lactamica* and *N. meningitidis* though HGT (Hotopp *et al.*, 2006), High similarities between different islands of *N. lactamica* and *N. meningitidis* give strong evidence that *N. meningitidis* arose before *N. gonorrhoeae* (Putonti *et al.*, 2013).

Prophages play a crucial role in maintenance of evolution in bacterial populations through enhancing genomic diversification (Fan *et al.*, 2016; Klimenko *et al.*, 2016). A filamentous family of prophages plays an important role in spread the virulence genes through population of bacteria and may be responsible for breaking the commensal relationship with their host and causing disease (Bille *et al.*, 2005). The TspB protein, encoded by prophage gene *ORF6*, mediates binding to IgG antibodies and formation of biofilms. This protein plays a significant role in protection of *N. meningitidis* against host immune responses through its ability to bind Ig and other immune effectors such as human factor H, serum albumin, fibrinogen human IgG, β(2)-glycoprotein I, and complement protein C3. *ORF6* is carried on a prophage that exhibits an association with invasive meningococcal strains, indicating that this prophage has a role in providing protection for some strains against the host immune system (Mller *et al.*, 2013).

1.10 Recombination as a driving force for evolution

High rates of recombination contribute to evolution of meningococcal lineages, enhancing persistence of these lineages for long periods even in immune populations (Buckee *et al.*, 2008). These high rates of recombination result in disappearance of the phylogenetic signal of species (Feil *et al.*, 2001) and increase diversification of clones with significant linkage

disequilibrium (Spratt *et al.*, 2001). It has been reported that the rate of recombination differs between meningococcal lineages. Thus, recombination rates of serogroup A strains appear to be significantly lower than other serogroups resulting in this lineage having a truly clonal population structure (Bart *et al.,* 2001). Based on MLST genes, the rate of recombination was estimated to be 8 times higher than the rate of mutation in the evolution of *N. meningitidis* lineages (Didelot *et al.,* 2009).

HR and selection drive genome variation of outer membrane proteins particularly *porA* and pilin genes (Smith *et al.,* 1995; Andrews and Gojobori, 2004). A genome comparison study carried out on two *N. meningitidis* serogroup C disease isolates with high fatality rates, showed that variation was mainly localized in genes involved in capsule biosynthesis through acquisition of sequences by HGT (Lavezzo *et al.,* 2013). Furthermore, high frequencies of HR have repeatedly been observed in many genes of *N. meningitidis* such as in the penicillin-binding-protein-2, capsule gene cluster, the RMSs and Maf3 adhesins (Bowler *et al.,* 1994; Hao, 2013; Joseph *et al.,* 2011; Feil *et al.*, 2001; Feil *et al.,* 1999; Kong *et al.,* 2013; Muzzi *et al.,* 2013; Lamelas *et al.*, 2014).

Thus, intra-species recombination is more frequent than inter-species recombination for *N. meningitidis* with the latter decreasing further for unrelated species (Linz *et al.*, 2000). However, some strains of *N. meningitidis,* known as hybrid groups, share homologous sequences that fit with more than one species (Corander *et al.,* 2011). These inter-species recombination events disturb the concept of a species and clonal population structures particularly when they occur with high frequency or involve large importation events (Linz *et al.,* 2000). These hybrid strains can contribute to vaccine resistance (Corander *et al.,* 2011). It has been reported that intra-species recombination is frequent for *tbpB*, *opa* and *IgA1* protease (Linz *et al.,* 2000; Morelli *et al.,* 1997) and that *porB2* undergoes interspecies recombination while *porB3* exhibits intra-species recombination (Urwin *et al.,* 2002).

A high frequency of variation is associated with specific loci containing large repeat arrays, indicating that these repeat arrays mediate localised recombination with externally acquired DNA (Parkhill *et al.*, 2000). A study carried out by Bilek *et al.* (2009) examined 11 *opa* genes from 14 unrelated isolates. The results showed that intragenomic recombination rather than mutation was the main cause of SNPs, insertion deletion events and other

patterns of variations such as gene duplications and mosaic structures (Bilek *et al.*, 2009).

While HR is a powerful mechanism for generating variation between closely related species and isolates, non- homologous recombination facilitates evolution through incorporation of sequences from distantly related organisms (Spratt *et al.*, 2001). Non-homologous recombination mainly mediates recombination in the accessory genome (Vos, 2009). While non-homologous recombination can induce beneficial or at least neutral effects on fitness of bacteria, some non-homologous recombination events result in random integration of foreign DNA, which increases the possibility of deleterious effects on fitness (Vos, 2009).

1.11 Mutation as a driving force for evolution

Mutation is a change in the genetic material of bacteria and transfers from one generation into another. This change leads to creating new alleles or affecting the expression profile and enhancing genetic diversity of a bacterial population (Lewis-Rogers *et al.*, 2004). Mutation hot spots are often associated the presence of direct or inverted repeat sequences where mutations produce (Lovett, 2004) structural variations such as insertions, deletions, inversions, translocations and duplications (Bryant *et al.*, 2012). Random mutations are usually deleterious or neutral rather than beneficial. The rate of accumulation of synonymous and intergenic substitutions is a measure of the basal mutation rate as these mutations are not subject to selection (Bryant *et al.*, 2012). Mutation rates are mainly determined by the balance of natural selection and the effects of genetic mechanisms of fidelity of DNA replication and repair but can be influenced by factors such as the strength of mutator alleles, bacterial population size and competition with other strains (Denamur and Matic, 2006).

Mutator phenotypes have been observed in Neisseria. Mutators are strains that have mutations resulting in inactivation of one of the cellular repair pathways controlling mutation such as MutS, MutL, and UvrD (Criss *et al.*, 2010). Mutators have higher rates of mutation than normal strains and may also exhibit increased frequencies of HGT and HR (Chopra *et al.*, 2003). High mutation rates induce high frequencies of deleterious mutations but this effect can be countered by high bacterial cell densities and selection (Chopra *et al.*, 2003). Mutators can also overcome the effects of deleterious mutations by reversion of the mutation or acquiring suppressor mutations through HGT (Denamur *et al.*, 2000). The rate

of recombination is higher in mutators due to the ability of mutators to recombine DNA sequences with low homology (Criss *et al.*, 2010). Mutational adaptation increases dramatically in populations when mutators are present; one study found that 57% of the epidemic isolates of *N. meningitidis* had elevated mutation rates (Richardson *et al.*, 2002).

1.12 Role of intergenic regions (IGRs) in changing the shape of evolution for bacterial populations of *N. meningitidis*

An IGR may contain two promoters in cases of head-to-head orientation between two genes and one promoter in cases of head to tail orientation between two genes (Hughes and Friedman, 2004). Congruent evolution has resulted in evolution of operonic regions while non-coding regions evolve on the principle of the minimization of the existence of non-functional DNA (Rogozin *et al.*, 2002). The presence of different repeats is considered as the largest source for diversification and evolution of intergenic regions and these sequences can play an important role in transcriptional activation or rewiring in prokaryotes (Matus-Garcia *et al.*, 2012). Additionally non-coding RNAs are subject to stronger selective pressures than promoters or rho-independent terminators (Thorpe *et al.*, 2016).

Recombination in IGRs is a crucial phenomenon in evolutionary terms as these regions reflect intrastrain differences in gene regulation. It has been observed that different types of selection such as purifying selection and positive selection act on promoter regions for the adaption of microorganisms to environmental stress conditions (Hughes and Friedman, 2004). Indeed selection is detected in intergenic regions even when regulatory elements are absence, this indicates the crucial role for IGRs in evolution of bacteria (Thorpe *et al.*, 2016). HGT can mediate acquisition of intergenic regions resulting in rapid changing or rewiring of regulatory networks such as movement of promoters into silent genes (Oren *et al.*, 2014; Matus-Garcia *et al.*, 2012). In a study carried out on *E. coli*, Oren *et al.* (2014) showed that 11% of the regulatory regions were subject to HGT and showed incongruence in phylogenetic trees as compared with trees of adjacent genes, which emphasizes the importance of bacterial regulatory networks.

In a study of genetic diversity in intergenic regions of meningcoccal strains, seven housekeeping genes never showed structural variation, such as deletions or frameshifts, however, their intergenic regions frequently exhibited insertion/deletion, genetic

replacement and transposition particularly for *opcA* and *opcB*.This indicates that Neisseria sp. possess highly dynamic IGRs (Zhu *et al.*, 1999).

1.13 Host-strain co-evolution

A crucial phenomenon in the biology of *N. meningitidis* is within-host evolution. This process generates phenotypic diversity by phase shifting in contingency loci, antigenic variation by recombination between silent loci with expressed loci, and localized recombination in repeat arrays (Meyers *et al.*, 2003; Schoen *et al.*, 2009). Evolution of virulence factors is a potentially important aspect of within-host evolution, however as these factors can increase fatalities and so result in evolutionary dead-ends, these factors must evolve for maintaining host or population colonisation rather than increasing their hosts morbidity or mortality (Meyers *et al.*, 2003; Schoen *et al.*, 2009). The next few paragraphs provide some examples of within-host evolution.

The capsule decreases the affinity of binding between pili and epithelial cells; therefore, phase variation in capsule expression may modulate binding through generation of acapsulate phenotypes during the early stages of infection (Hill *et al.*, 2010). Similarly, antigenic variation of pilin in *N. meningitidis* enhances variation of the pilus and so may mediate escape of immune responses (Obergfell and Seifert, 2015).

The presence of short repeats within the cassettes encoding alternative C termini of proteins may also trigger phenotypic changes. This mechanism of variation is found in *fhaB* encoding putative haemaglutinins and *mafB*, encoding adhesin proetins (Bentley *et al.*, 2007). The presence of RTX islands with three types (I, II, III) within the cassettes encoding alternative N termini of protein is another mechanism of variation observed in the *frpA/C* genes encoding iron-regulated type I secretion systems (Schoen *et al.*, 2009). Changes in these regions may be frequent during host persistence. Genome sequencing also aids to understand the within host evolution through the following examples. A study carried out to investigate the situation of phase variation during persistent carriage for 21 carriers representative of 1 to 6 months carriage. The result showed that the repeat tract was altered in two genes, *fetA* and *nadA* resulting in significant reductions in expression on their proteins during persistent carriage of some strains. The variant-specific PorA IgG antibodies, capsular group Y IgG antibodies and serum

bactericidal activity were used to investigate the immune responses in these carriers. The final conclusion was that localised hypermutation has evolved to facilitate host persistence through reduction in the expression of phase variation genes due to continuous exposure to antibody-mediated selection (Alamro *et al.*, 2014).

Several studies have showen that there was a replacement of ST5 to ST7 after 1990s in China and the African meningitis belt. Therefore, whole genome sequencing was carried out on strain NMA510612 belonging to serogroup A ST7 that was isolated from a patient in China. The genome comparison study showed that variation in this strain occurred in the type IV pilus and type I restriction enzymes. This variation in the ST7 was thought to have helped this clone to cope with the immune response established against ST5 strain in the population (Zhang *et al.*, 2014).

1. 14 Bioinformatics approaches for investigating genetic variation

Many bioinformatics approaches are used to predict HGT, recombination and mutation among *N. meningitidis* species. Compositional methods are based on searches for abnormal sequence composition compared to the rest of the genome to predict HGT. There is another method, which uses single transfer events to trace the genome location of putative HGTs in order to predict HGT (Azad and Lawrence, 2007). Phylogenetic methods include the search for incongruence between gene and the related species tree (Langille and Brinkman, 2009). The RDP program is the most widely used bioinformatics approach for recombination analysis. The identification of recombination sites is based on the three important steps: firstly ignoring those sequences not providing information on recombination sites in a phylogenetic tree, secondly comparing the position of different sequences in a UPGAMA dendrogram, thirdly calculation of average percentage identity (Martin *et al*., 2015). Recombination Analysis Tool (RAT) is a java-based application that depends on DNA and protein multiple alignment sequences (Etherington *et al.*, 2005). The Recco is another crucial program that can be used to identify recombination in sequence alignments (Maydt and Lengauer, 2006). Recently, ClonalFrameML was produced and works on the principle of maximum likelihood method for inferring recombination patterns in phylogenetic analyses within hundreds of bacterial whole genomes. Recombination parameters are then measured by Bayesian approach (Didelot and Wilson, 2015).

The Bacterial Isolate Genome Sequence Database (BIGSDb) is a tool that can be used to identify the variability among unlimited numbers of isolates and genes. It is useful for studying the population genetics of various strains through alignments of combinations of genes. Furthermore, it can also be used to identify variability of two genomes from two species by comparing their sequences pairwise enabling detection of point mutations, duplications, insertions, and deletions (Jolley and Maiden, 2010).

The study of diversity among a pool of strains (diversity among population) can be explored through various mathematical models. The principle behind measuring the diversity among different sequences involves estimating their evolutionary distance. An estimation of evolutionary distance is important as it gives clear evidence for substitutions occurring in genome sequences and phylogenetic trees of different strains. The various bioinformatics tools allow us to estimate variability and diversity of DNA or amino acid sequences and so identify types of selection such as purifying selection, positive selection and neutrality (Kosakovsky *et al.,* 2009; Sohpal *et al.,* 2011).

Finally, advances in the bioinformatics tools, next generation sequencing (NGS), computation speed and memory storage capability of computer are helping researchers to better understand the relationships such as host – strain interactions especially understanding the potential for within host evolution (Zhang *et al.*, 2014; Bansal, 2005).

1.15 Aims of the project

Advances in high-throughput sequencing technologies and bioinformatics tools have aided understanding of biology, diversity, evolution, and virulence in bacterial pathogens. Many comparative studies of *N. meningitidis* genomes have been performed and these studies are helping with discovery of the processes of host-Neisseria interactions and in preventing invasive meningococcal disease (IMD) (Zheng *et al.*, 2016; Gasparini *et al.*, 2015). One idea is that within-host evolution (WHE) may correlate with virulence of meningcoccal isolates but this is controversial. WHE is normally thought to enhance fitness of meningcoccal isolates in the nasopharynx rather than enhancing pathogenicity. Random mutations occurring in populations of meningcoccal isolates in the nasopharynx may result in increases in pathogenicity through selection for isolates capable of crossing the nasopharyngeal barrier. Klughammer *et al.,* (2017) have shown that virulence did not correlate with selection of invasive phenotypes during disease but correlated with coincidental evolution of invasive variants during carriage, however the number of isolates with four strains pairs used in this study was limited. Thus, the first aim of this project was to investigate the extent of within host evolution of *N. meningitidis* strains in carriers and to understand the dynamics and the spread of genetic traits in the meningococcal population in order to generate important information for vaccine design and immunization program with the general aim of reducing the occurrence of IMD. This general aim was approached through the following specific objectives;

1 – Analysis of genetic variation, comprising both de novo mutation and recombination, in a large meningococcal population from one carrier and paired meningococcal isolates from multiple carriers representing persistent carriage of up to six months for the purpose of finding the patterns and inferring the dynamics of variation associated with persistence of meningococci within the human host.

2 – Analysis of hypermutation mediated by CEs in meningococcal isolates representative of persistent carriage for the purpose of finding whether variation in CE patterns facilitate persistence of meningococci within the human host.

3 – Analysis of hypermutation mediated by SSRs for five phase variable genes from 15 isolates of the fourth time point of three carriers and comparison to data to for earlier time

points (as already detected by Alamro *et al.,* 2014), for the purpose of understanding how SSR patterns correlate with persistence of isolates within the host.

Many comparative genome studies have indicated that recombination in Neisseria can occur between species and between clonal complexes using strains from different geographical regions and years of isolation. For the second aim of this project, meningococcal disease and carriage isolates of CC-174 were used in order to examine micoevolution within a single clonal complex.

Chapter 2: Materials and Methods

2.1 Description of isolates

A comparative genome analysis was performed on next generation sequencing data (NGS) for three different groups of meningococcal isolates. This section describes the isolates while section (2.2) describes the type and source of the NGS. The first group consists of 40 meningococcal isolates collected from one asymptomatic carrier (V59) with ten isolates collected at four different time points: 0, 1, 3 and 6 months (with the latter three representing at least 1-6 months host persistence; the first time point was in November 2008 and hence the later three time points collected in December, February and May 2009) (See Table 3.1). The potential for within host strain microevolution was investigated using these isolates (Chapter 3). The second isolate group consists of 25 pairs of isolates of *N. meningitidis*. These 50 isolates were collected from 25 different asymptomatic carriers at two different time points (See Table 4.1). These isolates were used to study within host strain evolution of this species (Chapter 4). These two sets of isolates are a sub-set of isolates taken from a carriage study performed on 190 students from Nottingham University between November 2008 and May 2009 (Bidmos *et al.*, 2011). The samples were collected by taking a nasopharyngeal swab and spreading it on one half of a chocolate GC selective agar plate. In order to obtain multiple single colonies, the inoculum was spread with a loop on the other half of the selective agar plate. The plate was incubated at 37 ^{o}C to allow growth of *N. meningitidis* colonies. Single colonies were restreaked onto chocolate agar plates (Oxoid) and incubated overnight under the same conditions (Bidmos *et al.*, 2011).

The third group of isolates consists of 7 disease and 18 carriage isolates from CC-174. The carriage isolates were isolated from 18 different asymptomatic carriers (All Nottingham University students) with 17 collected between 2008 and 2011 (Bidmos *et al.*, 2011; Oldfield *et al.*, 2016) while an isolate NO0011039 (ID: 27497) was collected in 2000 (Oldfield *et al.*, 2016) (See Table 5.1). These isolates were previously shown to form a tight cluster (Oldfield *et al.*, 2016) and therefore, could be used to investigate microevolution focusing on recombination process as the main driving force, described in (Chapter 5).

2.2 Description the type and source of NGS of isolates

Genomic DNA was extracted from one isolate per carrier per time point using DNeasy purification kit (Qiagen) and stored at 4ºC for subsequent analysis. The *porA*, *fetA* and seven house-keeping genes were amplified and sequenced, then the sequencing products were further used to query the database (http://pubmlst.org/neisseria/) for detecting the typing method of each isolate. In the first group of isolates, all the 40 isolates had an identical serogroup, *PorA* type and ST type, in the second group of isolates (25 pair isolates), each pair of isolates had an identical serogroup, *PorA* type and ST type (Bidmos *et al.*, 2011).

In collaboration with Prof. Martin Maiden (University of Oxford), genome sequence data was generated by Illumina Hiseq and was assembled using Velvet (version 1.1) (Zerbino, 2010). Assembled sequence data was loaded into the pubMLST.org/neisseria database powered by the BIGSdb genomics platform (Bayliss, Harrison and Maiden, unpublished data). Prokka (Seemann, 2014) was used for further annotation of representative genomes (performed by Dr. M. Blades, BBASH, and University of Leicester). Sequences in BIGSdb range from partially assembled genomes to completely assembled genomes. The genome sequences of the all groups of isolates were partial and were contained in multiple contigs.

Initial studies were performed using the MLST database and commands contained in the BIGSdb platform. Due to concerns with the Velvet assembly process, most of NGS data was downloaded from the database and reassembled using SPAdes software (Bankevich *et al.*, 2012).

2.3 Checking the quality of the NGS

The quality of the paired end data set was checked using the Basic Statistics, Per Base Sequence Quality, Per Sequence Quality Scores and Overrepresented Sequence modules in the FastQC software program (Andrews, 2010). The Overrepresented Sequence module was used to check the forward and reverse paired end data to determine whether significant proportion of the reads were contaminated with adaptor sequences. This analysis showed that less than 1% was contaminated and therefore the Trimmomatic script (Bolger *et al.*, 2014) was not used to remove adaptor sequences.

The Per Base Sequence Quality and Per Sequence Quality Scores modules showed that in general, the quality of data was good for the forward and reverse paired end data with mean quality scores of > 35 (Figures 2.1, 2.2). The Basic Statistics modules hosted in the FastQC software program, however, showed that there could be some issues with data quality, especially in respect to contamination with forward and reverse single-end data (Appendix 1). Therefore, the subsequent assembly was undertaken using only forward and reverse paired end data.

The average of depth of coverage of reads data for whole genomes under the study was checked using Bedtools (Aaron and Ira, 2010). Bedtools used bam files that were downloaded from BIGSdb and the following command: bedtools genomecov -ibam bam files > output. The average of depth of coverage was 148X, 252X and 143X for the 40, 25 pairs and 25 CC-174 strains and that is considered as a good parameter for the quality of sequencing using NGS.

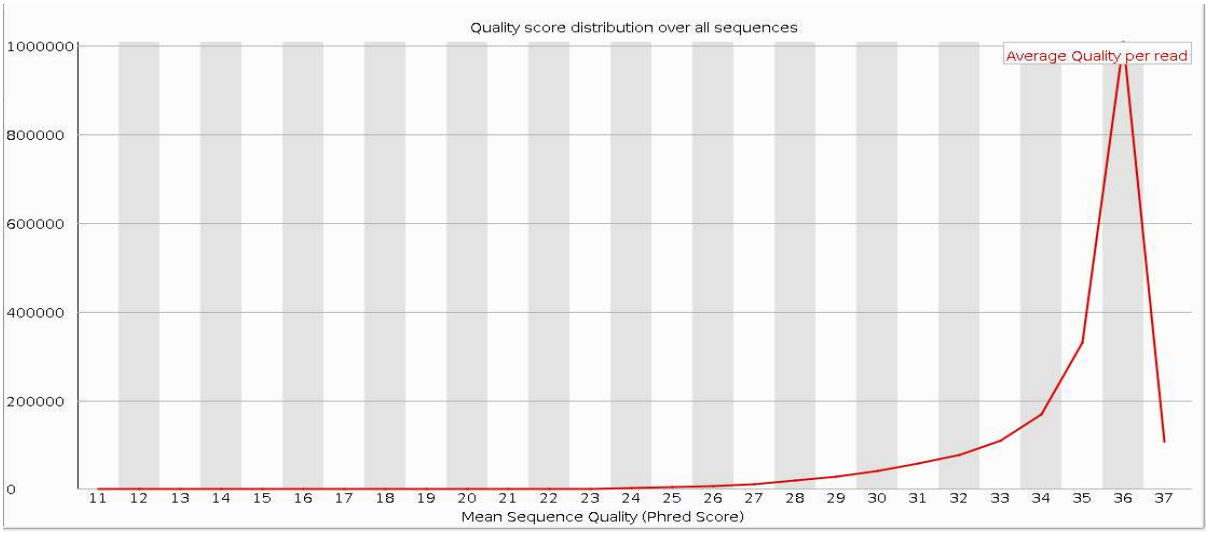


Figure 2.1: Per Sequence Quality Scores modules for the forward and reverse paired end data. The data in this figure is shown as an example for the N59.1 isolate obtained by Sequence Quality Scores module hosted in the FastQC software.

Figure 2.2: Per Base Sequence Quality modules for the forward and reverse paired end data. The data in this figure is shown as an example for the N59.1 isolate obtained by Per Base Sequence Quality module hosted in the FastQC software.

2.3.1 De-novo Genome Assembly using SPAdes

As the forward and reverse paired end data was contaminated with forward and reverse single end data, the Trimmomatic script was run to separate the paired end data from single end data. The following parameters were used (Table 2.1).

Table 2.1: Parameters used for separating paired end data from single end data using Trimmomatic script.

Phred33	Quality scores are 33 offset
"LEADING: 3"	Trim 5 prime bases with quality score <3
TRAILING: 3"	Trim 3 prime bases with quality score <3
MINLEN	Delete reads trimmed below length 36 (MINLEN)

The SPAdes software was used to reassemble the paired end data without using a reference sequence using the following parameters (Table 2.2).

Table 2.2: Parameters used to reassembling forward and reverse paired end data using SPAdes script.

--pe1-1	Name of fastq (forward reads)
--pe1-2	Name of fastq (reverse reads)
--careful	Reducing mismatches and indels
--cov-cutoff	The read coverage cutoff was set to auto
-k	Variable Kmer size
-o	Name of the directory (output)

The SPAdes software uses the De Bruijn graph algorithm for the purpose of reads assembly. The De Bruijn graph algorithm works on the principle of combinatorial mathematics therefore, the sequence reads split into short words (K-mers, that is short substring of read) with different sizes. The algorithm generates overlaps between different words to create the assembly. The software uses different K-mer sizes to optimize the assembly with best solution. Two output files were obtained: - contigs.fasta (multifasta file contains the name and sequence of all contigs) and scaffold.fasta (fasta file contains identifier and whole genome sequence) (Bankevich *et al.*, 2012).

2.3.2 Assembly assessment using QUAST

The Quality Assessment Tool for Genome Assemblies, part of the QUAST software (Gurevich *et al.*, 2013) was used to assess the SPAdes assembly. This tool uses the contigs.fasta and scaffold.fasta files and produces a number of metrics. The important criteria within QUAST are N50 and L50. N50 is the length of the contigs that form half (50%) of the bases of the assembly while L50 is the minimum number of contigs which form half (50%) of the bases of the assembly (Gurevich *et al.*, 2013). The total sequence length using the paired and single end reads, the percentage of removed bases, and the N50 and L50 for some isolates are shown in (Appendix 1). The new assembly for most isolates shows an increase in the value of N50 and decreasing the L50 values as compared with the old assembly, which is a good indicator that the new assembly is more accurate than the old assembly.

2.3.3 Genome assembly visualization using IGV and checking of variable genes

The paired end reads were mapped back to the assembly constructed using only paired end reads. The IGV program (Thorvaldsdttir *et al.*, 2013) was used to visualize the alignment. This was achieved in three steps: - first, an index of the assembly was created so that the Burrows-Wheeler Aligner (bwa) (Li and Durbin, 2009) could quickly access the assembly for mapping using the following command; *bwa index –a is assembly.fasta*. The assembly.fasta refers to the product file of SPAdes assembly while the (–a is) the parameter used to refer to the algorithm employed in the analysis; second, BWA was used to align the reads back to the indexed assembly using the following parameters (/left_trim.fastq: name of forward reads of fastq file, /right_trim.fastq: name of reverse reads of fastq file and >align_assembly.sam: name of output file); third, the BWA output file was converted into a file (binary alignment map file) then sorted and indexed using samtools commands; *samtools view –bus align_assembly.sam > align_assembly.bam*, *samtools sort align_assembly.bam > align_assembly_sorted* and *samtools index align_assembly.sorted.bam*.

Finally, the assembly.fasta file and align_assembly.sorted.bam were loaded into IGV software. The IGV window showed two parts; first, the coverage track displayed the depth of reads at each locus; and second, the alignment track displayed how reads mapped to the reference sequence (Figure 2.3).

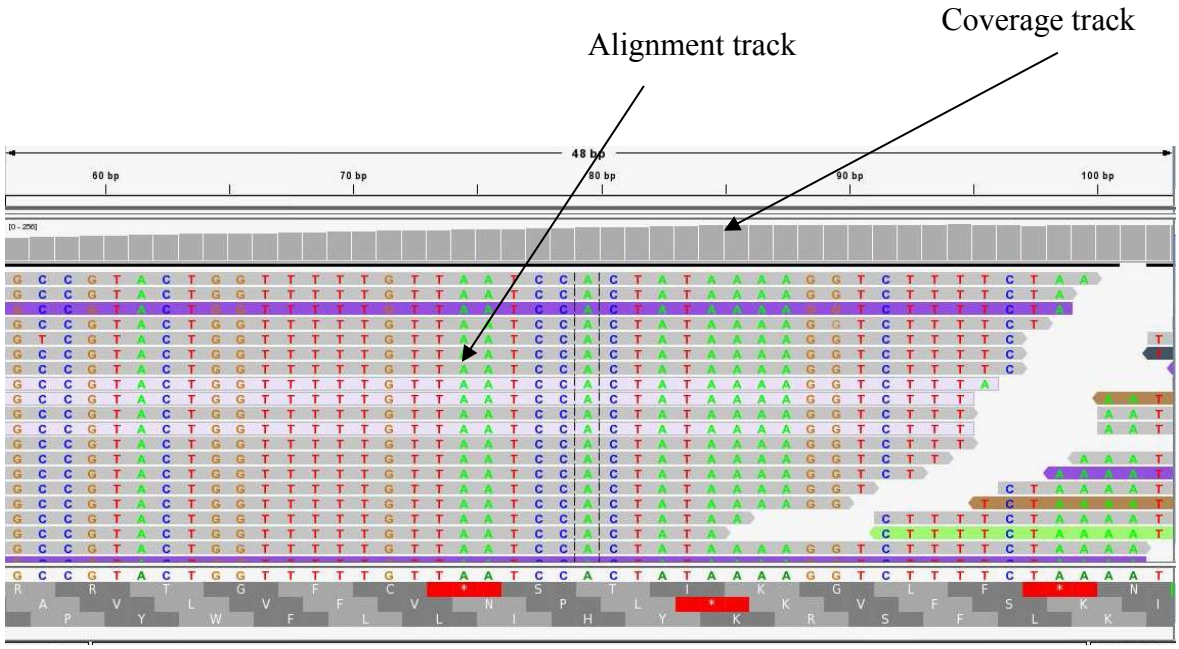


Figure 2.3: IGV window. The coverage track and the alignment track are shown for *NMC0202* of isolate N438.1. In this an example, all the reads were identical (not mix) with varied nucleotides and a very high base phred quality was observed.

Each variable gene from the old assemblies (as a query) was checked against the new assemblies (as a subject) using BLAST. From the BLAST result, the positions of each SNPs and indel were detected in the old assemblies from genome comparison among isolates were checked in the new assemblies. These SNPs and indel were not present in the new assemblies (means there is no identity between old and new assembly in the position of SNPs or indel), it was assumed to indicate that the variation was in a single end read in an old assembly and it was therefore removed.

Some variable genes had variation in the new assembly but were not real due to poor assembly as observed when checked with the IGV software. The *NMB1561* variable gene used as an example contains a mix of reads with varied nucleotides and a very low base phred quality (10).

2.4 Scripts with their general purposes for manipulation of DNA sequence

I have written different scripts that were used for different purposes in the current project, as listed in Table 2.3.

Table 2.3: List of scripts used in the current study with their aims.

No	Script	Description
1	extract_IGR.pl	Extract IGR from embl format file
2	Identify_var_IGR.pl	Extract the varied sequence from paired isolates and align them
3	align_format.pl	Insert the gene name in the header of alignment in each line of alignment format
4	fasta_format.pl	Convert the alignment format into a multifasta format
5	extr_gen_isolate.pl	Separate variable genes for a particular isolate of CC-174 into gene specific file
6	identify_nu_copy.pl	Identify variable genes with more than one copy in a genome
7	check_omittedgene.pl	Identify variable genes listed in appendix 9
8	extract_CD.pl	Extract genic region from embl format
9	allele_comp.pl	Detect if the variable genes identified by allele comparator had already been picked up by genome comparator
10	concate_embl.pl	Concatenate whole sequence from EMBL formats of each isolate
11	ext_seq.pl	Extract sequence of varied genes detected by allele comparator for each isolate
12	identify3_orgi.pl	Extract the varied sequence from paired isolates and reference genome and align them
13	final_gap_remov.pl	Identify gaps at the beginning and end of alignments and remove them
14	replace_fullseq.pl	Editing the reference genome to include the variable nucleotides within its sequence
15	extr_contg.pl	Extract the starting and ending points of all the contigs for each isolate of CC-174
16	extractespoint.pl	Extract the recombinant fragments from output of ClonalFrameML program
17	extractespoint2.pl	Extract the mutant fragments from output of ClonalFrameML program
18	extr.fasta.pl	Concatenate all the recombinant or mutant fragments of the isolates CC-174
19	index_var.pl	Extract the position of each varied base pair between two different files
20	sliding_win.pl	Inferring recombination patterns on probability of occurrence of SNPs density within sliding window
21	extract_aa.pl	Extraction amino acids from GenBank format MC58 genome sequence of sequences of each reading frame
22	Significant_CE.pl	Identify the presence of CEs within the IGRs and the starting and ending points of CEs pattern
23	variation_location.pl	Extract the index of variables from sequence alignment files
24	locate_CE.pl	Identify the presence of varied SNPs in the CEs patterns for the DNA sequence of the compared IGRs

2.5 Identification the genome variation of the isolates under the analysis

In general, genome variation was detected in the genic and IGRs of isolates under the study as follows:-

2.5.1 Methods detecting variation in the genic regions

Two approaches hosted in BIGSdb were used to detect the genome variation in the genic regions as follows:-

2.5.1.1 Genome comparator (GC) method for identification of variable loci

The comparative genome analysis was carried out using the GC (Jolley and Maiden, 2010) option on BIGSdb with MC58 (NC_003112), FAM18 (NC_008767) and N59.1 (a partial annotated genome provided by Dr. Bayliss) as the reference genomes. The default setting of the program was used with the minimum % identity was 70, the minimum % alignment was 50, the core threshold % was 90, and truncated loci were included in the analysis. The loci in the output files were annotated with the following codes; T, for truncated; X, for absent, and numbers for different alleles. The program also produces an alignment for each locus as well as phylogenetic trees.

The program assigned specific number for each allele by which allele belongs to reference genome assigned one for all variable loci. For each compared isolate, alleles similar to the reference genome were assigned as allele 1 but alleles different from the reference genome and different from each other were assigned 2, 3, 4 and so on.

2.5.1.2 Allelic comparator (AC) method for identification of variable loci

Another method to search for gene variability in BIGSdb is searching for allele variation (Jolley and Maiden, 2010). The default for the program was to include all the loci from the selected scheme (a total of 1437) and to export the allele numbers for each particular locus. The program compared all loci in the target isolates against each other, against alleles derived from multiple genomes hosted in the BIGSdb, and identified variable genes. The program assigned a specific number for each allele with identical allele's being assigned the same number while different alleles were assigned different numbers. The variable loci identified by the AC and GC methods were compared using Basic Local Alignment Sequence Tool (BLAST) (Altschul *et al.*, 1990).

2.5.2 Filtration process of the variable loci

Filtration of the data to exclude the spurious variable genes was achieved depending on different criteria. The first process of filtration was to exclude the loci that had an X signal in most of the isolates, since this indicates that the gene was absent. Filtration was also applied to the loci that have more than one copy in the genome (termed paralogous loci). A BLAST search was used to check if the variable loci had more than one copy using the MC58 genome sequence as a reference. If a particular variable locus had more than one gene with an alignment length greater than 85 and sequence identity greater than 90%, then the gene was considered to be a paralogous locus and excluded from the variable loci. The alignments for the variable loci were checked manually and loci that were variable due to a gap at the start or end of the alignment (Figure 2.4) or due to an SSR were removed. The loci that had a truncated allele in multiple isolates or consisted of a short sequence of less than 100 base pairs were also removed. Moreover, CEs were removed.

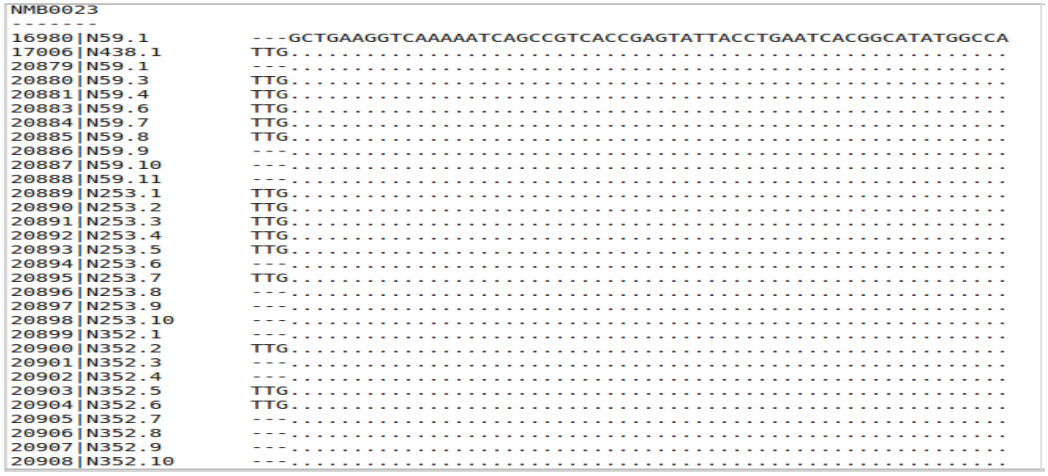


Figure 2.4: An example of a variable gene with gaps at the start of the alignment. The data in this figure is shown as an example for *NMB0023* gene, the sequence for the first isolate N59.1 has gaps due misalignment of (TTG).

2.5.3 Perl scripts for extracting and manipulating IGRs

Perl script extract_IGR.pl (Appendix 2) was used to extract an index of the starting and ending points of each coding region (CD) in the forward and reverse strands, with a full sequence and gene names, and then saving them in an array. Then the program was able to subtract the start point of the second CD from the end points of the first CD resulting in a list of all the IGRs for all coding sequences CDs in a particular sequence (genome or contig). The program was able to extract all the IGRs from multiple files in a directory. This script extracts all the IGRs of each persistent isolate and saves them in separate files for each isolate. The script distinguishes between IGRs located between head to head of adjacent genes or IGRs located between head to tail of adjacent genes. The script considers reverse complement for the DNA sequence of IGRs that are located on the reverse strand, to be in the same orientation with the DNA sequence of IGRs that are located on forward strand.

For the 40 isolates, the DNA sequence of all the IGRs were extracted from all the embl files of isolate N59.1, then the DNA sequence of each IGR was used for a BLAST search (hosted in the BIGSdb) to extract the DNA sequence of all the IGRs of the other 39 isolates.

For the 25 pair isolates, the identify_var_IGR.pl script (Appendix 3) was written to compare the DNA sequences for all IGRs from each pair isolates at a particular time point. The script was used to assign the ID and the sequence of each IGR in a key-value pair of hash. Then, the comparison was carried out by looping through the key of each hash (which represents the header of each IGR), such that if the keys were found in the pair isolates, then their values (which represent the sequence of each IGR) were compared. Those values found to be variable in each pair isolates, were printed out in one file with their keys. In this way, the number of output files depended on the number of variable IGRs between the two compared isolates. Finally, the script also used the MUSCLE program (Edgar, 2004) to align the two IGR sequences and printed these alignments into another file. The parameters of filtration described in section (2.5.2) were also applied to the variable IGRs detected by these scripts.

2.5.4 Building full genome sequence carrying the real variation in each contig

The reason for designing and using this method was to exploit the advantages of gene-by-gene comparison approaches in BIGSdb. The AC method utilises the system of classification of multiple genes within BIGSDb (See section 2.5.1.2) while the GC method involves comparison to a reference genome (See section 2.5.1.1). The GC method detects all variable SNPs that can be picked up through genome comparison to a small number of reference genomes. The AC method is used in addition to GC in order to detect SNPs in genes that are lacking in the reference genomes and hence would not be picked up by the GC analysis. Analysis of the outputs of both methods indicated that the genome sequences of individual isolates contained spurious SNPs. Thus, the 'real' variable SNPs had to be mapped into a reference genome in order to avoid the analysis of spurious SNPs during detection of recombination. Hence, Perl scripts were written and used to create artificial genomes that carried only the 'real' SNPs detected using the AC and GC methods. These scripts replace the existing sequences within the reference genome (isolate ID 27497 was used as the reference genome) with these SNPs.

All the contigs containing real variation (SNPs) for the 25 CC-174 isolates were constructed as follows:-

2.5.4.1 Detection SNPs in genic regions of 25 CC-174 isolates using BIGSdb and other scripts

The GC program hosted in BIGSdb was run to detect variation within the genic regions of multiple CC-174 isolates. A series of Perl scripts were written to automate filtration of spurious SNPs in these variable genes.

Firstly align_format.pl script (Appendix 4) uses the genic alignment files produced by BIGSdb as input and inserts the gene name in the header of an alignment into each line (header: the strain name, gene name) then prints them with DNA sequence. Later, the fasta_format.pl script (Appendix 5) converts the alignment format in previous step into a multifasta format (variable genes in all the isolates of CC-174 in one file). Then extr_gen_isolate.pl script (Appendix 6) extracts all the variable genes for a particular isolate of CC-174 into one file.

Secondly, the identify_nu_copy.pl script (Appendix 7) was used to identify variable genes with more than one copy in the genome. This script takes the sequences of variable genes and performs a BLAST search against MC58 genome. Sequences with more than one hit with an 85% length of alignment and 90% identity were considered as genes with multiple copies and were printed to an output file. Another script, check_omittedgene.pl (Appendix 8) was written to identify genes listed in (Appendix 9). This script concatenates the sequences of these 46 genes and performs a BLAST search against the multifasta file of the variable genes. Again, a hit with an 85% length of alignment and 90% identity was filtered and printed to an output file (Figure 2.5).

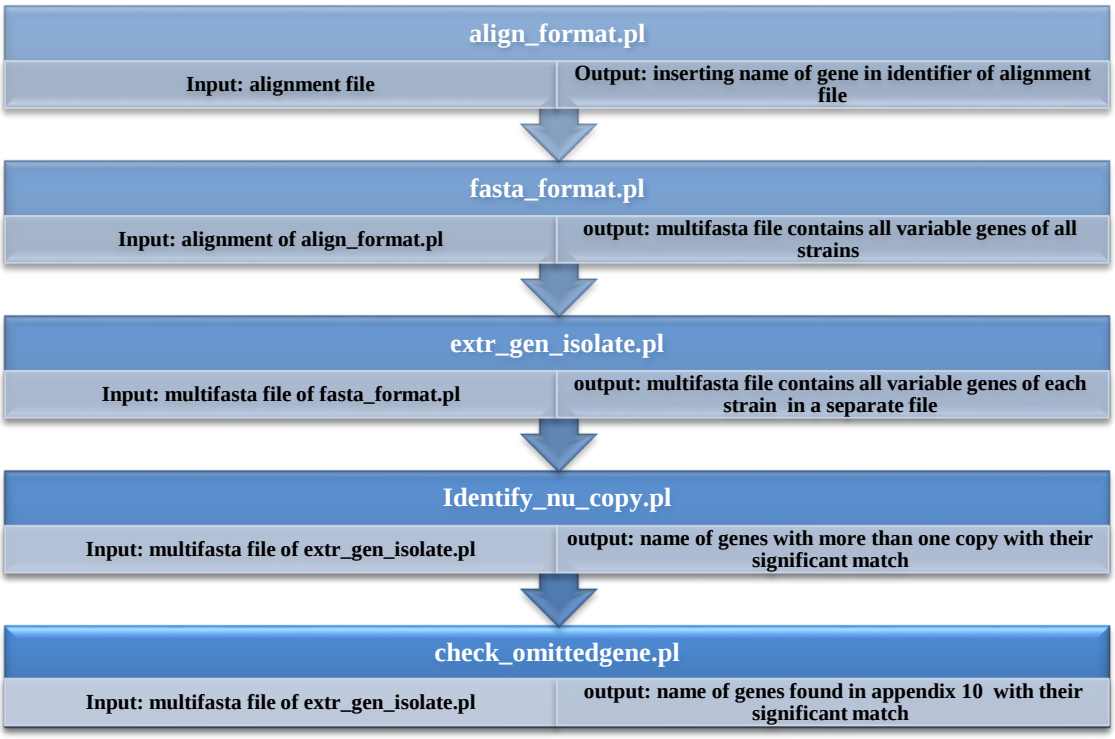


Figure 2.5: Flow diagram showing the Perl scripts for extraction and filtering of variable genes. These Perl scripts were used in a pipeline to filter the variable genes detected using GC software in BIGSdb.

The AC program hosted in BIGSdb was also run to detect variation within the genic regions of CC-174 isolates. These genes were subject to the same filtration processes as described above. Firstly, the AC program identifies only the name of varied genes, therefore, the sequence of each varied gene was extracted as following; the extract_CD.pl script

(Appendix 10) extracts all genic regions of isolate ID: 27497 as a reference genome using the same principle of extract_IGR.pl script but it extracts CD instead of IGRs. Then, samtools faidx command extracts only the sequence of the variable genes in fasta format detected by AC program for isolate ID: 27497 from file produced by extract_CD.pl script.

A Perl script allele_comp.pl (Appendix 11) which is modified version of check_omittedgene.pl script was used to detect if the variable genes identified by AC had already been picked up by GC. This script concatenates the sequences of variable genes detected by GC and then uses BLAST search to compare against genes identified by AC. The genes detected by both methods were identified as those with an 85% length of alignment and 90% identity.

To extract sequences of variable genes that were detected by AC in each isolate of CC-174 in multifasta file, different steps were carried out; the concate_embl.pl script (Appendix 12) concatenates the sequence from all EMBL files for each CC-174 isolate and prints them in fasta format (header; name of isolate, concatenated sequence of all embl files). Then ext_seq.pl, script (Appendix 13) extracts the sequences of varied genes detected by AC for all isolates of CC-174 (Figure 2.6).

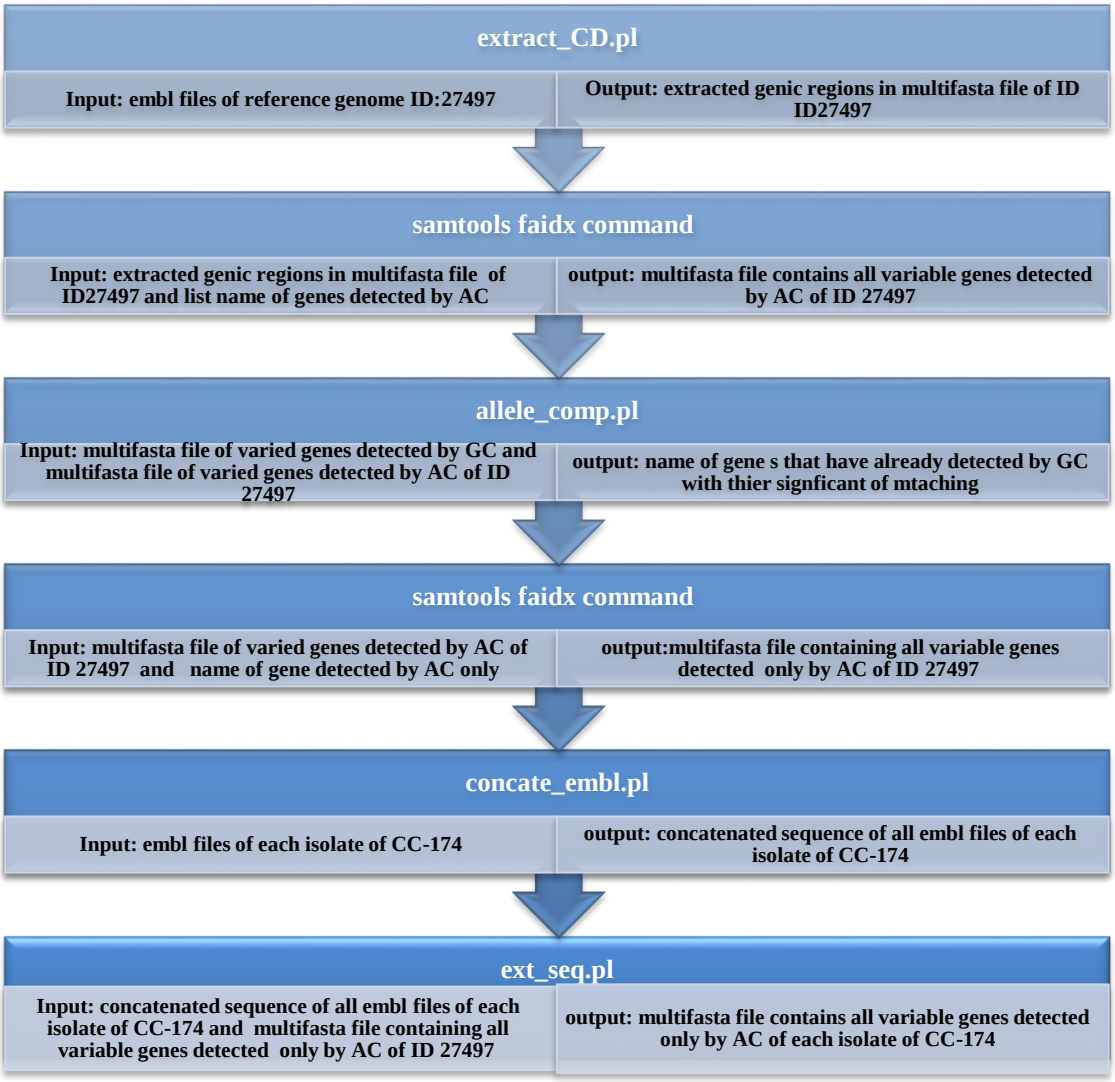


Figure 2.6: Flow diagram showing the Perl scripts used to filter the variable genes detected by AC. These Perl scripts were used in a pipeline to filter the genes detected using AC. Filtration included removing genes detected by GC sequences of each gene for multiple isolates.

2.5.4.2 Detection SNPs in IGRs of 25 CC-174 isolates using some scripts

The extract_IGR.pl script (See 2.5.3 section) was used to extract IGRs from the EMBL formatted contigs of each CC-174 isolate. Then, identify3_orgi.pl script (Appendix 14) was used to compare the sequences with the same identifier (gene name) from reference genome ID: 27497 and the genome sequences of other CC-174 isolates. Finally, the alignments of each variable gene were printed to a particular file.

The filtering of these IGRs was achieved with a series of scripts. The final_gap_remov.pl script (Appendix 15) was used to identify gaps at the beginning, or end of alignments, and remove these files from the list of variable regions.

As mentioned previously, other scripts align_format.pl, fasta_format.pl and extr_gen_isolate.pl scripts (See section 2.5.4.1) were used to produce a final output of all the variable IGRs of each CC-174 isolate in a separate multifasta file. These files were checked for the presence of CE pattern using significant_CE.pl, variation_location.pl and locate_CE.pl scripts (See section 2.9) which performs a BLAST search with the CE template sequence against variable IGRs of a reference genome and test isolate, and these scripts compare the position of variation in the IGRs with positions of CE matches and if the variation is located within CE positions, prints the names of these IGRs to a file for filtration (Figure 2.7).

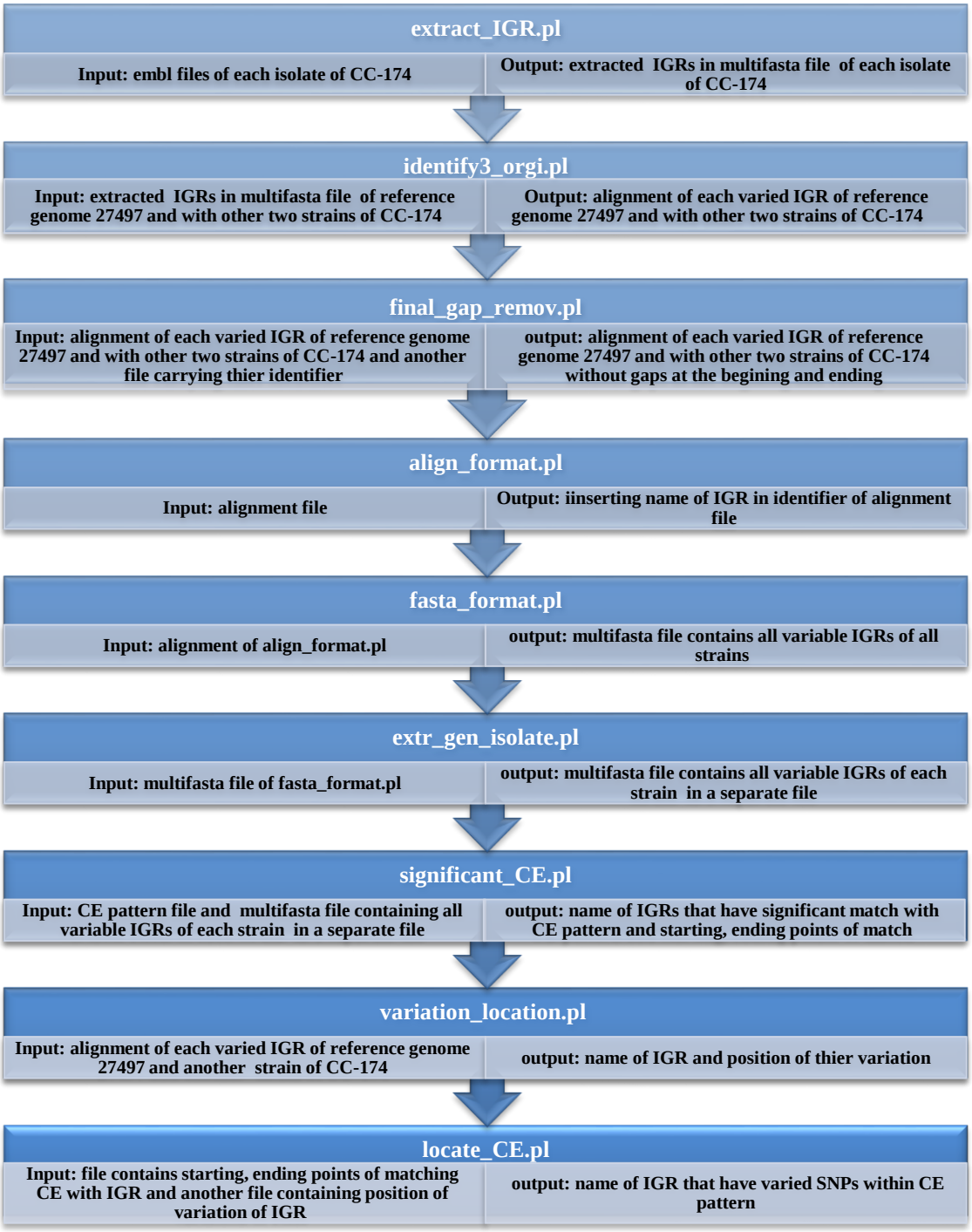


Figure 2.7: Flow diagram showing the Perl scripts to filter the IGRs.

2.5.4.3 Building contigs carrying the real variation (SNPs) of genic and IGRs

The variable genes/ IGRs identified by GC and AC were collected in one file for each CC-174 isolate using cat command.

Editing the full sequence of reference genome ID: 27497 to contain the real variation for each CC-174 isolate required two steps. First, each gene of reference ID: 27497 isolate was matched with its full sequence in the concatenated full genome sequence of same ID: 27497. Second, the matched gene sequence was replaced with the variable gene sequence from a particular CC-174 isolate. This process used Perl script replace_fullseq.pl (Appendix 16) and files containing the sequence of the variable gene of reference genome ID: 27497, the sequence of the variable gene of the target CC-174 isolate and the full genome sequence of the reference genome ID:27497.

The next step was to extract each contig of each modified isolate into a separate file. The PROmer (Delcher *et al.*, 1999) and ABACAS (Assefa *et al.*, 2009) scripts were used to order the contigs of ID: 27497 (See 2.6.2.1). The extr_contg.pl script (Appendix 17) was used to extract the sequences of different contigs for each isolate to remove N between contigs and to assign each contig name to the key and the sequence to the value before printing to a second file. The next step was to perform a BLAST search of each isolates that contained real variation against this second file. The starting and ending points of the alignment of each contig were reported and used by extractespoint.pl script (See later appendix 18) to extract the sequences of every contig of each isolate. These scripts generated one file containing all the contigs for each isolate of CC-174. Therefore, the final output was the full genome sequence carrying the real variation for each contig of CC-174 (Figure 2.8).

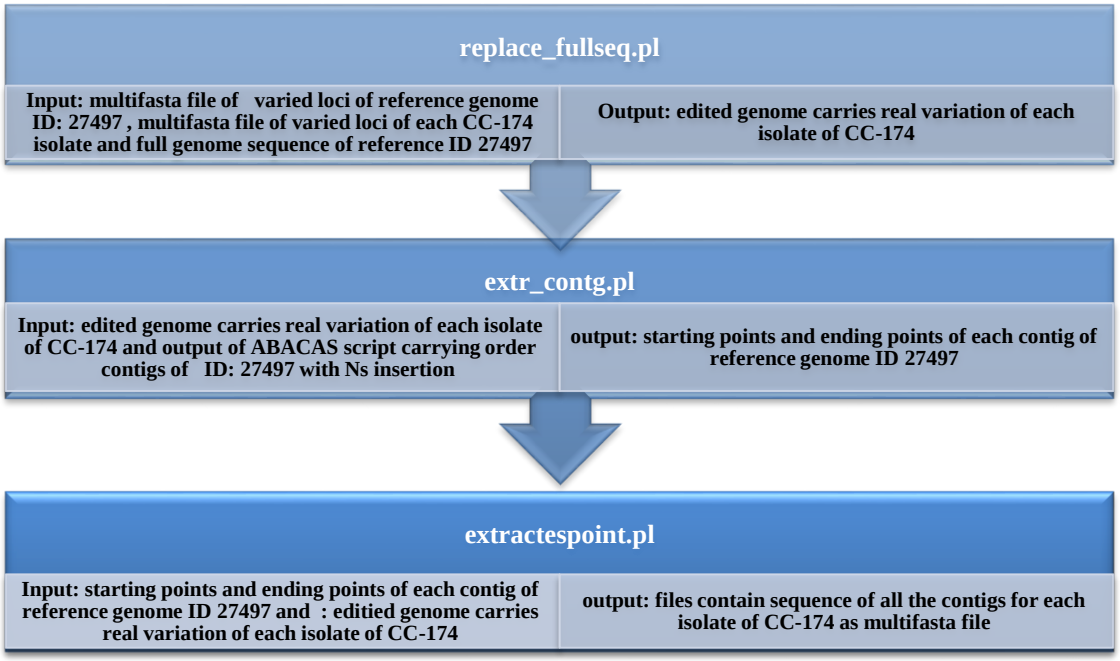


Figure 2.8: Flow diagram showing the Perl scripts used for building of a full genome sequence containing only the real variation. Variation was detected in the genic and IGRs of each contig of 25 CC-174. The 67 contigs containing only the real variation were analysed and used to infer recombination patterns among 25 CC-174.

2.6 Inferring evolutionary relationships between isolates

Evolution of groups of isolates was studied using phylogenetic trees (Section 2.6.1), estimation of recombination (Section 2.6.2), and analysis of selection (Section 2.6.3).

2.6.1 Construction of phylogenetic trees

2.6.1.1 Construction of phylogenetic tree using PhyML program

The PhyML program is hosted in the SeaView program but is considered an independent program from SeaView (Gouy *et al.*, 2010). Concatenated variable gene sequences were submitted to the program with application of specific parameters (Table 2.4). Bootstrap resampling (Felsenstein, 1985) was used to support each branch in the tree. The hill-climbing algorithm was used to adjust tree topology and branch length simultaneously. The program constructed trees using a fast distance based method with modification of the tree to improve its likelihood at each iteration (Guindon *et al.*, 2010).

Table 2.4: Parameters used to construct phylogenetic tree using PhyML program.

The model	General time reversible (GTR)
Branch support	Bootstrap method
Invariable site	None
Tree searching operations	Nearest Neighbor Interchange (NNI)
Starting tree	Bio neighbor joining (NJ) with random start: 5

2.6.1.2 Construction of phylogenetic tree using RAxML program

A phylogenetic tree was constructed as a first step for inferring recombination patterns for each contig using RAxML program (Stamatakis, 2014). The multifasta file of each contig was converted into Phylip format using the Sea View program and then used as input for RAxML program. The GTR substitution model with correction for among site rate variation was used and the program was set to perform 100 rapid bootstrap searches and then 20 maximum likelihood (ML) trees were searched to identify the best ML tree. These trees were saved as Newick format.

2.6.1.3 Construction of phylogenetic tree and a pairwise distance matrix using ParSnp package and MEGA program

The whole genome alignment was produced through comparing the user-defined isolates with a particular reference genome (i.e. one of the user-defined isolates used as a reference genome) as a first step to build phylogenetic tree using ParSnp V1.2 (Treangen *et al.*, 2014). A tree in the Newick format, the output of ParSnp, was submitted to FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/) for visualization and coloring of the isolates and the branches. The Gingr package was also used for interactive visualization of alignments, trees, and variants (Treangen *et al.*, 2014). A pairwise distance matrix was calculated using the MEGA program through genome pairwise comparison for the user-defined isolates (Tamura *et al.*, 2011).

2.6.1.4 Construction of phylogenetic trees for recombinant and mutant fragments

The extractespoint.pl script (Appendix 18) was written to extract the recombinant fragments from fasta files of each contig of a strain. The input files contained the start and

end points of all recombinant fragments detected using ClonalFrameML and the multifasta file of each contig. The script was designed to extract the sequence of each recombinant fragment and print all of them to one file. The extractespoint2.pl script (Appendix 19) was designed on same principle but the aim was to extract all the mutant fragments for each contig in a file. Therefore, the file containing start and end points of all recombination fragments detected using ClonalFrameML was modified to capture the mutation fragments. The recombinant and mutant fragments from all contigs were collected together in two separate files using cat command and then the extrfasta.pl script (Appendix 20) was used to concatenate all the fragments of each CC-174 isolate together. The final outputs were multifasta files for recombinant and mutant fragments that were used to construct phylogenetic trees using PhyML (See 2.6.1.1 section).

2.6.1.5 Phylogenetic network construction for the isolates

SplitsTree4 version 4.14.4 (Huson and Bryant, 2006) was used for the construction of a phylogenetic network. Input was character base or distance matrix and the network was drawn using median network or neighborNet algorithms respectively.

2.6.1.6 Haplotype network construction

The HapView program (Matschiner, 2016) was used to construct a haplotype network. The input data was a multifasta format for the aligned sequences of multiple isolates and the Newick format of phylogenetic tree construction.

2.6.2 Detection of recombination patterns

2.6.2.1 Inferring recombination patterns in genic and IGRs using a sliding window approach

The recombination signal and the length of recombination fragments in the genic and IGRs sequences of paired or closely related isolates were inferred through a series of five steps using a method adapted from (Kong *et al.*, 2013).

Firstly, an estimation was derived for nucleotide differences between compared isolates of the same sequence type arising by point mutation. As these isolates are from the same sequence type, the upper bound of nucleotide divergence (μ) was estimated assuming that

the seven MLST loci are identical for the compared genomes. Building on the fact that mutations are independent events and follow a binomial distribution, estimation of the μ for MLST genes was calculated for a P value 0.001 assuming no change in the MLST loci using the following formula $(1- \mu)^n = 0.001$ where n is the full length of concatenated sequences of MLST loci which was (3288 bp). Using this equation, μ was found to be 0.0021.

Secondly, the μ of MLST was used to calculate the threshold for the expected number of mutations for P values larger or equal to 0.001 using the following formula.

$$\sum_{K=X}^{N=y} f(K; \mu) \leq 0.001$$

The formula has three variables -x substitutions in y nucleotides and μ - and was applied using the R package for a range of window sizes. The threshold for the expected number of mutations was 5 in 500 bp. Therefore, if the number of mutations in a fixed window length of 500 bp was equal to or larger than 6, then this sequence would be deemed a recombinant fragment (Figure 2.9).

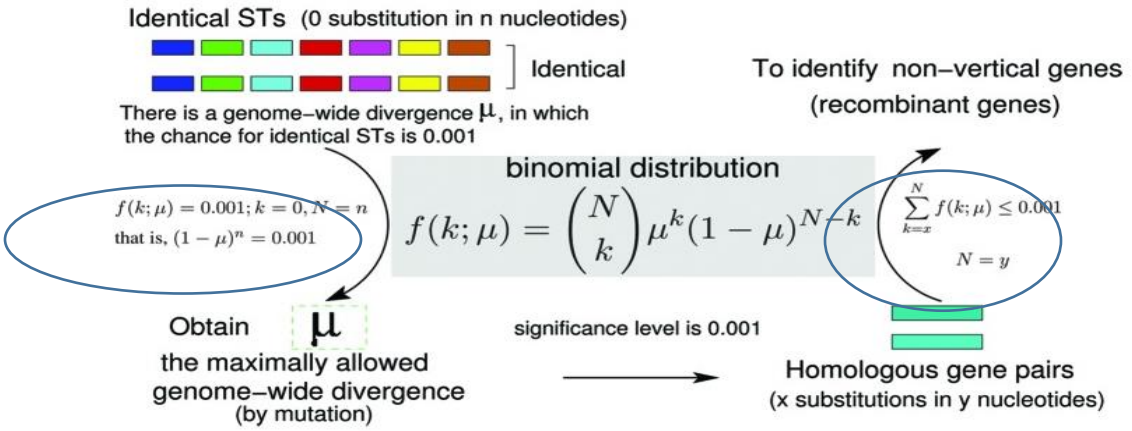


Figure 2.9: Diagram showing the equations for detection recombinant genes depending on (Kong *et al.*, 2013). The formula on the left side can be used to calculate the upper bound for divergence. The formula on the right side was used to calculate the expected number of mutations for definition of a recombinant sequence given a particular window size. This diagram has been taken from (Hao, 2013).

Thirdly, this step and the next step are explained as an example and these were done for one isolate of every pair. Co-linear genomes were required in order to examine recombination in two or more genomes. This required mapping of the positions of each variable nucleotide onto a reference genome. The multiple contigs of strain N420 (ID: 17012) was selected as a reference genome among 25 pair isolates. The first step of the mapping was used to align all the contigs of isolate 17012 against MC58 as a reference genome. The PROmer script was used because the alignment was dependent on the six frame amino acid translation of the query and hence can find more conserved regions than using DNA sequence. The output of PROmer is a promer.delta (this file refers to the alignment between reference and query isolate) which was then used as input for ABACAS. This program was employed to align, order and orientate all the contigs of N420 (ID: 17012) relative to the MC58 genome. This program inserts multiple Ns at the ends of each contig. Finally, the Artemis Comparison Tool (ACT) (Carver *et al.*, 2005) was used to visualize the outputs of the ABACAS script and to identify the positions of each varied nucleotide of all genic and IGRs of isolate to be compared. The input files were: - the output of ABACAS (order contigs of N420 (ID: 17012) with Ns insertion), the output of ABACAS (comparison file reference_query.crunch extension) and the MC58 genome in fasta format.

Fourthly, the 17012 contigs were edited to insert the variable nucleotides for genome to be compared, all the nucleotide differences in variable genic regions and IGRs were manually mapped using the Artemis program and manually edited. This resulted in two files containing the original sequence and varied sequence.

Fifthly, two scripts were written to automate detection of recombination signals using a sliding window approach. The index_var.pl script (Appendix 21) compares the DNA sequence from two different files the original sequence and the varied sequence of each contig of each genome. This script prints the position of each varied base pair between two different files into one file. The output of, the index_var.pl script was used as input for the second script sliding_win.pl (Appendix 22). The sliding_win.pl script assumed that at least six point mutations were required in a 500 bp sequence to infer recombination. The script prints the starting points and number of varied nucleotides for each window that passed the threshold. Having detected recombination, the script reports the length of each recombinant

fragment. The length was estimated by subtracting the starting point of first window that passed the threshold from the starting point of the first window below the threshold.

2.6.2.2 Detection of recombination fragments using ClonalFrameML

The Newick format of each phylogenetic tree and the fasta file of each contig were the two files required for running the ClonalFrameML program (Didelot and Wilson, 2015). The following parameters were used for running ClonalFrameML; -kappa 4.97 (relative rate of transitions versus transversions in substitution model), -emsim 100 (number of simulations to estimate uncertainty in the expectation-maximization algorithm (EM) result). There were five output files:-ML_sequence.fasta, a multifasta file of internal nodes of phylogeny and missing data; position_cross_reference.txt,mapped position of input files and ML_sequence.fasta file; em.txt,bootstrap values of R/theta and delta; important_status.txt , detected recombination events; and labeled_tree.newick, the output tree.

2.6.3 Identification of types of selection

The MEGA program (Tamura *et al.*, 2011) was used to search for evidence of neutrality, positive selection, and purifying selection. The analysis was done by calculating a codon based Z-test of selection (Table 2.5) (Nei and Kumar, 2000). The analysis was determined depending on the abundance of synonymous and non-synonymous substitutions in a pair of sequences belonging to a particular gene. The abundance of synonymous and non-synonymous substitution was calculated by estimating the following: - the number of synonymous substitutions per synonymous site (dS); the number of nonsynonymous substitutions per nonsynonymous site (dN); the variance of dS and dN. The null hypothesis (H0) was that dN = dS and this was tested using a Z-test:

$$Z= (dN-dS)\sqrt{(Var(dS)+Var(dN))}$$

Alternative hypothesis (AH) are that $dN>dS$ (positive selection) or $dN<dS$ (purifying selection). A one-tailed test was used to test the null hypothesis while a two-tailed test was used to test the alternative hypothesis with P-value (0.05). Bootstrap resampling was used to compute variance. For samples with more than two sequences, the average number of synonymous substitutions was used to perform the Z-test (Nei and Kumar, 2000).

Table 2.5: Parameters used to calculate codon based Z-test of selection.

The scope	Average number of sequences
Variance estimation method	Bootstrap method
The number of bootstrap replications	500
The model	Number of nucleotide different
Missing data treatment	Partial deletion

2.7 Estimation the genome diversity among the isolates

The genome diversity was estimated as follows:-

2.7.1 Identification of the mean of nucleotide diversity in the variable loci

Measuring mean diversity and phylogenetic tree construction required a concatenated sequence of all variable genes for each isolate. The fasta_format.pl script was used to concatenate the sequences of all variable genes for each isolate (See previous appendix 5). The mean diversity for multiple isolates is the estimation of the average frequency of i-th (number alleles in the population) and q (number of different sequences in the population). A comparison of the difference in mean diversity between two populations was carried out in the MEGA program, using the concatenated variable genes as input. For example, the concatenated sequences for each isolate belonging to the first and third time point were classified into two groups in MEGA. The analysis was then performed using the following parameters (Table 2.6) (Tamura *et al.*, 2011).

Table 2.6: Parameters used to estimate the mean rate of nucleotide diversity.

The scope	Average between populations
Variance estimation method	Bootstrap method
The number of bootstrap replications	500
The model	Number of nucleotide different
substitutions	To include, transitions and transversions
rate among site	Uniform rate

2.7.2 Determination of the genomic variation of genic regions for persistent carriage isolates

The genomic variation for persistent carriage was calculated by dividing the number of variable genes for a pair of persistent carriage isolates by the total number of analyzed genes. This was adjusted for time of carriage by dividing the genic variation by the months of carriage for each pair of isolates. Where multiple isolates were available for each time point, the mutation frequency was calculated by adding the total number of isolates with each variable gene in each time point then dividing them by the total number of analyzed genes for all isolates.

2.7.3 Estimation of the nucleotide diversity among 25 pair isolates

The nucleotide diversity was estimated for the 25 pair isolates and for each functional scheme as reported by KEGG. In each pair of isolates, the replace_fullseq.pl script (See 2.5.4.3 section) was used to edit the original sequence by inserting the variable nucleotides from the second genome after variation. Therefore, two co-linear genomes were generated for each pair of isolates. These genomes were submitted into MAFFT program (Katoh and Standley, 2013) for achieving alignment with Pearson/FASTA format. Then, the aligned co-linear genomes were submitted into a DnaSP program (Librado and Rozas, 2009) to estimate nucleotide diversity. The program was set up to estimate nucleotide diversity Pi (t). In the same way, the nucleotide diversity was estimated for each variable gene belonging into different functional schemes. Then all the nucleotide diversity of genes in each scheme was calculated and the average of nucleotide diversity for each scheme was compared.

2.7.4 Estimation the number of variable loci and SNPs using statistical analysis

Poisson mean confidence interval (www.evanmiller.org/lab-testing/Poisson-means.html) was used to test the statistical significance of data consisting of a count of the number of events occurring in a specific time (Ross, 2003). Therefore, this test was used to analyze difference in the number of CE with time points. While Mann-Whitney U test (Graphpad Prism version 5) was used to determine statistical differences in the medians of two groups. The null hypothesis was that there was no difference between the medians of the two samples while the alternative hypothesis stated that there was a difference (Milenović,

2011). This test was used to estimate the difference in the number of loci with time points for the 25 pair isolates. Furthermore, Chi-square test (Graphpad Prism version 5) was used to determine number of indel in genic versus IGRs. Finally, the oneway anova test (http://statpages.info/anova1sm.html) was used to estimate the difference in the nucleotide diversity for different functional schemes.

2.8 Parameter for analyzing the function of varied genes and intergenic under the study

2.8.1 Identification of synonymous and non-synonymous polymorphisms in the variable loci

To identify if a mutation would be synonymous, non-synonymous or produce an internal stop codon, an analysis was performed using the SNAP program (Korber, 2000) based on the method of (Nei and Gojobori, 1986). The statistical analysis used was developed by (Ota and Nei, 1994). The input file was a fasta file of two compared genes or genomes; the output file contains the position of each amino acid, type of change (synonymous, non-synonymous) with their significant values, the position and number of stop codons. In addition into summary of statistics for calculation of the dN/dS ratio. The set of codon-aligned nucleotide sequences were the principle data for calculation of synonymous and non-synonymous substitution rates by SNAP. The position of altered amino acids in the data under the current study was detected using the SNAP program.

2.8.2 Detection of the position of variable amino acids in conserved domains of each variable gene

A BLAST search was performed for each variable gene against NCBI's conserved domain database (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpb.cgi). The *E*-value equal or less than 1*E*-05 was considered as a statistically significant value for assigning the particular sequence to a matched conserved domain (Marchler-Bauer *et al.*, 2014; Marchler-Bauer *et al.*, 2011). The result of this search showed the starting and ending points of conserved domains within each query gene. The position of each variable amino acid within each conserved domain, type of amino acid change, the name of the conserved domain and identifier were recorded.

2.8.3 Visualization of the three dimensional structures of proteins encoded by variable genes of 40 isolates

Visualization of the three dimensional structures of proteins that were coded by some of the variable genes was carried out using the PyMOL tool (Delano, 2002). In the case of those proteins whose structures were not been solved by x-ray or NMR (nuclear magnetic resonance), a three dimensional structure that has already detected by the homology modeling, was taken from Protein Model Portal (PMP) (www.proteinmodelportal.org). The ID of each variable gene was submitted into Universal Protein Resourse (UniProt), then the pdb file containing three dimensional structure of protein with the highest percentage identity was taken from PMP hosted in the (UniProt) (Bairoch *et al.*, 2005) (www.UniProt.org). PyMOL was then used to visualize the structure and the sequence of each protein. A cartoon visualization was selected and the protein was colored depending on the named/known domains. Finally, the position of the mutation (i.e. variable amino acid) was highlighted by drawing a sphere at that position.

2.8.4 Detection of functional schemes and proportional effect of variation of variable genes

Functional schemes were identified for each gene using the Kyoto Encyclopedia of Genes and Genomes (KEGG) scheme a database that assigns biological function for genes and stores high-level functions in the form of networks of molecular interactions (Kanehisa *et al.*, 2015). KEGG was used to detect the functional scheme of each variable gene. The proportional effect of variation on different functional groups was determined as follows:-

The amino acids sequences of all the genes were extracted in GenBank format for isolate MC58 using extract_aa.pl script (Appendix 23). The script is designed to loop through all the CDs and to match the identifier of each gene and their amino acids. Finally, each identifier and their amino acid sequences were printed in fasta format. These amino acid sequences were then submitted to a BLAST search hosted in KEGG (Kanehisa *et al.*, 2016). This search assigns each gene to a KEGG functional group. Finally, the percentage of variable genes in each functional group was calculated by dividing the number of variable genes in each group by the total number of genes in the whole genome belonging into this particular group.

2.8.5 Prediction of promoters and sRNA in IGRs

A potential promoter was predicted using the Fgenesb_annotator pipeline (Solovyev and Salamov, 2011). This application can be additionally used to predict tRNA sequences, rRNA terminators and operons. The program is designed to recognize sigma 70 promoter sequences with 80% accuracy and specificity. This prediction is more accurate if the region is located immediately upstream of the open reading frame (ORF) (Solovyev and Salamov, 2011). The input file was a multifasta file containing variable IGRs, the output file contained a log-likelihood score (LDF) for each pattern of each IGRs.

The presence of sRNA was predicted using a BLAST search. The query was each variable IGR against sRNA sequences that have been taken from (Del Tordello *et al.*, 2012).

2.9 Perl scripts for extracting and manipulating CEs in the intergenic DNA sequence

CEs vary in length. Complete CEs span 153 to 157 bp with a 26 bp inverted repeat at both ends. Partial CEs span 104-108 bp and have a deletion of around 50bp in the middle. Alternatively, partial CEs span 60-62 bp and have a big deletion in the middle with a 26 bp inverted repeat at one side (Snyder *et al.*, 2009).

CEs were searched within the variable IGRs (See section 2.5.3). The fasta_format.pl script (See previous appendix 5) was used to concatenate sequences for all variable IGRs in alignment format for each isolate. Then the extr_gen_isolate script (See previous appendix 6) was used to extract variable IGRs for each isolate from this alignment file and save them separate multifasta files.

To detect the CEs, the Significant_CE.pl script (Appendix 24) was designed to achieve a BLAST search for CE template (complete CE 157 bp) as a query against the multifasta file of every gene and IGR sequence of each isolate (output of extr_gen_isolate.pl script). The script reports the significance values of matching sequences (E-value > 1e- 20, identity > 90%, length of alignment > 50pb) and the starting and ending points of matches within each IGR. The variation_location.pl script (Appendix 25) was used to detect and report the position of varied nucleotides in all variable IGRs for each pair of isolates.

Finally, the positions of varied nucleotides of IGRs that showed significant matching with

template CEs were further checked using locate_CE.pl script (Appendix 26). The input files for the locate_CE.pl script was the output of both Significant_CE.pl and variation_location.pl scripts. If the position of varied nucleotides detected by variation_location.pl script was located between the starting and ending points of a CE, the variation was reported as due to the presence of CE patterns. These variable CEs were then examined for the presence of inverted repeats, IHF patterns, Black, Snyder promoters and (AC or AT) ending sites.

2.10 Phase variation and SNPs validation

2.10.1 Bacterial isolates and growth conditions

The alteration of SSR in five phase variable genes during persistent carriage was examined using DNA extracts provided by Dr. C. D. Bayliss.

2.10.2 Study primers

The primers used for the purpose of repeat tract analyses and validation of SNPs were either taken from (Alamro *et al.*, 2014; Tauseef *et al.*, 2011) or designed using Clone manager 9 software (Table 2.7).

Table 2.7: List and sequences of primers used in the current study.

Name	Sequence	Comment
HpuA - vic for	ATGCGATGAAATACAAAGCCC	Fluorescent labelled and non-fluorescent forward primers to amplify repeat tract in *hpuA*.
HpuA- 350 rev	GGATGAAAGGGCGTATTGCGC	
HpuA C for	ATGCGATGAAATACAAAGCCC	
nadA for fam	TCGACGTCCTCGATTACGAAGGC	Fluorescent labelled and non-fluorescent forward primers to amplify repeat tract in *nadA*.
nadA rep. rev	TGGCTGTGGTCAGTACTTTGGATGG	
nadA for	TCGACGTCCTCGATTACGAAGGC	
Opc for fam	GAGAATAACAATTCGTTGTA	Fluorescent labelled and non-fluorescent forward primers to amplify repeat tract in *opc*
Opc rep rev	CTCATTAGCGGTTTGAAGCTCTTGTGCAG	
Opc for	GAGAATAACAATTCGTTGTA	
*nalP*F1	GTTGCAACAACACTTTCTGCCTGC	Fluorescent labelled and non-fluorescent forward primers to amplify repeat tract in *nalP*
*nalP*R1 for sequecing	GCAGGTTGTCGTTGCTCATCCACG	
*nalP*R2 for sequecing	CAGGCGCTTCCTTCCGCATATACG	
*nalP*F2-FAM	AAATGTGCAAAGACAGAAGCATGC	
hmbr-RF3 for sequecing	TGCCAACCTCTTTTACGAATGG	Fluorescent labelled and non-fluorescent forward primers to amplify repeat tract in *hmbr*
hmbr-RF4 for sequecing	GCTACTGAACACGTCGTTCC	
hmbR-for 2	CGGCATTCAGGTCAAAATCCC	
hmbR-Rev 7	GCCGAAGGATCCAATTTGGG	
NMB1390: Wildtype Forward Primer	ACCCGCCAAACTGATGAC	for SAP method detection *NMB1390*
NMB1390: Mutant Forward Primer	ACCCGCCAAACTGATGAT	
NMB1390: Common Reverse Primer	TGTCGAGAGCCGCAGCCG	
NMB1390: Forward primer	TACGCCGAATAAACAAGCCG	For sequencing *NMB1390*
NMB1390: Reverse primer	GAACGGCGCTAAGGGCAA	
NMB0329: Wildtype Reverse Primer	CAGCTCGAACACAGCCTAC	for SAP method detection *NMB0329*
NMB0329: Mutant Reverse Primer	CAGCTCGAACACAGCCTAA	
NMB0329: Common Forward Primer	ATGAGCGTAGGTTTGCTGAG	
NMB0329: Forward primer	ATGAGCGTAGGTTTGCTGAG	For sequencing *NMB0329*
NMB0329: Reverse primer	GATTAACCTGCCGGGCATC	
NMB0329: Reverse primer	CGCTGATGGGCATAACCTC	

2.10.3 Polymerase chain reaction (PCR)

The target DNA sequences were amplified by PCR with the following components, 1 µl of a 1:10 dilution of DNA (~15 ng/µl) as template, 0.2 µM final concentration of forward and reverse primers (2µM stock concentration), 1 µl of a 10x PCR buffer (10x KAPA Taq buffer A with 15mM $MgCl_2$), 0.4 µl of 25mM $MgCl_2$, 0.25 µl of 10mM dNTPs, 0.1 µl of a 5U per µl stock of Taq DNA polymerase (KAPA Biosystems) and 5.25 µl of sterile distilled water. The reaction conditions for each cycle were:-95 °C for 30 seconds, an annealing step (52-60) of 60 seconds and an elongation step at 72 °C for 1 minutes.

2.10.4 A-tailing

An A-tailing step was used when PCR products were subject to GeneScan. The mixture contains 10 µl PCR reaction and A-tailing mix solution that is 0.4 µl PCR buffer, 0.05 µl Taq, DNA polymerase and 3.55 µl dH_2O, and then incubated for 45 minutes at 72 °C.

2.10.5 GeneScan

The variation in the length of PCR products spanning the repeat tracts of five genes was measured using GeneScan. Repeat tracts were amplified with specific primers, one of which was labeled with a fluorescent dye. The GeneScan components were:-0.5 µl of the 1:10 dilution of the PCR products (1:10 in sterile distilled water), 9.25 µl of formamide and 0.25 µl of DNA size standard GS500 LIZ (Life Technologies). Samples were analyses on an ABI3730 machine by the Protein and Nucleic Acids laboratory (PNACL). Finally, Peak scanner software v1.0 (Applied Biosystems) and Microsoft Excel were used to analyze the sequencer data.

2.10.6 DNA sequencing

The reactions for target DNA sequences were set up with the following components, 0.5 µl of DNA PCR product or 1 µl of plasmid DNA, 4 µl of sequencing mix (1:8 Big Dye v3.1 , 1:5 5x Sequencing buffer), 1 µl of primer required (forward or reverse), and water to make the total volume of 10 µl. The reaction conditions of each cycle were 96 °C for 30 second, 50°C for 15 second and 60 °C for 4 minutes. Reactions were analysed on an ABI 3730 DNA Sequencer (ABI, Applied Biosystems) at PNACL.

2.10.7 Agarose gel electrophoresis

PCR products were analyzed on gels containing 1% agarose (Seakem LE Agarose, Cambrex), 1x TAE buffer (40mM Tris acetate, 1mM EDTA, pH8.2), and 0.5 Mg/ml ethidium bromide. The DNA samples were mixed with 6x loading Dye (0.25% bromophenol blue). A DNA standard marker, Hyper ladder 1kb (Bioline Reagents Ltd.) was used to calculate the size of PCR fragments. Gel visualization was performed using a transilluminator gel documentation system (Syngene).

2.10.8 SNPs Validation using Simple Allele discriminating PCR (SAP) method

Validation of variable genes was achieved using a SAP method (Bui and Liu, 2009). The SAP method relies on having a mismatch in forward primers that result in a weak destabilization between the allele-specific (AS) primer and its non-template target combined with a strong destabilizing mismatch at the penultimate site or some variant thereof as illustrated in (Table 2.8). A step-by-step approach was used to design SAP primers (Figure 2.10):-1, forward primer designed as complementary to Wild Type (WT) template sequence; 2, forward primer was designed for Mutant Type (MT) template sequence; 3 the terminal mismatch (GT) was determined from pairing of wild type primer with mutant template; 4 the terminal mismatch (AC) was determined from pairing of mutant type primer with wild template. As a weak destabilization effect was produced from the terminal mismatches (GT) and (AC), a G was inserted at the penultimate site because this produces the strongest destabilization mismatch (i.e. GA).

Table 2.8: The all combinations of nucleotide pairing used to design primers in SAP method.

Base Pairing	Destabilization Strength
GA, CT, TT	Maximum
CC	Strong
AA, GG	Medium
CA, GT	Weak
AT, GC	None

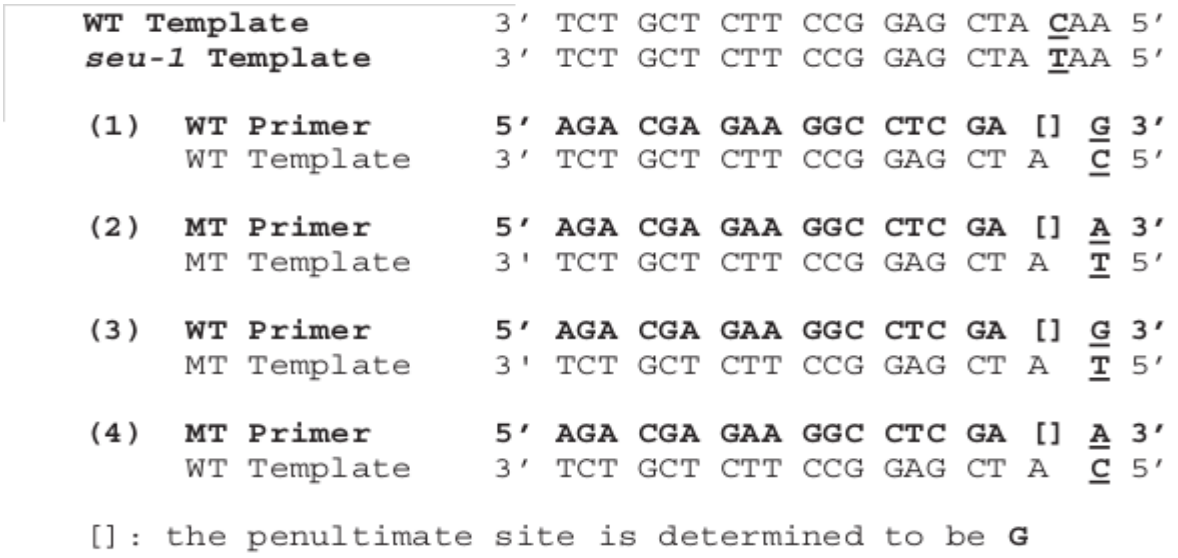


Figure 2.10: The step-by-step approach used for designing primers in SAP method. A common reverse primer was designed to pair with the two AS primers (wild and mutant forward primers). A web-based computational design tool using SAP method is available at http://bioinfo.biotec.or.th/WASP [17].

Chapter 3: Analysis of genome variation in 40 isolates

3.1 Introduction

A large number of genome comparator studies of *N. meningitidis* and other Neisseria species have been carried out and have revealed key attributes of this species and genus. Firstly, several studies have investigated the genomic content of commensals and their evolutionary relationship with disease isolates. In this context, it has been reported that commensal Neisseria play a significant role as reservoirs of virulence alleles for their pathogenic counterparts as well as other bacteria and are frequently involved in genetic exchange with pathogenic species through recombination (Marri *et al.*, 2010). Moreover, to gain insight into the evolution of virulence traits in disease isolates, comparative studies revealed that the difference between disease and commensals was not only in the presence of specific types of capsule polysaccharide in the disease isolates but that a filamentous prophage (involved in rearrangements and the translocation of some candidate virulence genes) and insertion sequence IS1655 were over represented in the disease isolates. Furthermore, a study showed that some highly exchangeable genetic islands were present in a set of virulent strains of *N. meningitidis* and *N. gonorrhoeae* but absent from other species such as *N. lactamica*. These islands are also involved in the translocation of some candidate virulence genes (such as *pilC1*) that are responsible for control of the type IV pili adhesion structure (Perrin *et al.*, 2002). Evolutionary relationships between disease and commensal species has been reported for *hmbR, exl2,* or *exl3* (genes required for heme utilization). The *hmbR* gene was found in the population of *N. meningitidis* whereas commensal neisserial species were found as a reservoir for all the three types of heme utilization genes (Kahler *et al.*, 2001).

Secondly, genome comparator studies have compared new serogroups, disease isolates or commensal isolates of *N. meningitidis* with historic isolates that have already been characterized. This study found differences in repeat patterns, repeat types, other mobile elements and a range of other genomic features. One study showed that comparing an ST-4821 isolate with other CCs revealed that there were thousands of repetitive elements and simple sequence repeats, numerous phase variable genes, and

similar virulence-related factors that were shared by isolates from different CCs (Peng *et al.*, 2008). Subsequently, other studies have shown that variation in specific part of the genome is enhanced by localized recombination. In this content, variation of surface antigens was found to be due to repeat elements (neisserial intergenic mosaic elements) which act as a target for intergenic recombination in some strains of *N. meningitidis* (Bentley *et al.*, 2007).

Thirdly, some genome comparison studies were conducted to assess presence, distribution, variation, and conservation of some proteins used as vaccine components such as factor H-binding protein (fHBP), neisserial heparin-binding antigen (NHBA), and *N. meningitidis* adhesion A (NadA). It has been reported that these proteins were conserved in *N. meningitidis* colonizing a particular host (human). Regarding fHBP, three evolutionary variants were identified, variant 3 was ancestral while variant 1 transferred from *Neisseria cinerea* into *N. meningitidis*. Variant 2 was the product of recombination of variant 1 and variant 3 (Muzzi *et al.*, 2013). Similarly, HmbR was mainly found within *N. meningitidis* and *N. gonorrhoeae* while HpuAB was distributed in different Neisseria species. Recombination and phase variation was the main processes that generated variation within HpuAB (Harrison *et al.*, 2013).

Finally, epidemiological studies of *N. meningitidis* isolates from different geographical regions have been investigated to aid in the formulation of novel vaccines and assess their potential coverage across countries and continents. The EU-MenNet programme aimed to carry out molecular typing — MLST, *porA* and *fetA* — of European disease isolates. It is crucial for epidemiological studies (Brehony *et al.*, 2007).

The previous mentioned studies provide overview of the evolution and diversity at a given time and they do not introduce information on the evolution and diversity for the bacterial population within their host during the colonization. Therefore, in this chapter, a study was conducted with 40 isolates of *N. meningitidis* from carrier V59 collected from four time points with 10 isolates at each time point: 0, 1, 3 and 6 months (Table 3.1). Genomes were sequenced, assembled and annotated by Illumina Hiseq, Velvet and Prokka respectively (See chapter two 2.2). The novelty of the current study is the analysis of genetic variation using genome comparison on the population of

isolates (10 isolates in each time point) with the purpose of inferring evolution of isolates within their host and for the purpose of finding the patterns and dynamics of variation in such regions that may have a role in persistence of isolates for months.

Table 3.1: Listing the name, ID and time point of isolation of 40 isolates of *N. meningitidis.*

NO	IDs	Isolate	Time (months)	NO	IDs	Isolate	Time (months)
1	**28299**	N59.1	0	**21**	**28320**	N352.2	3
2	**28300**	N59.3	0	**22**	**28321**	N352.3	3
3	**28301**	N59.4	0	**23**	**28322**	N352.4	3
4	**28303**	N59.6	0	**24**	**28323**	N352.5	3
5	**28304**	N59.7	0	**25**	**28324**	N352.6	3
6	**28305**	N59.8	0	**26**	**28325**	N352.7	3
7	**28306**	N59.9	0	**27**	**28326**	N352.8	3
8	**28307**	N59.10	0	**28**	**28327**	N352.9	3
9	**28308**	N59.11	0	**29**	**28328**	N352.10	3
10	**28309**	N253.1	1	**30**	**28329**	N438.1	6
11	**28310**	N253.2	1	**31**	**28330**	N438.2	6
12	**28311**	N253.3	1	**32**	**28331**	N438.3	6
13	**28312**	N253.4	1	**33**	**28332**	N438.4	6
14	**28313**	N253.5	1	**34**	**28333**	N438.5	6
15	**28314**	N253.6	1	**35**	**28334**	N438.6	6
16	**28315**	N253.7	1	**36**	**28335**	N438.7	6
17	**28316**	N253.8	1	**37**	**28336**	N438.8	6
18	**28317**	N253.9	1	**38**	**28337**	N438.9	6
19	**28318**	N253.10	1	39	**28338**	N438.10	6
20	**28319**	N352.1	3	40			

To achieve these aims; this chapter analyzes the genome variation in genic and IGRs sequences; firstly, the importance of variable genes in the evolution of isolates with their host was predicted depending on three criteria which are function of gene (hypothetical or functional genes), type of variation (synonymous gene or non-synonymous genes) and the location of variation within the conserved domains. This chapter was extended to look for the number of varied nucleotides using non-parametric statistical tests and by estimating mean of nucleotide diversity, and mutation

frequency. Finally, the complete picture of isolate evolution was examined by comparing the evolutionary relationships inferred from genic and IGR sequences.

3.2 Analysis of variability in the genic regions of 40 meningococcal isolates using GC method

To start with, a phylogenetic tree was drawn from the whole genome sequence for all the 40 isolates using ParSnp v1.2. The entire genome assembly of all the 40 isolates was submited to the ParSnp along with entire genome assembly of N59.1 as a reference genome. The ParSnp package was able to align the entire genome assembly of all the 40 isolates and to build the tree from extracted SNPs. The percentage of core genome alignment among all sequences of 40 isolates using ParSnp was 97.2% (2045820 bps out of 2104753 bps). The tree visualisation using FigTree v1.4.2 and Gingr package showed that nine isolates at the second time point cluster together and separately from the other isolates. Contrastingly, the isolates from the first, third and fourth time points were mixed with each other (Figure 3.1).

Distance matrices were built for the actual number of SNPs differences between each pair of genomes using the NJ algorithm hosted in MEGA7 program. The pairwise distance matrix showed that the allelic differences were highest for the pairwise genome comparison between isolates in the second time point with isolates in other time points (Table 3.2). This result was compatible with clustering of isolates (See Figure 3.1) where the isolates in the second point cluster together and separately from the other isolates due to high genome variation comparable with other isolates in other time points.

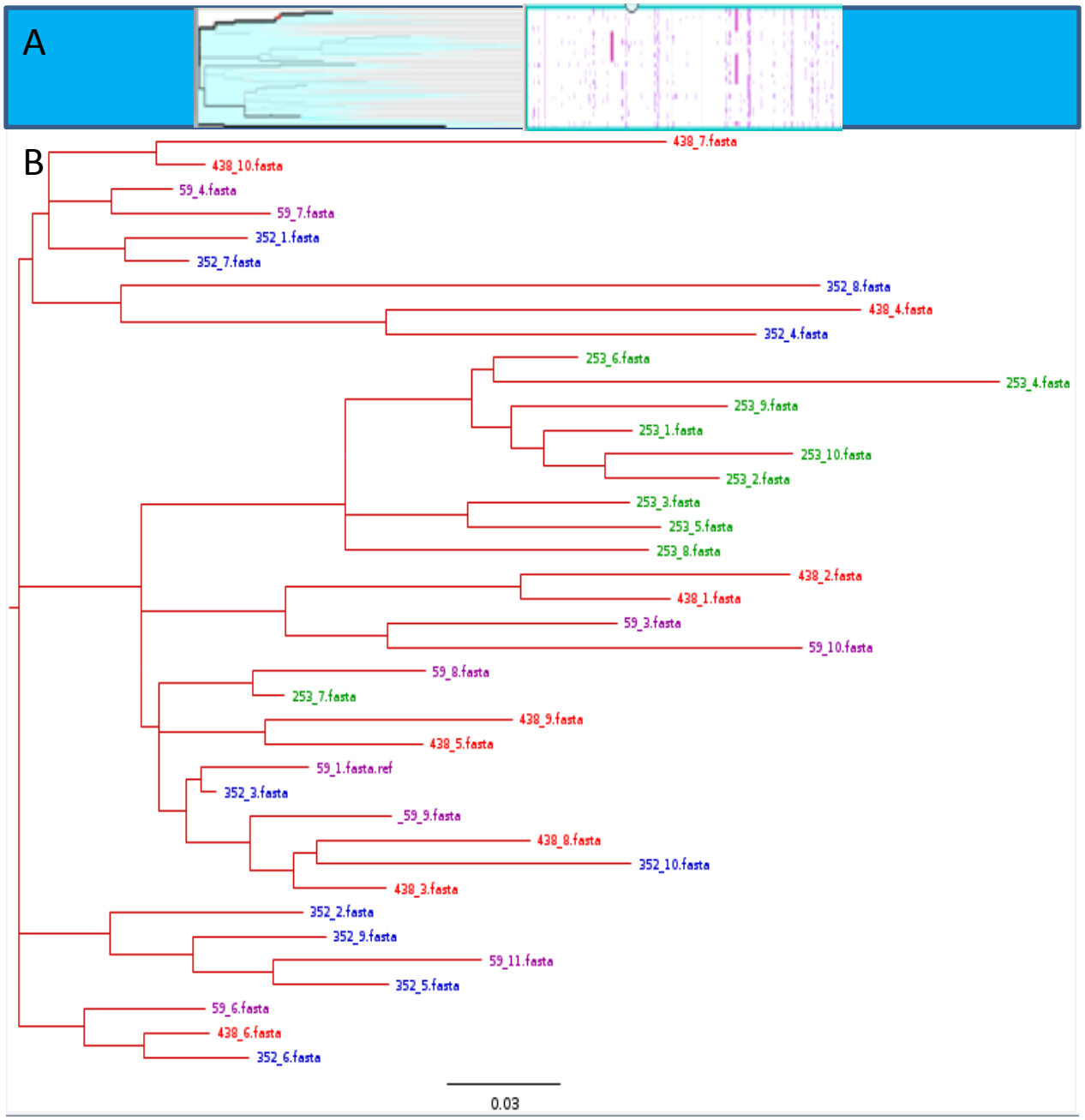
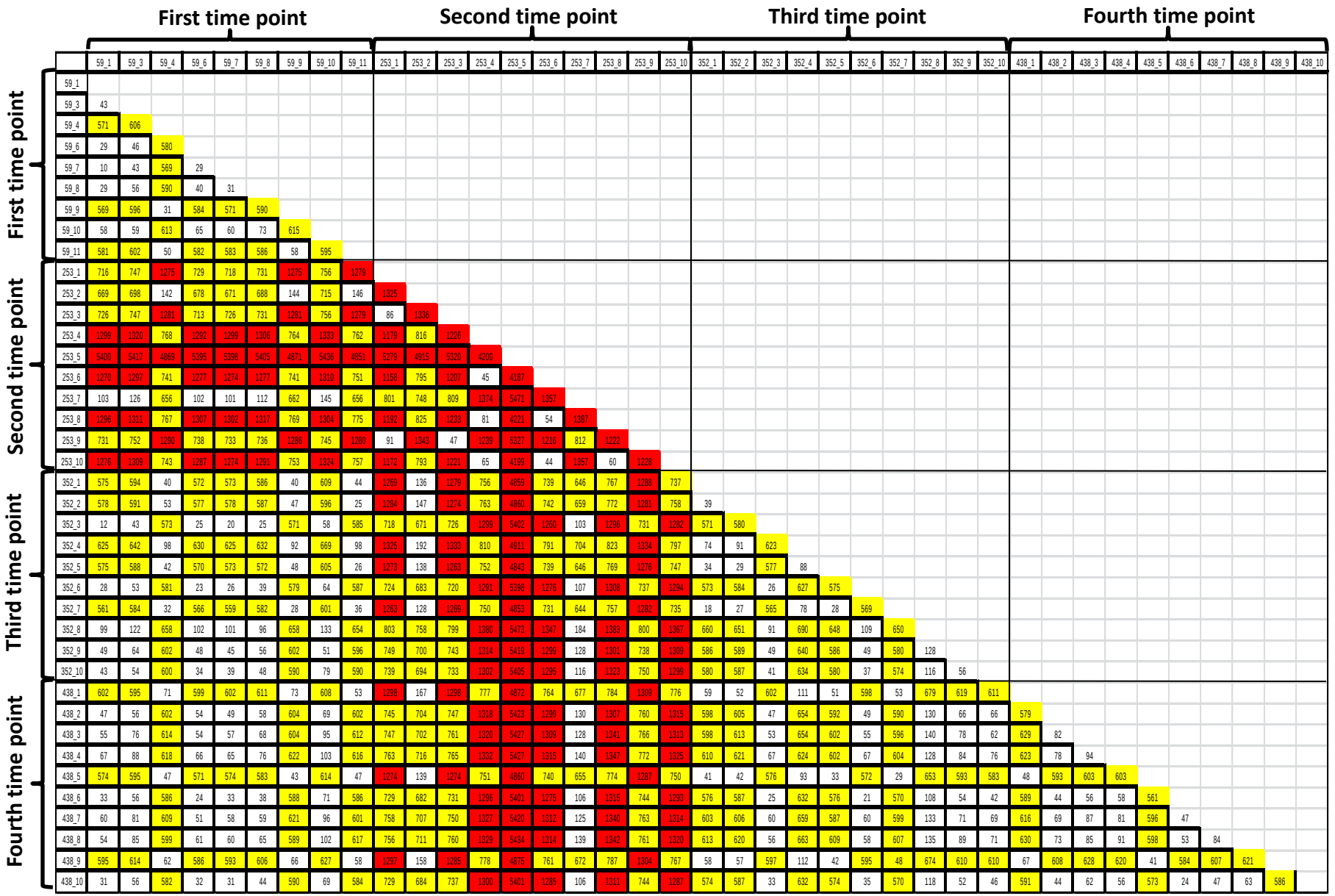


Figure 3.1: Phylogeny of 40 carriage isolates from a volunteer V59.1 constructed from whole genome sequences using ParSnp package. Panel A: schematic representation of phylogeny along with the genome variation depicted with pink line along each isolate in the phylogeny. Panel B : schematic representation of phylogeny. Purple color: isolates in first time point. Green color: isolates in second time point. Blue color: isolates in third time point. Red color: isolates in fourth time point.



	59_1	59_3	59_4	59_6	59_7	59_8	59_9	59_10	59_11	253_1	253_2	253_3	253_4	253_5	253_6	253_7	253_8	253_9	253_10	352_1	352_2	352_3	352_4	352_5	352_6	352_7	352_8	352_9	352_10	438_1	438_2	438_3	438_4	438_5	438_6	438_7	438_8	438_9	438_10
59_1																																							
59_3	43																																						
59_4	571	606																																					
59_6	29	46	580																																				
59_7	10	43	569	29																																			
59_8	29	56	590	40	31																																		
59_9	569	596	31	584	571	590																																	
59_10	58	59	613	65	60	73	615																																
59_11	581	602	50	582	583	586	58	595																															
253_1	716	747	1275	729	718	731	1275	756	1279																														
253_2	669	698	142	678	671	688	144	715	146	1325																													
253_3	726	747	1281	713	726	731	1281	756	1279	86	1336																												
253_4	1299	1320	768	1292	1299	1306	764	1333	762	1179	816	1226																											
253_5	5400	5417	4869	5395	5398	5405	4871	5436	4851	5279	4915	5320	4200																										
253_6	1270	1297	741	1277	1274	1277	741	1310	751	1158	795	1207	45	4187																									
253_7	103	126	656	102	101	112	662	145	656	801	748	809	1374	5471	1357																								
253_8	1296	1311	767	1307	1302	1317	769	1304	775	1192	825	1233	81	4221	54	1387																							
253_9	731	752	1290	738	733	736	1286	745	1280	91	1343	47	1239	5327	1216	812	1222																						
253_10	1276	1309	743	1287	1274	1291	753	1324	757	1172	793	1221	65	4199	44	1357	60	1228																					
352_1	575	594	40	572	573	586	40	609	44	1269	136	1279	756	4858	739	646	767	1288	737																				
352_2	578	591	53	577	578	587	47	596	25	1284	147	1274	763	4860	742	659	772	1281	758	39																		
352_3	12	43	573	25	20	25	571	58	585	718	671	726	1299	5402	1260	103	1296	731	1282	571	580																	
352_4	625	642	98	630	625	632	92	669	98	1325	192	1333	810	4911	791	704	823	1334	797	74	91	623																
352_5	575	588	42	570	573	572	48	605	26	1273	138	1263	752	4843	739	646	769	1276	747	34	29	577	88															
352_6	28	53	581	23	26	39	579	64	587	724	683	720	1291	5398	1276	107	1308	737	1294	573	584	26	627	575													
352_7	561	584	32	566	559	582	28	601	36	1263	128	1269	750	4853	731	644	757	1282	735	18	27	565	78	28	569												
352_8	99	122	658	102	101	96	658	133	654	803	758	799	1380	5473	1347	184	1383	800	1367	660	651	91	690	648	109	650											
352_9	49	64	602	48	45	56	602	51	596	749	700	743	1314	5419	1299	128	1301	738	1309	586	589	49	640	586	49	580	128										
352_10	43	54	600	34	39	48	590	79	590	739	694	733	1302	5405	1295	116	1323	750	1299	580	587	41	634	580	37	574	116	56									
438_1	602	595	71	599	602	611	73	608	53	1298	167	1298	777	4872	764	677	784	1309	776	59	52	602	111	51	598	53	679	619	611								
438_2	47	56	602	54	49	58	604	69	602	745	704	747	1318	5423	1299	130	1307	760	1315	598	605	47	654	592	49	590	130	66	66	579							
438_3	55	76	614	54	57	68	604	95	612	747	702	761	1320	5427	1309	128	1341	766	1313	598	613	53	654	602	55	596	140	78	62	629	82						
438_4	67	88	618	66	65	76	622	103	616	763	716	765	1332	5427	1315	140	1347	772	1325	610	621	67	624	602	67	604	128	84	76	623	78	94					
438_5	574	595	47	571	574	583	43	614	47	1274	139	1274	751	4860	740	655	774	1287	750	41	42	576	93	33	572	29	653	593	583	48	593	603	603				
438_6	33	56	586	24	33	38	588	71	586	729	682	731	1296	5401	1275	106	1315	744	1293	576	587	25	632	576	21	570	108	54	42	589	44	56	58	561			
438_7	60	81	609	51	58	59	621	96	601	758	707	750	1327	5420	1312	125	1340	763	1314	603	606	60	659	587	60	599	133	71	69	616	69	87	81	596	47		
438_8	54	85	599	61	60	65	589	102	617	756	711	760	1328	5434	1314	139	1342	761	1320	613	620	56	663	609	58	607	135	89	71	630	73	85	91	598	53	84	
438_9	595	614	62	586	593	606	66	627	58	1297	158	1285	778	4875	761	672	787	1304	767	58	57	597	112	42	595	48	674	610	610	67	608	628	620	41	584	607	621	
438_10	31	56	582	32	31	44	590	69	584	729	684	737	1300	5401	1285	106	1311	744	1287	574	587	33	632	574	35	570	118	52	46	591	44	62	56	573	24	47	63	586

Table 3.2: Whole genome pairwise comparison of 40 meningococcal carriage isolates isolated from a volunteer V59.1 before filtration. Square colored in red indicates more than 1000 allelic differences while square colored in yellow indicates allelic differences between 500 and 1000.

The analysis of whole genome sequences was likely to include polyaligned CDs variants due to the presence of multicopy CDs, repetitive elements, truncated sequences or the lack of coverage of certain sequences. In order to overcome these obstacles, the genomes were analysed by alternative methods in order to detect only variants in single copy genes. The first approach was to run the GC program that can be accessed through a search navigator of BIGSdb using MC58, N59.1, and FAM18 as refrence genomes. The first run of the program utilized the full-length MC58 genome sequence as the reference genome and 2074 genes were compared overall. 540 variable loci and 174 missing loci were identified. There were 98 loci that were identical in all isolates (i.e. including in the MC58 sequence); meanwhile the number of loci that were exactly the same in all the test genomes except for the reference genome was 447. There were 804 loci that were truncated in some isolates and, finally, there were 11 paralogous loci (loci that are present in more than one copy within the genome).

To distinguish between real variation and spurious variation among 540 variable loci, several filtration processes were run. The first level of filtration was removal of the data for the isolate with ID 20882 (N59.5). This isolate was the most variable among the 40 isolates and its genome sequence appeared to be incomplete because it was truncated or lacking multiple genes relative to the other isolates. This filtration run reduced the number of variable loci from 540 to 87. The second filtration run was achieved by removing genes that had more than one copy in the genome (these genes mostly encode transposes), phase variable genes and CEs. Only two phase variable genes and two genes carrying CE were found and these genes had mismatches in the repeat tract (Appendix 27). The third run of the filtration removed loci that were variable due to a mismatch in the ends or the beginnings of the alignments, but were otherwise not really variable. The total number of filtered loci was 2054 (Table 3.3).

Table 3.3: Different processes used to filter genes from GC approach.

filtration runs	Filtration processes	Number of filtered genes
First run	missing loci	174
Second run	identical loci	545
Third run	truncated loci	804
Fourth run	paralogous loci	11
Fifth run	variable loci due variation in ID 20882 (N59.5)	453
Six run	variable loci with more than one copy	28
Seven run	variable loci due phase variation	2
Eight run	mismatch in alignment	35
Nine run	variable loci due CEs	2
Total		**2054**

This left 15 variable genes found by this GC approach with 14 genes having only one nucleotide difference and one with multiple differences (Figure 3.2).

Figure 3.2: *NMB1926* gene was varied in three positions.

An overall AC program (found in Export dataset in BIGSdb) was run using 1024 genes annotated in BIGSdb at that time. The allele search detected 19 variable loci among the 40 isolates. The first filtration run detected a mismatched alignment for 5 genes. The second filtration run was achieved by running a BLAST search for each variable locus against the complete genome sequence for strain MC58 and excluded 6 loci with more than one copy (See section 2.5.2). The third filtration run was carried out for phase variable genes and one was detected (Table 3.4).

Table 3.4: Different processes used to filter genes from AC approach.

filtration runs	Filtration processes	Filtered genes	Counts of filtered genes
First run	mismatch in alignment	*BACT000065 , NEIS2109, NEIS1901, NEIS0469, NEIS0046*	5
Second run	variable loci with more than one copy	*NEIS0046, NEIS0581, NEIS0297, NEIS0957, NEIS0976, NEIS0904*	6
Third run	variable loci due phase variation	*NEIS2011*	1

Finally, the variable loci identified in the AC were compared with the variable loci found through the GC approach, resulting in two new loci, *NEIS2163* having only one nucleotide difference while *NEIS0035* (*NMB0051*) had multiple differences.

Five loci found by the GC method using MC58 as the reference were confirmed by the AC method (*NMB0086 (NEIS0071)*, *NMB1982 (NEIS1957)*, *NMB0557 (NEIS0498)*, *NMB1577* (*NEIS1497*), and *NMB1926 (NEIS1901)).*

Next, the GC approach was performed again, but now using FAM18 (complete genome sequence for a serogroups C, ST-11 complex/ET-37 complex) and N59.1 (partial genome sequence consisting of all the contigs downloaded from BIGSdb and annotated with Prokka; Blades and Bayliss, unpublished data) as the reference genomes; this identified 3 additional variable genes (*NMC0202 (NEIS0202)*, *NMC1056 (NEIS2537)*, and *NMB0003 (NEIS2141))* and a phase variable gene – *NMC1156* (*NEIS1156*). The variable loci confirmed by all methods were 13, while 7 loci were only found by comparison to one of the reference genomes or by the AC method (Figure 3.3).

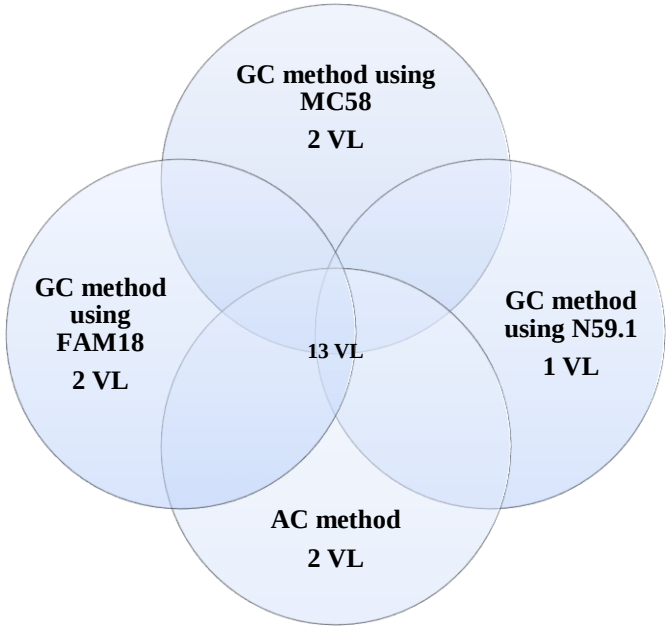


Figure 3.3: Number of variable loci detected using all methods. The number of isolates was 40 from one carrier V59 and one CC with 10 isolates in each time point.

The quality of the data was checked using the FASTQC program. The poor quality data was filtered using the Trimmomatic script. Then the SPAdes script was used to reassemble our isolates using only paired end forward and reverse reads (See chapter two 2.3). After assembly, the variations in *NMC0202* and *NMB1926* genes disappeared due to the location of variability on single end forward and reverse. Surprisingly, the variation in *NMB0846* and *NMB1220* also disappeared presumably because the new assembly was achieved using SPAdes script which produces more accurate assemblies than the Velvet script. Finally, variation in *NMB1561* was due to a point mutation in a repeat tract therefore the variation was checked on read data, there was a mix of reads with and without the varied nucleotide and with very low base phred quality (10) therefore the variation seemed to be due to a sequence error. Overall, therefore, there were 15 variable genes that exhibited variation in these 40 isolates.

The variable genome, formed of SNPs in the CDs regions of the 40 isolates, is presented as a schematic representation of the number of variable mutations across time (Figure 3.4).

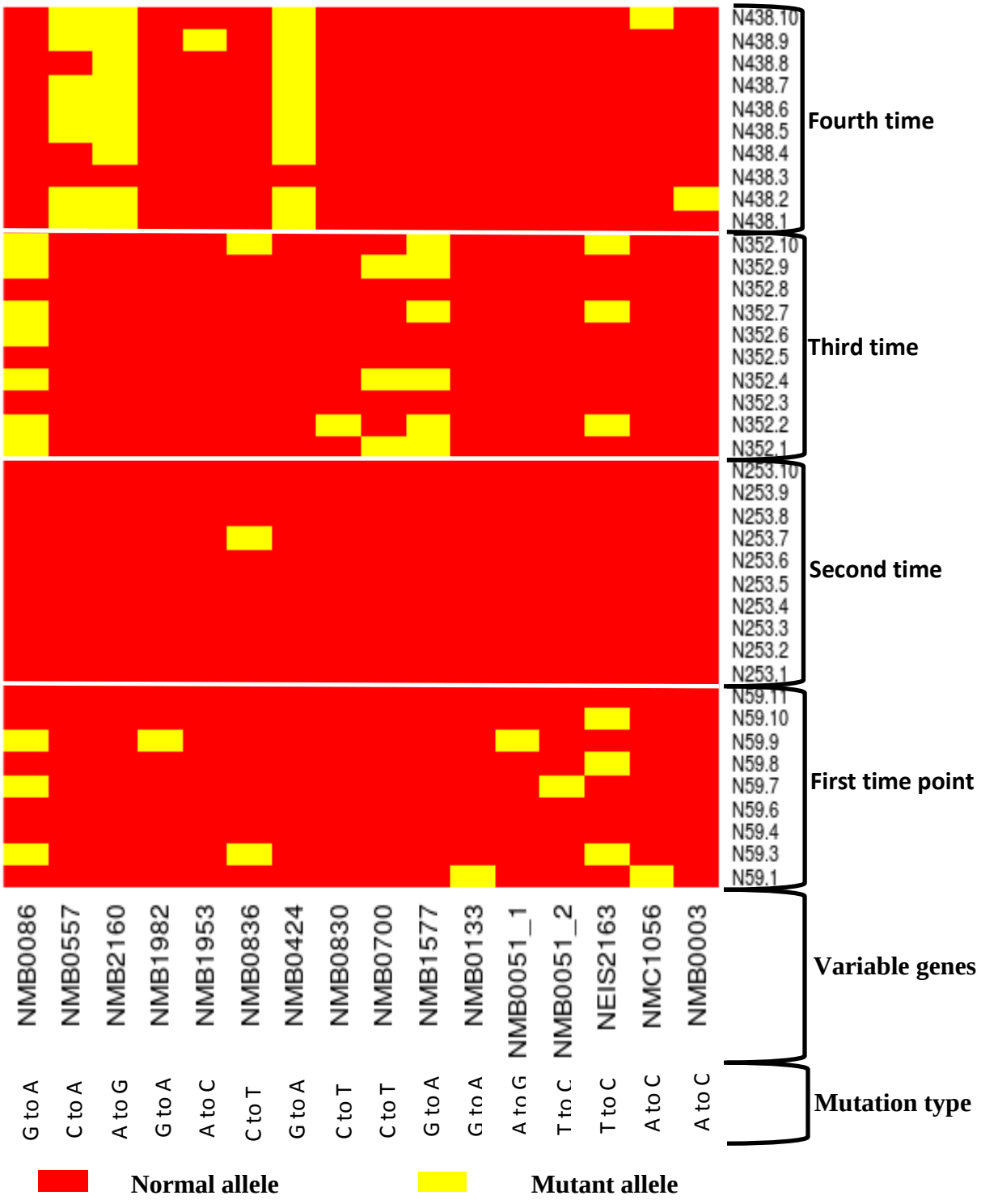


Figure 3.4: Portraying the chronology of the accumulating mutations in the genic regions of 40 isolates from a volunteer V59.1. This schematic shows the variable genes along with their SNPs. The yellow colored boxes depict variable alleles and red colored boxes depict non-variable alleles. The panel on the right shows the isolates with different time points

3.3 Analysis of the functions of specific variants

The potential functional effect of the variability in different genes was examined. The variable loci were classified into two different groups using specific criteria. The first group was assumed to have a strong effect on the persistence of this *N. meningitidis* population and the criteria were:- variable loci varied in multiple isolate; or variation located within functional residues. The second group was deemed less likely to contribute to persistence as the variation was not located within functional residues, or the variation resulted in a synonymous mutation and hence does not cause a change in the amino acids or gene functions or was within pseudogenes.

There were 8 genes that showed variation within a functional domain and showed variation in more than one isolate. Three-dimensional structures were drawn for some of these gene products based on homology modelling and the PyMOL tool (Table 3.5).

Table 3.5: Variation in genes with a high potential to effect persistence of *N. meningitdis*.

Gene	Nucleotide changes	Amino acid change	Position of residue	Number of varied isolates	Protein function
NMB0051	T to C, G to A	Arg to Cys and Arg to His	166 and 321	2	Twitching mobility
NMB0086	G to A	Glu to Lys	84	10	Lipoprotein
NMB0836	C to T	Trp to Arg	411	3	ATP-binding subunit ClpA
NMB0424	G to A	Cys to Tyr	206	10	D-alanine-D-alanine ligase
NMB0557	C to A	His to Thr	98	8	Iron-sulfur cluster insertion
NMB0700	C to T	Phe to Leu	686	3	IgA specific serine endopeptidase
NMB2160	A to G	Thr to Ala	193	10	DNA mismatch repair
NEIS2163	T to C	Ile to Leu	433	5	Capsule region A

There were three genes coding for OMPs or coding for enzymes mediating modification of OMPs. The first of these genes was *NMB0700 (NEIS0651)*, which varied in three isolates and encodes an IgA specific serine endopeptidase. This enzyme has a major role as it cleaves IgA1 (isotype 1). *NMB0700* comprises two domains, the C-terminal domain, which is membrane integrated with the Iga beta core and facilitates substrate access through the outer membrane into the second N-terminal domain. This N-terminal domain contains the functional unit of the enzyme that is represented by its activity as a chymotrypsin-like fold (Klauser *et al.*, 1993). According to the KEGG classification, the function of this gene is related with the hydrolases (acting on peptide bonds), transporters and Type V secretion system. Here, the variation Phenylalanine (F) to Leucine (L) was seen near to the C terminal domain that is exposed on the outer membrane and, therefore, it seems possible that this change alters the shape of the surface –exposed domain of this protein (Figure 3.5).

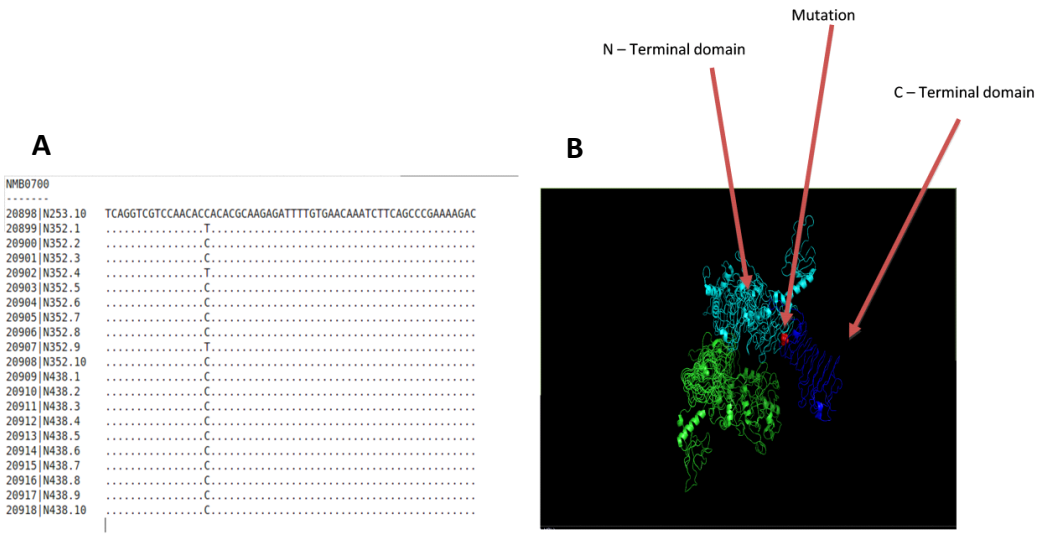


Figure 3.5: Three-dimensional structure of protein encoded by *NMB0700*. Panel A: alignment of the nucleotide sequences of the variable region of *NMB0700*. **Panel B:** three-dimensional structure of the neisserial IgA-specific serine endopeptidase obtained from UniProt (homology modeling) ID (Q9K0B4). This structure contains two domains: C-terminal domain in Blue that carries the mutation with red sphere; and N-terminal domain in green.

NEIS2163 is one of seven genes present in capsule region A of the serogroup Y strains (annotation was taken from BIGSdb website). The variation was detected in five isolates outside the glycosyltransferase_GTB_type super domain and converts Isoleucine (I) to Leucine (L) at residue 499. Therefore, this change may have no obvious effect on persistent of *N. meningitidis*.

The gene *NMB0051 (NEIS0035)* varied in two isolates. It codes for a twitching motility protein. The variation converted an arginine (R) into Cysteine (C) and an arginine (R) into Histidine (H) at residues 166 and 321, respectively. The mutation was located within the Tfp pilus assembly protein (PilU domain) (15-1224 bp). This domain has several functions such as adhesion and twitching motility (Giltner *et al.*, 2012; Gault *et al.*, 2015; Carbonnelle *et al.*, 2005). Therefore, the change in residue may have an effect on colonization and survival of *N. meningitidis* within their host.

There were three genes coding for proteins that have different metabolic function. The *NMB0557 (NEIS0498)* gene varied in eight isolates and encodes an iron- sulfur cluster insertion protein, ErpA. ErpA has a major role in the formation of metallo-sulfur clusters through the conjugation of iron with sulfur metals. This protein contains one domain, PRK13623. This domain spans residues 4-108, and the active site of the domain is rich in Cys residues that could carry the iron-sulfur cluster (Bandyopadhyay *et al.*, 2008). The variation was found in residue 103 and converted Histidine (H) to Threonine (T) (Figure 3.6). As the variation was found in the active site besides the Cys-x-Cys polymer and near the iron-sulfur cluster, therefore this change may impact on persistence of *N. meningitidis* in their host.

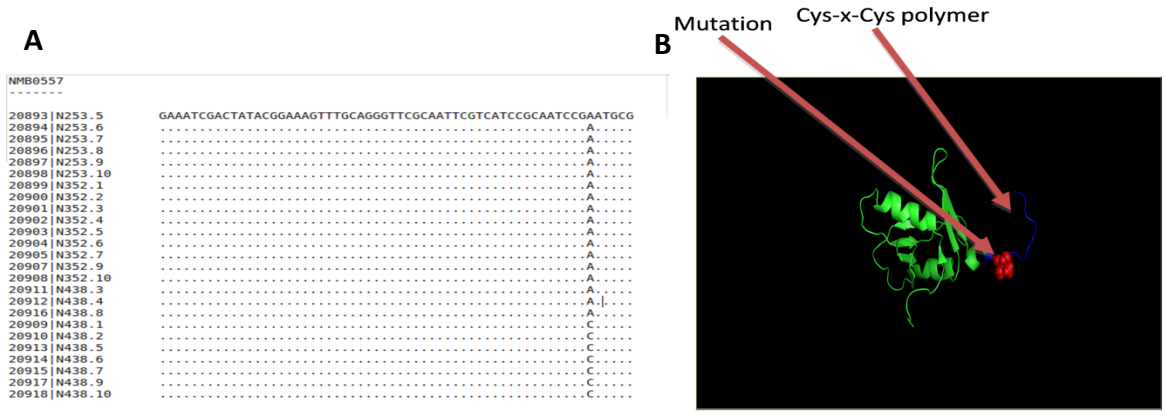


Figure 3.6: Three-dimensional structure of protein encoded by *NMB0557*. Panel A: alignment of the nucleotide sequences of the variable region of *NMB0557*. **Panel B:** three-dimensional structure of the neisserial iron-sulfur cluster insertion protein ErpA obtained from UniProt (homology modeling) ID (Q7DDN1). This structure contains one known functional domain and the mutation was seen in the active site Cys-x-Cys polymer and near to the (Fe-S) cluster indicated by the red sphere.

A variant of *NMB0424* (*NEIS1740*) gene was found in nine isolates. According to KEGG, the function of this gene is related to D-alanine metabolism, peptidoglycan biosynthesis, vancomycin resistance and metabolic pathways. This enzyme consists of two parts; the first part that spans residues 1-100 contains the putative catalytic domain, and the second part, called the Mur ligase domain, spans residues 100-316. The variation was seen in position 206, which is within the ATP binding site and converts a Cysteine (C) to Tyrosine (Y) (ID: Q9K0Y0 UniProt). The variation in this residue may be important for the activity of this enzyme; hence, this change could affect the physiological activities of *N. meningitidis*.

The *NMB0836* (*NEIS0774*) gene varied in three isolates. This gene codes for the ATP-binding subunit ClpA that is found in the bacterial cytosol. This protein has a major role in targeting proteins for degradation by the ClpAP protease. The *NMB0836* works as a chaperone for the assembly and disassembly of proteins in process that requires ATP hydrolysis (ID: Q9JZZ6 UniProt, Rusniok *et al*., 2009). This protein has three domains (217-361, 498-647, and 664-755) that encode for the ATP-dependent peptidase activity and ClpBD2. Variation was located outside these domains and converted a Tryptophan (W) to

an Arginine (R). This alternation may have no effect on protein activity and altering the folding process necessary for protein formation.

The *NMB2160* (*NEIS2138*) gene was found to vary in nine of the isolates. This gene has a major role in recognizing and repairing mismatches during DNA replication and recombination (Davidsen *et al.*, 2007). *NMB2160* codes for the DNA mismatch repair protein MutS that contains two main domains: the DNA binding and the ATPase domains. The protein also consists of two subunits: subunit "A", shown as green in (Figure 3.7), and subunit "B", shown as blue. Subunit A consists of four domains. The variation was connected with a residue which was located on the domain (PRO_0000115112) (1-864 bp) of the A subunit converting Threonine (T) to Alanine (A) and hence is likely to affect the activity of this enzyme. In addition, Richardson and Stojiljkovic (2001) showed that the mutations between residues 39 and 473 in the PRO_0000115112 domain highly affect the protein function. This variation in MutS may lead to a defect in the repair mechanisms and generate variations in other genes hence this gene is the most important variable gene (Davidsen *et al.*, 2007).

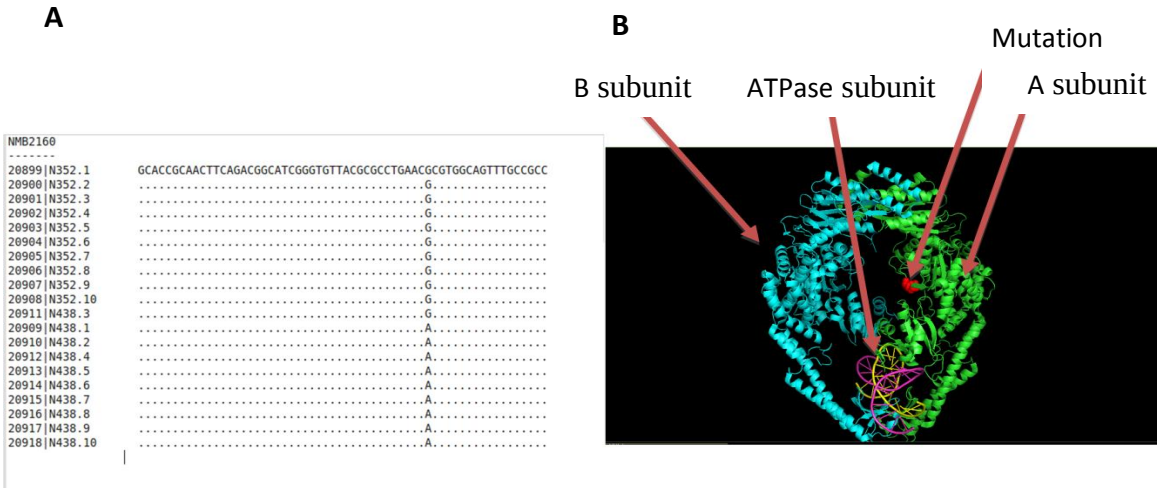


Figure 3.7: Three-dimensional structure of protein encoded by *NMB2160*. Panel A: alignment of the nucleotide sequences of the variable region of *NMB2160*. **Panel B:** three-dimensional structure of the neisserial DNA mismatch repair protein MutS obtained from UniProt (homology modeling) (ID: Q9JX94). This structure contains subunit A shown as green and subunit B shown as blue. Mutation was conducted with PRO_0000115112 domain with red sphere.

The last gene in this first group is *NMB0086* gene. This gene was also found to vary in nine isolates. This gene codes for a lipoprotein. The function of this protein is unknown (ID: Q9K1M3 UniProt).

In order to refine that the SNPs of variable genes with a high potential to effect in persistence of *N. meningitidis* were located in conserved regions between multiple proteins from different species. We performed a BLASTp search against the database in NCBI website. Seven out of eight (Except *NMB0557)* were located within the conserved regions, this may suggest that these SNPs may enroll in functional activities of proteins.

The other seven genes that are predicted to have less significant effects on adaption had variation due to a synonymous change, varied within one isolate, or encoded pseudogenes (Table 3.6).

Table 3.6: Variation in genes with a low potential to effect in persistence of *N. meningitidis*.

Gene	Nucleotides change	Amino acid change	Position of residue	Number of varied isolates	Protein
NMB0003	A to C	Threonine to Proline	77	1	Glutamyl-tRNA ligase
NMB0133	G to A	synonymous			DNA-directed RNA polymerase
NMB0830	C to T	synonymous			Hypothetical protein
NMC1056	A to C	Pseudogene			phage related
NMB1577	G to A	synonymous			Acetolactate synthase isozyme III
NMB1953	A to C	Threonine to Asparagine	110	1	SspA
NMB1982	G to A	Glycine to Serine	558	1	DNA polymerase I

The *NMB1953* (*NEIS1925*) gene varied in one isolate. This gene encodes a protein stringent starvation protein A (SspA), which is induced when the cell is experiencing starvation conditions. This change may affect the ability of *N. meningitidis* cope with environmental stress conditions. *NMB1953* showed variation in residue 107 converting Threonine (T) into Asparagine (N). A conserved surface-exposed pocket that consists of a loop between helix 3 and helix 4 mediates the main function of SspA, which is transcriptional activation of the phage P1 promoter and acid resistance (Hansen *et al.*, 2005). According to KEGG classification, the function of this gene is related to RNA polymerase-associated proteins. The variation was located within the surface-exposed pocket (Figure 3.8).

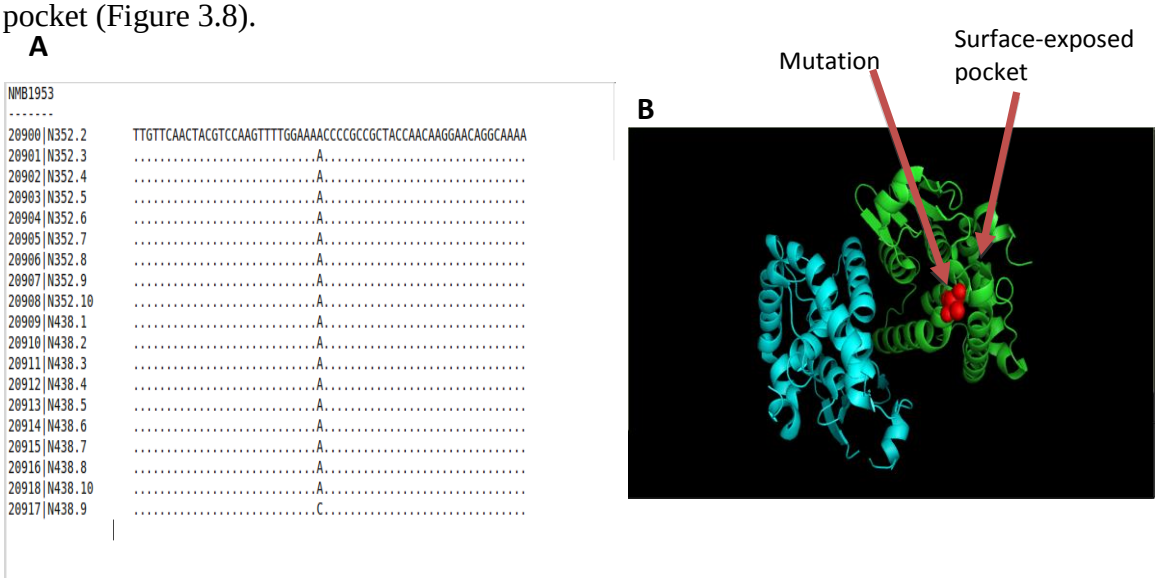


Figure 3.8: Three-dimensional structure of protein encoded by *NMB1953*. Panel A: alignment of the nucleotide sequences of the variable region of *NMB1953*. **Panel B**: three-dimensional structure of stringent starvation protein A obtained from UniProt (homology modeling) (ID: Q9JXN8 UniProt). This structure contains conserved surface-exposed pocket the mutation was seen in the domain with red sphere.

NMB0003 (NEIS2141) encodes a glutamyl-tRNA ligase. This enzyme has three domains, whose functions are ATP binding, glutamyl-tRNA ligase and tRNA binding (ID: Q9K1R6) UniProt). The variability was located in residue 82 converting Threonine (T) to Proline (P). According to the KEGG classification, the function of this gene is related to translation.

The *NMB1982* gene varied in one isolate. This gene codes for a part of the DNA polymerase I (Tettelin *et al.*, 2000). The variation of the DNA polymerase I resulted in transferring amino acid Glycine (G) to Serine (S) at residue position 558. According to the KEGG, the function of this gene is related to DNA replication and nucleotide excision repair.

The last four genes were pseudogenes or have synonymous changes. The *NMC1056* is pseudogenes, it codes for methylase, and putative phage related protein. The *NMB1577*, *NMB0830* and *NMB0133* have synonymous mutation and code for acetolactate synthase isozyme III, hypothetical protein and DNA-dirtced RNA polymerase respectively.

3.4 Analysis of the dynamics of variation in genic regions of the multiple isolates with time of isolation

The level of variability and temporal changes in the distribution of allelic variants for each variable gene were compared within and between time points by counting the number of isolates with a particular variable gene. In the first time point, there were seven variable genes with *NMB*0086 being variable in multiple isolates. At the second time point, one month after the first, there was one variable gene. At the third time point, three months after the first, there were six variable genes with high variability in *NMB*1577, *NMB0086*, *NEIS2163* and *NMB*0700. At the fourth time point, six months after the first, there were 6 variable genes with high variability in *NMB0557*, *NMB*2160 and *NMB0424*. These results also showed that *NMB0424*, *NMB0557*, *NMB0086*, *NMB1577*, *NEIS2163* and *NMB2160* each have a major role in shaping the variation among the 40 isolates and more than 10% of the 40 isolates were variable in one of these genes (Figure 3.9).

Allelic variation was detected as sporadic events occurring in 10% (1/10 colonies) of a specific population for 7 out of 15 variable genes. A partial sweeps of 90% (i.e. 9/10 colonies) were observed for two variable genes (*NMB*2160 and *NMB0424*) in the fourth time point and sweeps of 60-80% for three other genes in the third and fourth time points. The persistence of some variable gene into later time points was detected for *NEIS2163* and *NMB0836.* These variable genes were initially detected in the first and persisted into the third time point. However, in the fourth time point the allelic variants were dominated by a

new lineage. This result indicates that meningococcal carriage is adynamic process with no fixation of allelic variation.

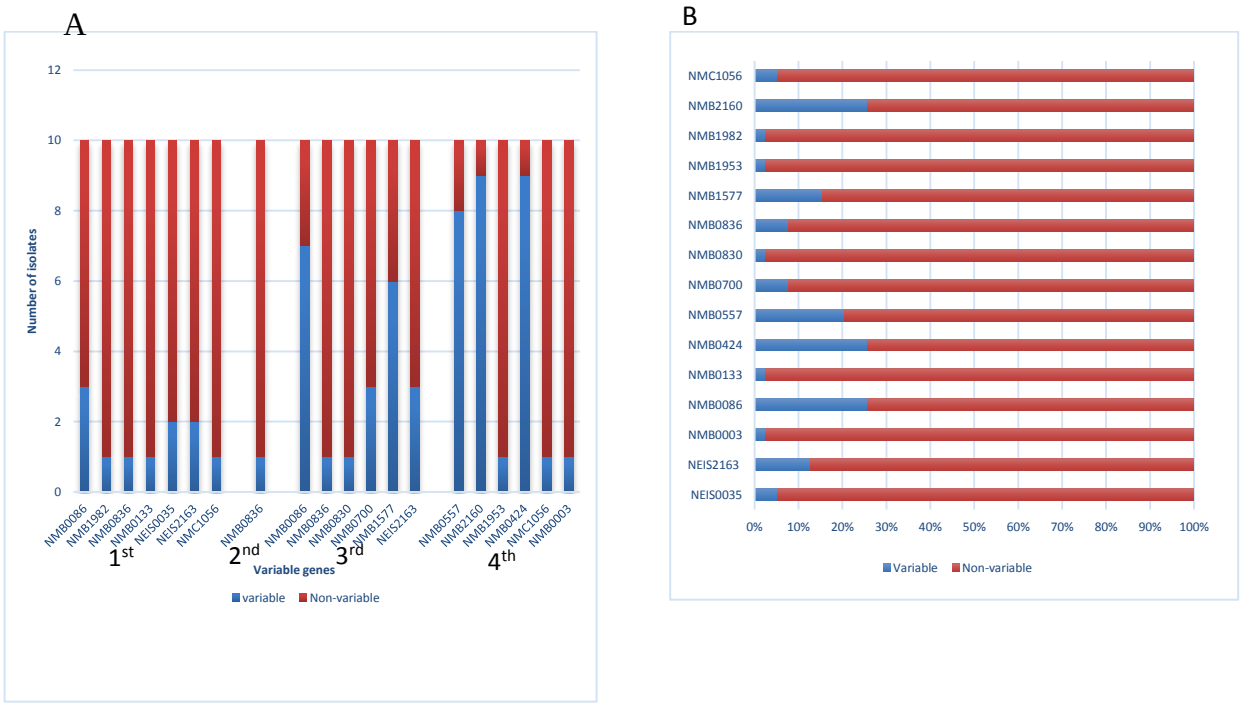


Figure 3.9: Level of variability in the variable genes of 40 isolates from V59. Panel A: This panel shows the identity of variable genes detected in the four time points and the number of variable isolates for each of these genes. **Panel B:** This panel shows the percentage of isolates that have a particular variable gene among the 40 isolates.

3.5 Determination of summary statistics per time point of genic regions for meningococcal genomes

The nucleotide diversity and the mean number of variants per CDs after normalization by the length of CDs were measured for all the variable genes between different strains (Table 3.7). These parameteres were calculated for each time point and the average of these four rates was 0.000006885 and 0.00163 respectively. This indicated that nucleotide diversity and mean number of variants per CDs were very low among all isolates within different time points.

Table 3.7: Listing of different statistical parameters of the genome variation in form of SNPs within the genic regions of 40 isolates.

Time point	Size of total CDs analysed	Total number of variants found	*Mean number of variantes per coding sequence (CDs)	Nucleotide diversity (pi)
First time point	1555629	9	0.00077285	0.0000047
Second time point	1555629	1	0.00042571	0.00000048
Third time point	1555629	19	0.00126278	0.00000903
Fourth time point	1555629	29	0.00406858	0.00001333

*** means total number of variants in each time was divided by the length of variants conding sequences in each time point**

After filtration of spurious variable genes, the pairwise distance matrix was constructed and the allelic differences were the highest for a pairwise genome comparison between isolates in the fourth time point with other isolates mostly in the first and third time points (Table 3.8). This result indicated that the variation was increased with persistence of isolates, however allelic variants were temporal and may not have introduced an advantage for survival of *N.* meningitidis.

As mentioned previously, the distance matrix estimation from whole genome sequences showed that the allelic differences in the second time point was high (See Table 3.2), however in this time point most of the variable genes were present in more than one copy in the genome (these genes mostly encode transposes) and were excluded from the subsequent analysis of genome variation in these 40 isolates. Therefore, as a result of the filtration processes the allelic differences in the second time point were low for pairwise genome comparison of isolates.

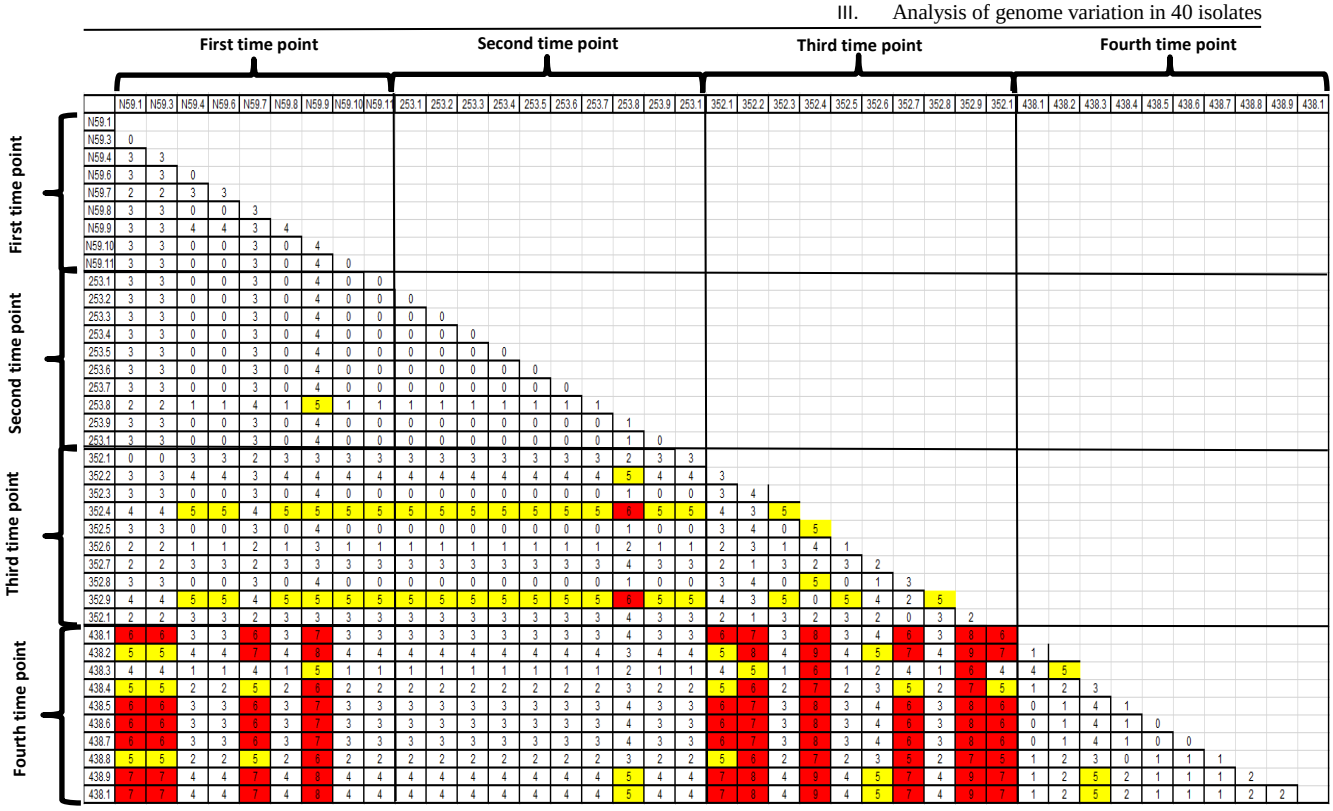


Table 3.8: Whole genome pairwise comparison of 40 meningococcal carriage isolates isolated from a volunteer V59.1 after filtration. Square colored in red indicates the allelic differences were more than five while square colored in yellow indicates the allelic differences were equal to five.

3.6 Identification the type of selection acting on the variable loci

A series of tests were performed to assess the types of selection acting on each gene (Table 3.9) using a codon based Z-test selection model and a bootstrap method. This method computes the average number of synonymous and non-synonymous substitutions to conduct the Z-test and the bootstrap method was used to estimate variance of the difference between these two quantities (See chapter two 2.6.3). None of the P values of genes were significant, however for all variable genes, the average number of non-synonymous changes (1.3) was 5.7 times more than the average number of synonymous changes (0.2) suggestive of the action of diversifying selection, no enough time for the purifying selection to remove deleterious mutation or the combination of both.

Table 3.9: Estimation of selection type for each variable gene. The neutrality, positive and purifying selection were calculated using the MEGA program.

Genes	Gene length	Average number of synonymous	Average number of non-synonymous	P neutrality	Standard error	P value positive selection	Standard error	P value purifying selection	Standard error
NMB0003	1395	0	0.025	0.147	1.052	0.151	1.036	1	-1.04
NEIS0035	1227	0	0.05	0.115	1.588	0.06	1.497	1	-1.47
NMB0086	1017	0	0.25	0.298	1.04	0.16	0.99	1	-1.1
NMB0133	4176	0.025	0	0.34	-0.9	1	-0.9	0.167	0.96
NMB0424	915	0	0.225	0.33	0.9	0.15	1.029	1	-0.9
NMB0557	339	0	0.2	0.297	1.03	0.158	1.006	1	-1.03
NMB0700	5448	0	0.075	0.3	1.03	0.14	1.06	1	-1.06
NMB0830	1011	0.025	0	0.3	-1.03	1	-1.09	0.14	1.06
NMB0836	2262	0	0.075	0.31	1.009	0.157	1.012	1	-1.07
NMB1577	1728	0.15	0	0.3	-1.027	1	-1.028	0.151	1.035
NMB1953	606	0	0.025	0.3	1.01	0.15	1.02	1	-1.029
NMB1982	2797	0	0.025	0.26	1.1	1.34	1.12	1	-1.1
NMB2160	2625	0	0.225	0.3	1.002	0.15	1.01	1	-1.05
NMC1056	629	0.025	0	0.32	-0.9	1	-1	0.16	0.9
NEIS2163	3114	0	0.125	0.1	1.3	0.09	1.2	1	-1.3

Average number of synonymous: number of synonymous divided by 40, average number of non-synonymous: number of non-synonymous divided by 40

3.7 Measurement of the mean nucleotide diversity for the variable loci of carrier V59

The MEGA program was also used for the calculation of the mean diversity between populations using the number of nucleotide differences as a model and bootstrap as a method. Nucleotide diversity reflects the degree of polymorphism among different isolates in different time points (See chapter two 2.7.1). This analysis was used to determine if the amount of nucleotide diversity increases with host persistence (i.e. length of carriage).

Short –term (first-second) population was not estimated due the only one variable gene detected in the second time point. For the populations of the first and third time points, mean diversity was 2.6 with a standard error (S.E.) of 0.7 while for the first and fourth time points the value was 3.2 with a S.E. of 0.9. Comparing diversity between medium-term (first and third) against long-term (first and fourth) populations generated a P value of 0.03, which is statistically significant.

3.8 Construction of a haplotype network for the variable genes of carrier V59

To improve visualization of the genome variation, the variation was converted into a haplotype network-using program HapView. There are 15 different haplotypes but 17 individuals shared the same haplotype (nine from the second time point, five from the first time point and two from the third time point). The isolates from the fourth time point were clustered on one end of a separate branch while the isolates from the third time point were mostly clustered on another end of a separate branch. The isolates from the first and second time points were dispersed randomly (Figure 3.10). An important observation was seen from the haplotype construction was that haplotypes were temporal; an example could be explained through the fact that allelic variants appeared in third time but disappeared in fourth time point. In addition, identical or near identical isolates were detected during persistence.

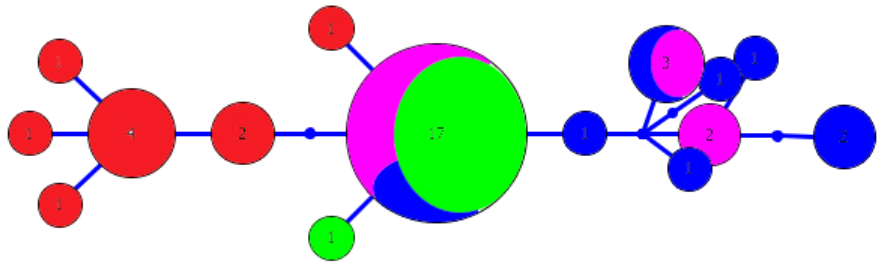


Figure 3.10: Haplotype network of 15 variable genes for 40 isolates from four isolation times from carrier V59. In this tree, each ball represents a single haplotype. The different haplotypes are connected by a line that refers to the branches in the phylogenetic tree. Some branches have a dot that represents that variation was caused by more than one SNP for a particular strain. The purple color: first time point. Green color: second time point. Blue color: third time point. Red color: fourth time point.

3.9 Perl scripts for extracting and manipulating intergenic DNA sequence

The extract_IGR.pl script was written to extract IGRs from whole genomes of 40 isolates. The script was first run on the genome of strain MC58. The overall number of IGRs detected for MC58 was 1952; there were 928 IGRs on the forward strand while there were 1024 IGRs on the reverse strands. The script was also run on N59.1, there were 1880 IGRs in total with 978 on the forward strand and 902 on the reverse strand.

The extracted IGRs of N59.1 were used in BIGSdb to BLAST search and to extract the IGRs of the other 39 isolates. The variable IGRs for the 10 isolates of a particular time point were aligned using muscle (See section 2.5.3).

Prior to filtration, putative variable IGRs were located in 129 loci for the 40 isolates. Truncated sequences were found in eight loci while three loci were excluded due to a mismatch at the end of the alignment. Interestingly, the majority of variation was located within IGRs containing repetitive elements, especially CE with 71 intergenic loci, NIME with 28 loci and REP4 with one locus, while there was no variation associated with REP

elements 1, 2, 3 and 5. As repetitive elements have a strong potential for assembly errors the variation was in these loci filtered (See chapter six). Eighteen out of the 129 intergenic loci, however, had variation due to allelic changes in non-repetitive elements.

As mentioned previously, poor data quality was filtered using Trimmomatic script (See chapter two 2.3). It was observed that the variation disappeared in the IGRs of *NEIS0233*, *NEIS0784*, and *NEIS2833* genes when the assembly was carried out on only pair end read data. Overall, therefore, there were 15 IGRs that exhibited variation in 40 isolates (Table 3.10). To determine whether this variation was located in promoters, promoter positions were predicted using the BPROM program. None of the variable nucleotides were predicted to be located within the promoter regions.

Table 3.10: The variable IGRs found in 40 isolates. The IGRs in red were filtered due to poor assembly.

Intergenic regions	Allelic variation	Number of changes	Isolates
NEIS0233	C to T	1	N438.2
NEIS0321	T to C	1	N253.4
NEIS0384	A to G	6	N59.1,7,9, N253.3,N352.2,N438.4
NEIS0422	C to T,G to A,A to G	10	N59.11,N253.3,10,N352.4,5,9,10,N438.2,4,10
NEIS0595	G to A,T to C	3	N59.6,N253.2,N438.9
NEIS0784	G to A	1	N438.7
NEIS0825	T to A	11	N59.1,9,10,N253.1,2,10,N352.1,3,7,N438.3,8
	G to T	1	N253.6
NEIS0905	T to C	1	N253.1
NEIS1018	A to G	4	N253.3,4,N352.,6,,10
	C to T	4	N253.5,N352.4,8,N438.1
NEIS1448	T to G	1	N438.9
NEIS1576	G to A	1	N59.11
NEIS1669	C to T	1	N352.8
NEIS1745	A to C	1	N438.9
NEIS1768	G to T	4	N253.6,N352.10,N438.4,5
	T to C	2	N253.1,N438.3
NEIS1791	C to T	3	N352.2,7,10
NEIS1959	C to T	1	N352.6
NEIS2101	C to T	1	N438.2
NEIS2833	C to T	1	N352.2

Variable genome in form of SNPs in the IGRs regions of 40 isolates within different time points also presented as a schematic representation of number of variable mutations across time (Figure 3.11).

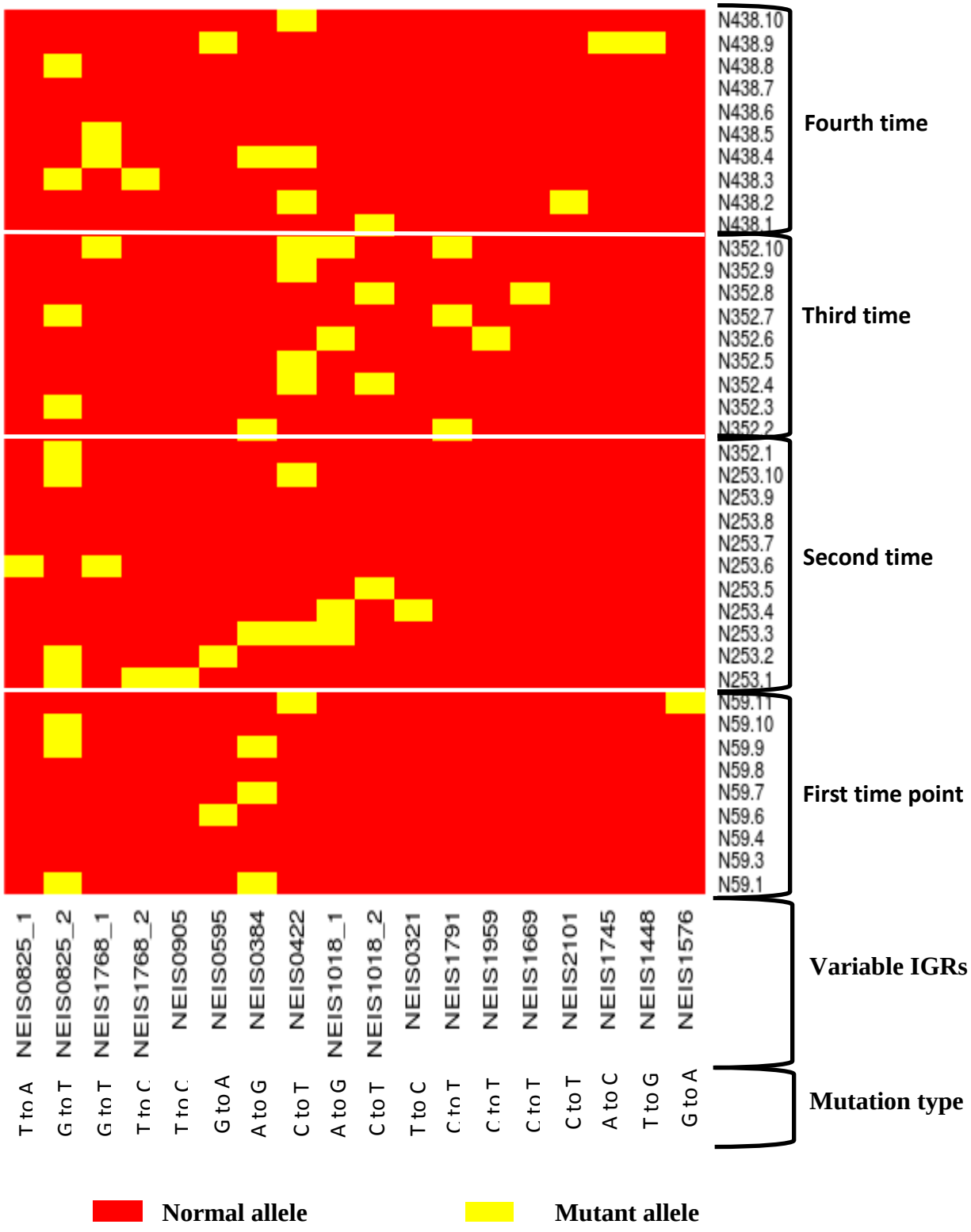


Figure 3.11: Schematic representation of the chronology of the accumulating mutations in the IGRs regions of 40 isolates from volunteer V59. This schematic shows the variable IGRs along with their SNPs with yellow colored boxes depicteing variable alleles and red colored boxes depicting non-variable alleles. The panel on the right shows the isolates with different time points.

3.10 Analysis of the temporal variation in the IGRs of the 40 isolates

The level of variability for each variable IGR was compared within each time point. In the first time point, there were five variable IGRs with *NEIS0825* and *NEIS0384* being variable in multiple isolates. In the second time point, there were seven variable IGRs, again with high variability in *NEIS0825* in addition to *NEIS1018*. In the third time point, there were nine variable IGRs, again with high variability in *NEIS0825* and *NEIS1018* in addition to *NEIS0422* and *NEIS1791*. In the fourth time point, there were also nine variable IGRs again with high variability in *NEIS0825* and *NEIS0422* in addition to *NEIS1768*. The percentage of overall contribution of the different IGRs to the variability in the 40 isolates of *N. meningitidis* is illustrated in (Figure 3.12). The result shows that *NEIS0825, NEIS0384, NEIS1018, NEIS0422,* and *NEIS1768* each have a major role in shaping the variation in 40 isolates affecting more than 10% of the population.

As seen for genic regions, a dynamic population was also seen for the allelic variants of IGRs. Sporadic variation (1/10 colonies) was detected in 11 out of 15 IGRs while partial sweeps of 40% (i.e. 4/10 colonies) were observed for *NEIS1018*, *NEIS0825* and *NEIS0422* IGRs in the second and third time points. However, the expansion of the same IGR within different points was detected in *NEIS0384*, *NEIS0422*, and *NEIS0825*. The potential function of IGRs were examined and the result showed that all IGRs were located in tail to tail organization between two adjacent loci or upstream of nearby genes coding for hypothetical proteins, however, two IGRs (*NEIS0384* and *NEIS1959*) located upstream of nearby genes coded for lipoprotein signal peptidase and IgA specific serine endopeptidase respectively.

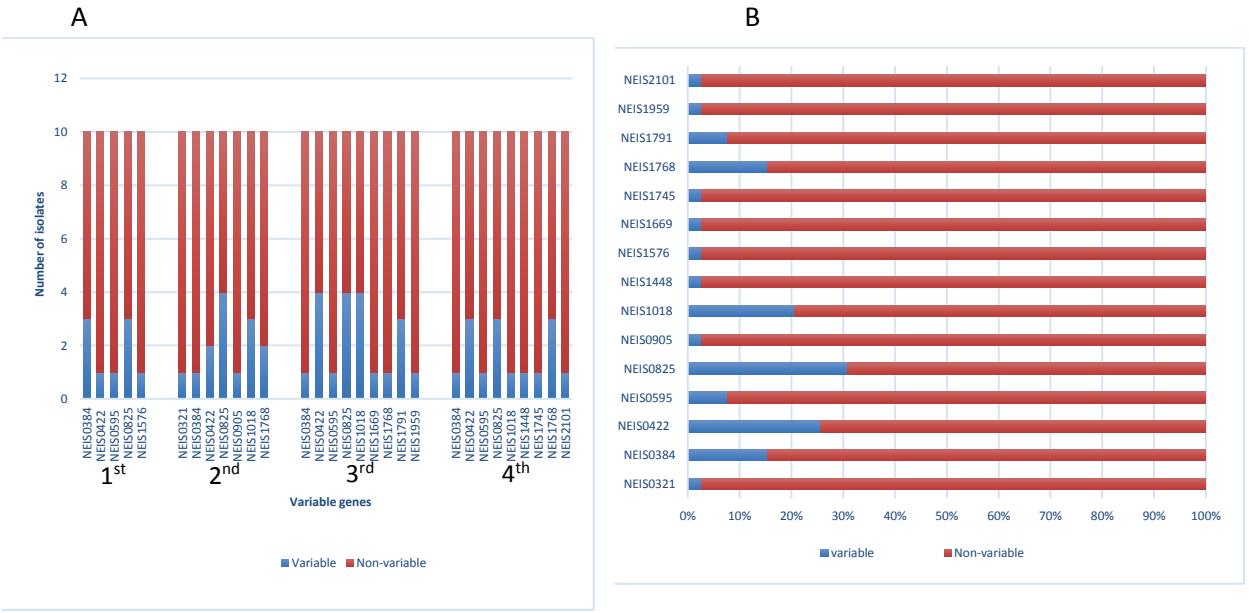


Figure 3.12: Detecting the level of variability of variable IGRs in the 40 isolates. Panel A: identity of variable intergenic detected in different time points. The number of variable isolates for each intergenic in each time point was shown. **Panel B:** percentage of isolates that have a particular variable intergenic among the 40 isolates from carrier V59 of *N. meningitidis.*

3.11 Identification of summary statistics per time point for complete variable loci

The nucleotide diversity and the mean number of variants per IGR region after normalization by the length of IGRs were calculated for each time point. The average for the four time points was 0.0000066 and 0.001195 (Table 3.11). Again, the nucleotide diversity and the mean number of variants per IGR regions after normalization by the length of IGRs were low.

Table 3.11: Listing of different statistical parameters of the genome variation within the IGR regions of 40 isolates.

Time point	Size of total IGRs analysed	Total number of variants found	Mean number of variantes per IGR sequence	Nucleotide diversity (pi)
First time point	422040	9	0.001386108	0.00000426
Second time point	422040	15	0.001199328	0.00000711
Third time point	422040	18	0.001229508	0.00000855
Fourth time point	422040	14	0.000965118	0.00000662

*** means total number of variants in each time was divided by the length of variants conding sequences in each time point**

3.12 Haplotype network construction for IGRs of 40 isolates

A haplotype network was constructed for the 15 IGRs of the 40 isolates. There were 26 different haplotypes. Comparing with the haplotype network produced from the genic region, the intergenic network showed that all the isolates were dispersed randomly without any clusters for isolates from specific time points. Two further points are that there is no increase in diversity with time and that the number of haplotypes for the IGRs is higher than the number of haplotypes found in genic regions (Figure 3.13). A significant finding was that temporal haplotypes and the expansion of the same haplotypes were detected during persistence.

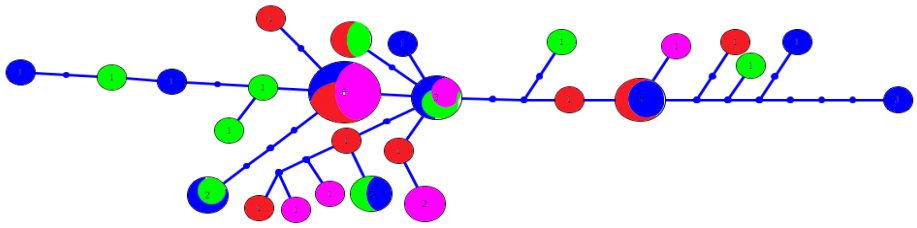


Figure 3.13: Haplotype network of 15 variable IGRs for 40 isolates from four isolation times from carrier V59. The pink color: first time point. Green color: second time point. Blue color: third time point. Red color: fourth time point.

3.13 Determination of the mean of nucleotide diversity and statistical testing of the temporal changes in the intergenic loci of V59 carriage isolates

As mentioned previously in the genic regions, the mean diversity between populations of variable IGRs was estimated for concatenated IGRs of isolates from the first and third time points (See chapter two 2.7.1) and the first and fourth time points. For the first and third time points, mean diversity was 4.7 with a S.E. of 0.85 while for the first and fourth time points the value was 4.0 with a S.E. of 0.84. Comparing diversity between these two populations generated a P value of 0.01, which is statistically significant for the compared populations (first and third time point) against (first and fourth time point). The analysis of diversity in IGRs showed that there was no relation between increasing diversity with increasing persistence of *N. meningitidis.*

3.14 Summary of main findings

A series of analyses were conducted to detect and investigate variability in a pool of 40 meningococcal carriage isolates from a single carrier. Mutation usually occurs as a series of constant events over time. In contrast, recombination incorporates DNA segments with many varied genes or a particular gene with many SNPs and so variation has an irregular distribution over time. The initial analysis showed that the varied loci were mainly occurred due to a single nucleotide polymorphism, therefore the assumption was drawn that variation was caused by de novo mutation in the genic and IGRs sequences. Our analyses showed that the number of SNPs increased from 11 at the first time point to 29 in genic region after six months. In addition, mean diversity was significantly different between medium (first-third isolates) and long-term (first-fourth isolates) carriage with p-value (0.03). This may suggest there was a correlation between increasing variation with increasing persistence of isolates in genic regions. On the contrary, the number of SNPs increased slightly from 9 at the first time point to 15 in IGRs after six months. Moreover, mean diversity between first and fourth time point did not exhibit a correlation between variations in IGRs and persistence of isolates. Furthermore, in spite of a low number of haplotypes in genic regions (15), the isolates from the fourth time point were clustered on one end of a separate branch while the isolates from the third time point were mostly clustered on another end of a separate branch. The IGR network, haplotypes were higher (26) but were dispersed randomly without any clusters in specific time points. Increasing variation with persistence of isolates for genic versus IGRs may indicate the role of genic regions in persistence of meningococcal population for 6 months.

The analysis also showed that the average number of non-synonymous changes was 5.7 times more than the average number of synonymous changes for all variable genes; this may due the action of diversifying selection, no enough time for the purifying selection to remove deleterious mutation or the combination of both.

Another key finding was that dynamic temporal fluctuations were detected within the meningococcal population with specific allelic variants and haplotypes appearing and disappearing during persistence. As an example, dynamic temporal fluctuation was observed through appearing the haplotypes in the third and disappearing in the fourth time

points (Figure 3.10). This indicates that the haplotypes were temporal and may introduce no advantages for persistence of *N. meningitidis*.

Regarding the function, in the genic regions, the variable genes assumed to have a strong effect on persistence of *N. meningitidis*, included genes coding for outer membrane proteins or for enzymes which mediate modification of outer membrane structures, these are *NMB0700* (IgA specific serine, endopeptidase), *NEIS2163* (Capsule region A in serogroup Y) and *NMB0051* (Twitching mobility). These changes may alter the shape of the protein, or structure so that it cannot be recognized by the antibodies mediated adaptive immune response. In comparison with the current study, studies of genetic variation as a mechanism to overcome immune selection associated with herd immunity have been investigated frequently at a population level. Meningococcal isolates were collected in Ghana and Burkina Faso with time points, but showed no point mutations or insertional/deletion (indel) events in genes encoding OMPs. However, ST7 and ST2859 strains were different in the simple sequence repeat (SSR) of *opa* genes (Huber, 2011). On the other hand, a study achieved by Krauland *et al.* (2012) on a population of disease isolates of *N. meningitidis* collected between two different time points (June 1999 and October 1999) showed that in the later population serogroup Y disease had emerged due to variation in pilin and iron acquisition of OMPs.

In general, the genes that have a high level of variability within the 40 isolates code mainly for metabolic enzymes. Epidemiological study showed that differences in meningococcal transmission fitness correlated with variation in metabolism genes (Buckee *et al.*, 2008).

Interestingly, variation was found in *NMB2160* encoding MutS which is involved in (Repairing mismatch during DNA replication and recombination). The change in MutS may lead to a less effective damage repair mechanism and generation of variation in other genes. One study showed that mutation rates were highly increased in epidemic serogroup A isolates due to defects in mismatch repair pathways caused by mutations in mismatch repair genes. This work suggested that hypermutability may help in transmission of this pathogen (Richardson *et al.*, 2002).

The *NMB0424*, *NMB0557*, *NMB0086*, *NMB1577*, *NEIS2163*, and *NMB2160* genes have shaped the variability within the 40 isolates and hence these genes may have essential physiological activities that facilitate host persistence of Neisseria. However, the haplotypes may be temporary and so have no fitness advantage. In IGRs, the *NEIS*0825, *NEIS0384*, *NEIS1018*, *NEIS0422* and *NEIS1768* have shaped the variability within the 40 isolates, however the positions of variation in overall IGRs were located outside the core promoter as predicted using promoter prediction program. Nevertheless, the prediction from promoter programs may be inaccurate due to limitations in the algorithms therefore the variation needs to be analyzed practically to determine if the variation affects on gene expression. Interestingly, most of the variations in IGRs were found in the different repeat elements (See chapter six) suggesting that these elements may have the greatest impact on host persistence.

Chapter 4: Analysis of genome variation in persistent isolates from multiple carriers

4.1 Introduction

In the previous chapter (Chapter three), the genome comparison study was carried out on isolates from an individual carrier, the complete picture of isolate evolution within their host can be inferred through genome comparison of persistent isolates from multiple carriers and that is the aim of this chapter. Therefore, the current study was conducted with 25 pairs of meningococcal carriage isolates belonging to several CCs collected from different volunteers at two different time points: 0 and 2/3 months and 0 and 5/6 months (Table 4.1). As mentioned in chapter three, Illumina Hiseq, Velvet and Prokka were used for sequencing, assembling and annotating the genome sequences of these isolates. The novelty of the study is the analysis of genetic variation using genome comparison on pairs of isolates with two different periods of carriage for inferring the potential evolution of the isolates within their host and for finding the patterns and dynamics of variation in such regions that may have a role in persistence of isolates for months.

To reach to our aims, this chapter analyzes the genome variation on genic regions and IGRs by mutation or recombination. As mentioned in the chapter three, mutation usually occurs as a series of constant events with time. In contrast, recombination incorporates DNA segments with many variable genes or a particular gene with many SNPs so that it is important to examine genome variation at the gene and SNPs level to draw a complete picture of the evolutionary shape of genome variation.

This chapter starts by comparing the number of variable genes and IGRs among pairs of isolates belonging to different CCs. Genome variation was then analyzed for different time points (short and long periods of carriages).

Secondly, at the nucleotide level; genome variation was analyzed by estimating nucleotide diversity for each isolate pair and by estimation of a diversity score for each functional scheme.

Finally, recombination was detected for all the genic and IGRs of all 25 pair isolates using a sliding window approach. Patches of recombination containing three adjacent loci were deeply characterized.

Table 4.1: Listing the name, ID, volunteer number and time point of isolation of 25 pair isolates of *N.* meningitidis.

Volunteer	Time (months)	Isolate Pair	Isolate IDs
V43	2	N241-N349	16975_17001
V51	6	N51-N424	16976_17002
V52	3	N52-N342	16977_17003
V54	3	N54-N343	16978_17004
V58	6	N58-N429	16979_17005
V59	6	N59-N438	16980_17006
V88	6	N88-N449	16981_17007
V138	3	N138-N331	16982_17008
V64	3	N64-N348	16983_17009
V117	6	N117-N417	16984_17010
V124	3	N124-N336	16985_17011
V128	6	N128-N420	16986_17012
V69	5	N258-N431	16987_17013
V93	2	N264-N359	16988_17014
V96	5	N259-N445	16989_17015
V222	6	N222-N459	16990_17016
V114	3	N114-N330	16991_17017
V134	3	N134-N333	16992_17018
V185	6	N185-N456	16993_16994_17019_17020
V82	6	N262-N446	16995_17021
V73	6	N73-N450	16996_17022
V199	3	N199-N378	16998_17024
V176	6	N176-N408	16999_17025
V188	6	N188-N462	16997_17023
V86	6	N86-N447	17000_17026

4.2 Analysis of variability of the genic regions of isolates belonging to different CCs

The comparative analysis of the genomes was achieved by both comparisons to reference genomes using GC and AC as available on BIGSdb. Each analysis involved pairwise genome comparison between isolates from each volunteer. These analyses identified a number of identical loci for all compared isolates and a number of loci that were identical with each other but different from the reference genome, as well as truncated and missing loci. In the first instance, the MC58 genome was used as a reference genome, N59.1 was also used as reference genomes to pick up genes that were missing in the MC58 genome.

A phylogenetic tree was drawn from the whole genome sequence for all the pairs of isolates belonging to 25 pairs of isolates using ParSnp and Gingr packges. The variable genomes in form of SNPs were used to construct a tree after aligning entire genomes sequences of 25 pairs of isolates. The percentage of core genome alignment among all sequences of 25 pairs of isolates using ParSnp was 78.2% (1625916 bps out of 2079177 bps).Again Gingr and FigTree programs were used to visualize and coloring the tree and the well supported tree showed that isolates clustered according to CCs while paired isolates from the same individual were closely related (Figure 4.1).

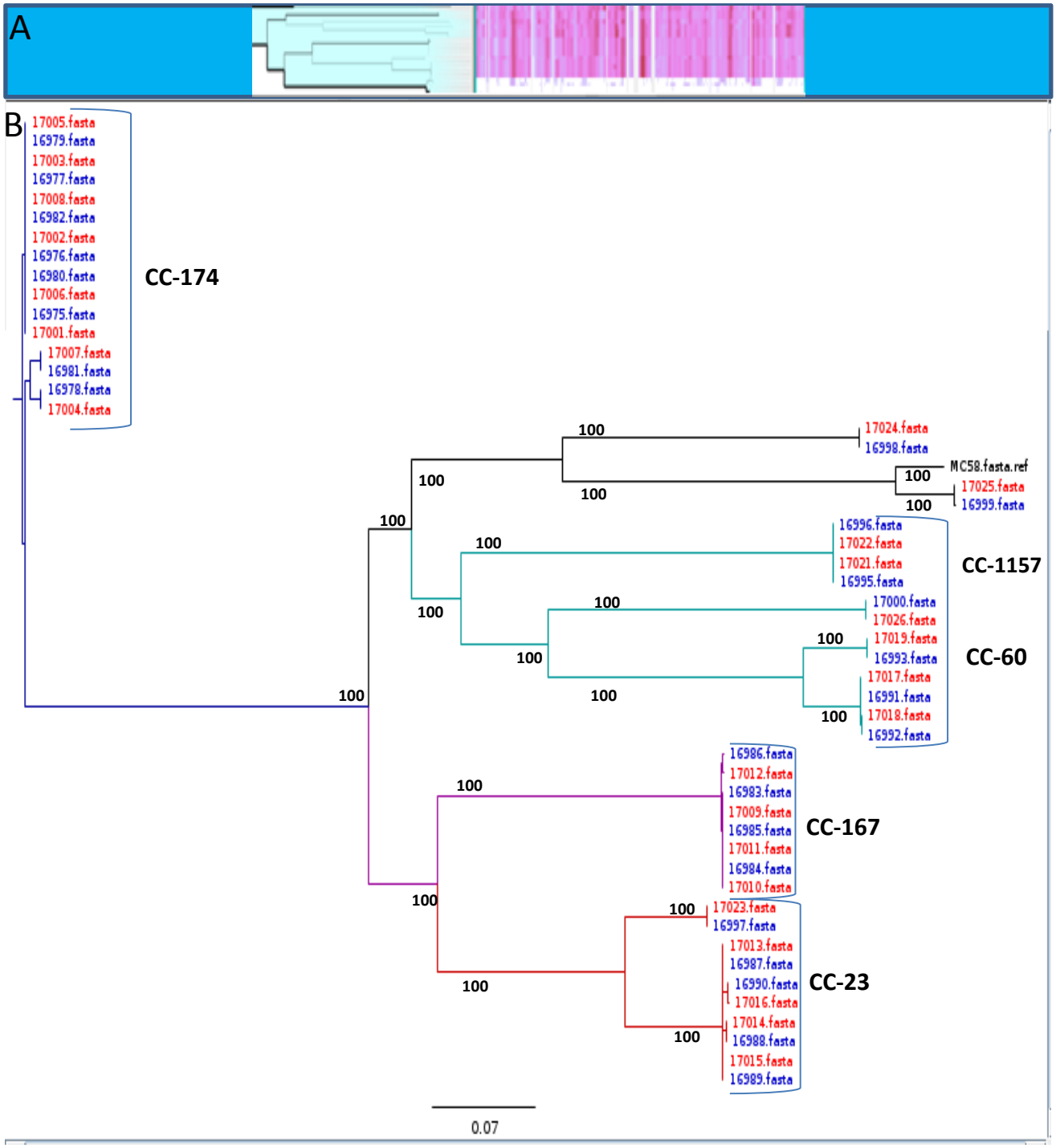


Figure 4.1: Phylogeny of 25 pairs carriage isolates constructed from whole genome alignments using ParSnp package. Panel A: schematic representation of phylogeny along with the genome variation depicted with pink line along each isolate in the phylogeny. Panel B: schematic representation of phylogeny. Blue color depicted isolates in initial time point and red color depicted isolates in latter time point. The branch of isolates belongs to each CC depicted with specific color.

4.3 Filtration process to distinguish between real variation and spurious variation for the genic regions of 25 pair isolates

Filtration processes were carried out for all the putative variable loci as errors in detecting variation were noted due to both the detection method and inaccurate genome assembly.

The filtration processes on the genic regions were as follows; firstly removing the loci that have more than one copy in the genome; secondly, removing the loci that were truncated or have a poor alignment; thirdly, removing the loci that varied within a repetitive sequence or have short sequences (i.e. <100bps).

Most loci were filtered due to a poor alignment or were found with more than one copy in the genome (mainly coding for a transposase). Spurious variation was highest in loci containing SSR repeats among different types of repeat tracts in *N.meningitidis* (Table 4.2).

Table 4.2: Number of variable genes filtered using different filtration processes.

CCs	More than one copy	CE	SSR	NIME	REP 1,2,3 ,4,5	Poor alignment	Truncated sequence	Short sequence chunk
CC-174	1-22	1	4	0	0	19-36	1-5	0
CC-167	0-3	0	4	0	0	27-33	0-2	1
CC-23	3-8	0	2	0	0	26-29	1-7	0
CC-60	4-11	0	1	0	0	16-65	0-7	0
CC-1157-32-269	1-6	0	5	0	0	2-33	0-1	2

FASTQC program was also used for checking the quality of data. As mentioned in chapter two (2.3), Trimmomatic script was used for filtering the poor data quality. Then the SPAdes script was used to reassemble our isolates using only pair end forward and pair end reverse.

The variation in those genes located on single end forward and reverse was removed by this assembly. The unique numbers of genes filtered due to poor quality data were 20 genes; CC-174 has the highest with 8 out of 23 (Table 4.3).

Table 4.3: List of variable genes filtered due to poor quality data in 25 pair isolates.

Variable loci	Time	Isolate Pair	Isolate Pair (IDs)	Volunteer	CCs
NMB0836	6	N51-N424	16976_17002	V51	CC-174
NMB0701 *NMB1410* *NMB1529* *NMB1733*	3	N54-N343	16978_17004	V54	
NMB0836 *NMB1812* *NMB0533*	6	N88-N449	16981_17007	V88	
NMB0846 *NMB1157* *NMB1883*	3	N64-N348	16983_17009	V64	CC-167
NMB0648	6	N128-N420	16986_17012	V128	
NMB1641 *NMB1235*	2	N264-N359	16988_17014	V93	CC-23
NMB2001 *NMB0281*	5	N259-N445	16989_17015	V96	
NMB1880	6	N262-N446	16995_17021	V82	CC-60
NMB0977 *NMB1464* *NMB1880* *NMB1857*	6	N73-N450	16996_17022	V73	CC-1157-32-269
NMB1857	6	N176-N408	16999_17025	V176	
NMB0319	6	N86-N447	17000_17026	V86	

4.4 Variation in the genic regions of 25 pair isolates

For CC-174, CC-167, CC-23, CC-60 and CC-1157-32-269, the number of real variable genes were 12, 11, 6, 10 and 4 respectively according to the AC method while there were 25, 35, 9, 16 and 3 real variable genes, according to the GC method. The total numbers of real variable genes identified according to AC and GC were 37, 46, 15, 26 and 7 respectively for each CC (Table 4.4).

Table 4.4: The overall variable genes captured by AC and GC methods.

CCs	Number of isolates	Variable genes captured by AC	Variable genes captured by GC	Total number of variable genes
CC-174	8	12	25	37
CC-167	4	11	35	46
CC-23	5	6	9	15
CC-60	4	10	16	26
CC-1157-32-269	4	4	3	7

The total number of genes detected by GC and AC were 88 and 43 respectively. In general, the number of variable genes captured by the GC method was higher than identified by AC due to an incomplete classification of meningococcal genes in the BIGSdb database.

The number of variable loci was higher in CC-174 and CC-167 compared with other CCs (Figure 4.2). The high number of variable loci in carrier V54 (i.e. isolate pair 16978 and 17004) is the cause of the high number of variable loci within CC-174 while similarly

carrier V128 (i.e. isolate pair 16986 and 17012) has a high number of variable loci and is responsible for the high number of variable loci within CC-167. These isolates were predicted to be subject to multiple recombination events with DNA from isolates of other CCs.

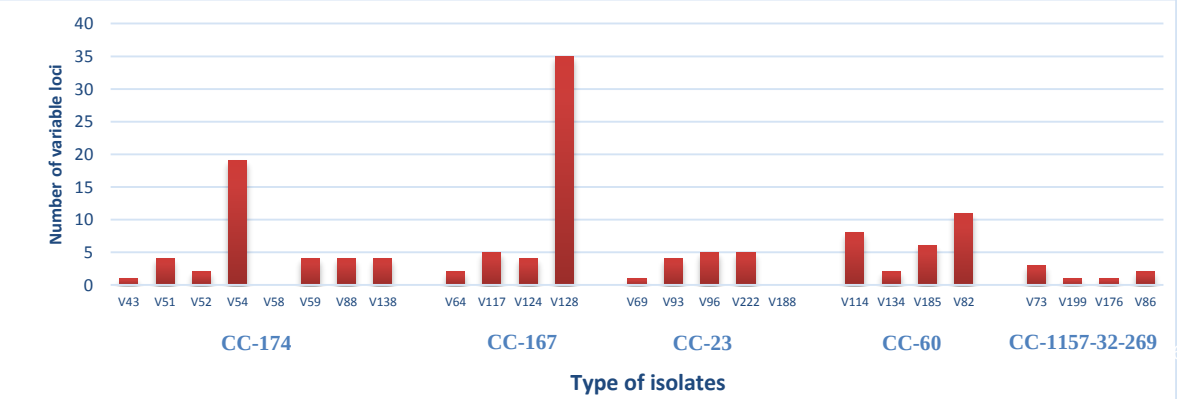


Figure 4.2: The number of real variable genes for each pair of isolates detected by GC and AC method among different CCs.

To more precisely illustrate the effect of variation, the mean number of variants per coding sequence (CDs) was detected, then it was normalized by the length of the CDs. In addition, the number of nucleotide differences was used to estimate the nucleotide diversity. After normalization of the mean number of variants per coding sequence (CDs), the isolate from V128 (16986-17012), V82 (16995-17021) and V222 (16990-17016) scored the highest with 1.82E-07, 1.29E-07 and 1.16E-07 respectively. In addition, the results of nucleotide diversity indicated that the isolates from V128 (16986-17012) scored the highest with 0.0026, then the isolates from V82 (16995-17021) and V222 (16990-17016) which scored 0.00016 and 0.00014 respectively. Finally, the isolates from V54 (16978-17004), V51 (16976-17002) in CC-174 and V114 (16991-17017) in CC-60 all scored 0.00001, while the other isolates pairs scored zero (Table 4.5; for nucleotide diversity see appendix 28). The assumption was drawn that these high nucleotide diversities and mean number of variants per coding sequence (CDs) were due to recombination and therefore, a further analysis of recombination within all paired isolates was undertaken in the next sections.

Table 4.5 : Listing different statistical parameters of the genome variation within the genic regions of 25 pairs isolates.

Volunteer	Isolate IDs	Size of total CDs analyzed	Total number of variants found	*Mean number of variants per coding sequence (CDs)	Nucleotide diversity (pi)
V43	16975_17001	1513620	0	0	0
V51	16976_17002	1546612	19	0.00349	0.00001
V52	16977_17003	1536548	6	0.00285	0
V54	16978_17004	1529000	19	0.00063	0.00001
V58	16979_17005	1505988	0	0	0
V59	16980_17006	1558173	4	0.00049	0
V88	16981_17007	1548859	5	0.00137	0
V138	16982_17008	1550735	4	0.00070	0
V64	16983_17009	1519820	1	0.00064	0
V117	16984_17010	1513305	7	0.00159	0
V124	16985_17011	1524820	4	0.00204	0
V128	16986_17012	1531401	554	0.02100	0.0026
V69	16987_17013	1509343	2	0.00128	0
V93	16988_17014	1539457	4	0.00199	0
V96	16989_17015	1496423	4	0.00032	0
V222	16990_17016	1302373	296	0.0400	0.00014
V114	16991_17017	1503078	11	0.00086	0
V134	16992_17018	1510446	4	0.00236	0
V185	16993_16994_17019_17020	1501067	4	0.00094	0
V82	16995_17021	1511372	397	0.02664	0.00016
V73	16996_17022	1490448	2	0.00064	0
V199	16998_17024	1512228	3	0.00900	0
V176	16999_17025	1546006	0	0	0
V188	16997_17023	1474991	0	0	0
V86	17000_17026	1468128	2	0.00116	0

*** means total number of variants in each time was divided by the length of variants conding sequences in each time point**

The variation in the isolates belonging to the various CCs was analyzed for two/three as compared to five/six month's carriage (Table 4.6). The former were for carriers where isolates, were collected at 0 and 2/3 months and the latter at 0 and 5/6 months. The total number of isolates between the two different periods was 10 and 15 for 2/3 months carriage and 5/6 months carriage respectively. Therefore, to compare the two different time points, normalization of the number of carriers was carried out through dividing the variable genes by the number of carriers at each time point.

Table 4.6: Number of pairs of isolates analysed for each time point and CC.

CCs	Number of carriages Short period	Number of carriages long period	Number of variable genes Short period	Number of variable genes long period	Number of variable genes in total
CC-174	4	4	25	12	37
CC-167	2	2	6	40	46
CC-23	1	4	4	11	15
CC-60	2	2	10	16	26
CC-1157-32-269	1	3	1	6	7

In general, non-parametric analysis (Mann-Whitney test) comparing two/three and five/six months carriage showed that there was no statistically significant difference in the number of variable genes between these time periods (P value 0.96) Similarly, there was also no statistically significant difference for genes with synonymous changes, genes encoding hypothetical proteins or genes encoding proteins with conserved functions (P value 0.98, 0.99 and 0.4) respectively.

However, for CC-174 carriers, the number of variable loci was higher after two/three month's carriage (25) than 5/6 months (12 loci). After two/three months carriage, there

were 14 out of 25 (56%) synonymous changes compared to 3 out of 12 (25%) after five/six months carriage. The Chi square test showed no significant difference in number of synonymous changes between short and long periods with P value 0.09. However, the high proportion of synonymous changes at the early time point indicates that the variation was random without any adaptive advantage in *N. meningitidis*. Conversely, the high proportion of non-synonymous changes in conserved functional proteins (75%) after five/six month's carriage suggests that more directional adaptive changes are occurring in isolates that have persisted for a significant time or insufficient time for purifying selection to remove the deleterious mutations.

For CC-167 and CC-60, the number of variable loci was higher after five/six months carriage than 2/3 months. Additionally, 33% (2 out of 6), and 50% (5 out of 10) of variable genes for CC-167 and CC-60 respectively at two/three months carriage coded for functional proteins with non-synonymous changes**.** Contrastingly, 37.5% (15 out of 40), and 81.25% (13 out of 16) of variable genes for CC-167 and CC-60 respectively at five/six months carriage coded for conserved functional proteins with non-synonymous changes suggesting also more directional adaptive changes in isolates that have persisted for a significant time or insufficient time for purifying selection to remove the deleterious mutations (Figure 4.3). Although, many variable genes encode conserved functional proteins. It is not clear whether this variation may have a role in persistence of *N. meningitidis*.

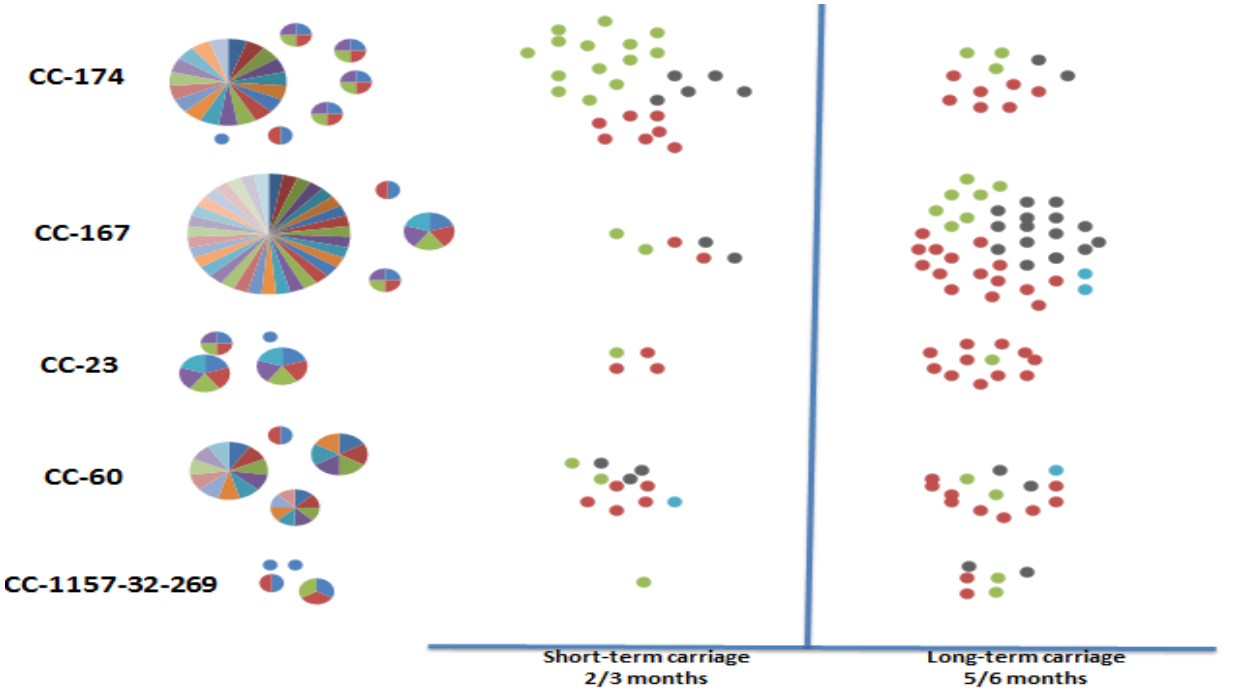


Figure 4.3: The variable loci within the different carriers of 25 pairs isolates and different time points along CCs. The panel on the left with pie charts shows the number of variable genes in each carrier by which each colour represents a particular variable gene. The panel on the right showed the variable genes in short-term carriage and long term carriage with the following; green color: synonymous variable loci, red color: non-synonymous functional variable loci, black color: hypothetical and sky color: Pseudogene.

4.5 Analysis of the type of genetic variation

In general, the analysis of variable loci among isolates belonging to different CCs revealed substitutions higher than indel changes. The overall variability among all the strains showed that there were 76 non-synonymous changes, 35 synonymous changes and 22 genes with other kinds of variability (11 internal stops, 2 non-triplet indel and 7 triplet indel).

The analysis of the proportion of variable loci with synonymous changes revealed high in CC-174 compared with other CCs. There are 37 variable loci in CC-174 of which 17 (45.9%) have synonymous changes.

As mentioned previously, the number of variable loci was high in CC-174 due to high variability in the paired isolates of V54 (16978-17004). This isolate was predicted to be subject to recombination with DNA from isolates of other CCs. On the other hand, analysis of variation also revealed that variable loci were higher in CC-167 due to high variability in the paired isolates of V128 (16986-17012) probably also due to recombination. Surprisingly, only 19.5% (9 out of 46 of variable genes) were under synonymous changes in CC-167 while 65% (30 out of 46) were subject to non-synonymous changes. However, a high fraction of genes with non-synonymous changes coded for hypothetical proteins 39.1% (18 out of 46) in CC-167 (Figure 4.3). Therefore, the conclusion was that significant enrichment in genes with synonymous change or genes with non-synonymous changes in hypothetical protein were observed due to the action of purifying selection in the donor genome to reduce the effect of deleterious mutation on survival of *N. meningitidis*.

For all variable genes, single nucleotide polymorphisms were detected in 54.1% (71/131) of variable genes and there were 1.8X (46/25) more non-synonymous changes than synonymous changes suggestive of the action of diversifying selection, insufficient time for purifying selection to remove deleterious mutation or the combination of both.

For all variable genes, single nucleotide polymorphisms were detected in 54.1% (71/131) of variable genes and there were 1.8X (46/25) more non-synonymous changes than synonymous changes suggestive of the action of diversifying selection, insufficient time for purifying selection to remove deleterious mutation or the combination of both.

4.6 Analysis of the functions of specific variants

The functions of the variable loci were determined from annotation data for each locus and then loci were assigned to functional groups by different schemes (KEGG).

Overall, 39 variable loci in all the CCs coded for hypothetical proteins (29.7%), 37 for metabolic pathway functions (28.2%) and 32 (24.4%) for environmental information processing and 18 (13.7%) for genetic information processing. These proportions do not however account for differences in the numbers of genes in each scheme.

The proportional effect of function of variable genes was therefore estimated (See chapter two 2.8.4). The proportions for hypothetical proteins and environmental information processing genes were the highest with (18.7 and 15.4) respectively (Table 4.7 for detailed section see appendix 29). The high variability in the environmental information processing may indicate that proteins belonging into those schemes are either OMP or encode enzymes that can modify the structure of surface determinants. Therefore, as these proteins and determinants in direct contact with the immune system, the variation may help in establishing resistance by changing outer membrane antigens. In the environmental information processing, variation was mainly found in pilin, porin and iron acquisition schemes which indicates a major role in the colonization by *N. meningitidis* and its persistence for long periods in the host. These results suggested that surface antigens are subject to diversifying selection mostly for pilin and porin during persistent carriage. These proportions do not however account for differences in the length of genes in each scheme.

Table 4.7 Functions of all variable loci all CCs. The function was grouped using KEGG scheme for all 131 variable genes found in the 25 paired isolates.

Different schemes	Number of non-synonymous variable genes	Number of synonymous variable genes	Number of genes in all genome	Percentage of proportion effect %
Genetic Information Processing	12	6	388	4.6
Metabolism	25	12	695	5.3
hypothetical protein	31	8	208	18.7
pseudogenes	5	1		
Environmental Information Processing (membrane protein)	24	8	207	15.4
Total	131		1498	

Percentage of proportion effect : number of gene/ number of genes in all genome in each scheme

The diversity score for each functional scheme was also detected. The nucleotide diversity was measured for each gene. Then the average of all the nucleotide diversity for the genes under each scheme was determined (See chapter two 2.7.3). The diversity score of environmental information processing was the highest with 0.027 but with no significant difference and a P value of 0.9, while scores for hypothetical proteins, metabolism, and genetic information processing were 0.011, 0.009 and 0.005 respectively. In addition, there were five genes belonging to environmental information processing schemes, one metabolism gene and one genetic information processing gene that carried an indel (Appendix 30). The overall result of the analysis of functional schemes and their diversity support the fact that the variation was highest due to SNPs and indel in environmental information processing functions.

There are different criteria to predict the effect of each variable gene on the adaptation of *N. meningitidis* to stressful conditions. As mentioned previously, these criteria include the fact that, genes undergoing synonymous mutation with 35 out of 131 loci are unlikely to contribute to persistence of strains. Conversely, genes important for the survival of *N. meningitidis*, or those encoding determinants located on the outer membrane are more likely to have a strong effect on the persistence of *N. meningitidis*. The overall result of the application of these different criteria to the variable genes showed that the percentage of loci in all CCs deemed less likely to contribute to persistence of *N. meningitidis* (synonymous, hypothetical and pseudogenes) was (70/131) 53.4% while the percentage of loci in all CCs (coded for OMPs) deemed more likely to contribute to persistence of *N. meningitidis* was (24/131) 18.3% and loci coded for other functions were (37/131) 28.2% (Figure 4.4).

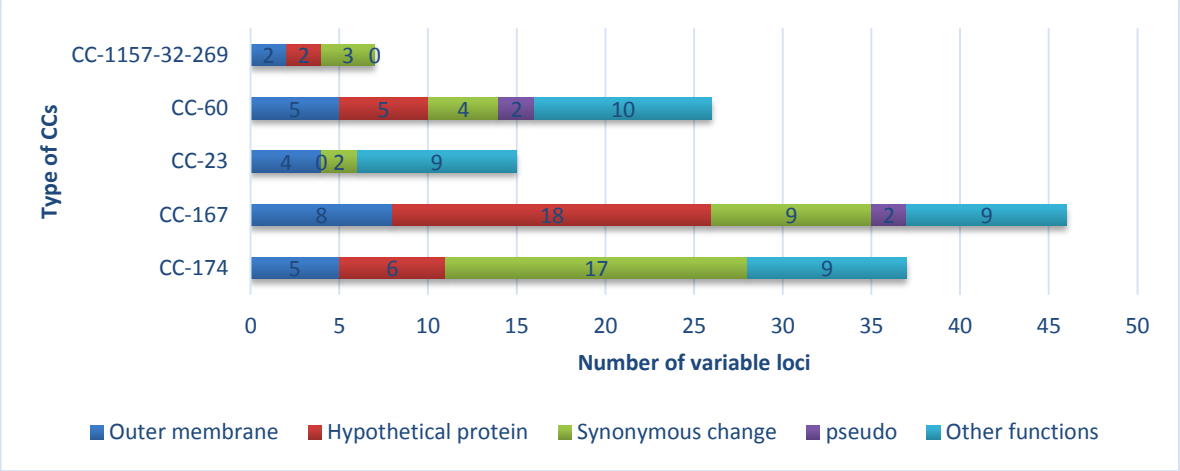


Figure 4.4: The variable genes in the isolates belonging to the CC-174, CC-167, CC-23, CC-60 CCs and the CC-1157-32-269complexes are depicted using a color scheme. Red color: hypothetical protein, green color: synonymous change, blue color: OMP, purple color: pseudogene, sky blue: Other functions.

The position and functionality of each non-synonymous mutation was checked for each variable gene by performing searches in a conserved domain database (NCBI's conserved domain databases). The mutations in varied amino acids located in conserved domains were considered further as potentially essential for the adaptation of *N. meningitidis* against stress condition. The results showed that, a high number of variable genes had mutations within a conserved domain with 24 in CC-167 (Appendix 31) while there were 9, 11, 14, and 2 in the isolates belonging to the CC-174, CC-23, CC-60 CCs and the CC-1157-32-269 (Appendix 32). Most variable genes showed variation within conserved domains of metabolism processes with 42%, membrane domains with 33% and genetic information processing with 22% (Figure 4.5). Genic variation was mainly found in porin and iron acquisition then pilin genes, in addition to mutS gene. This also confirms a major role of these genes in the colonization and persistence of *N. meningitidis* in their host.

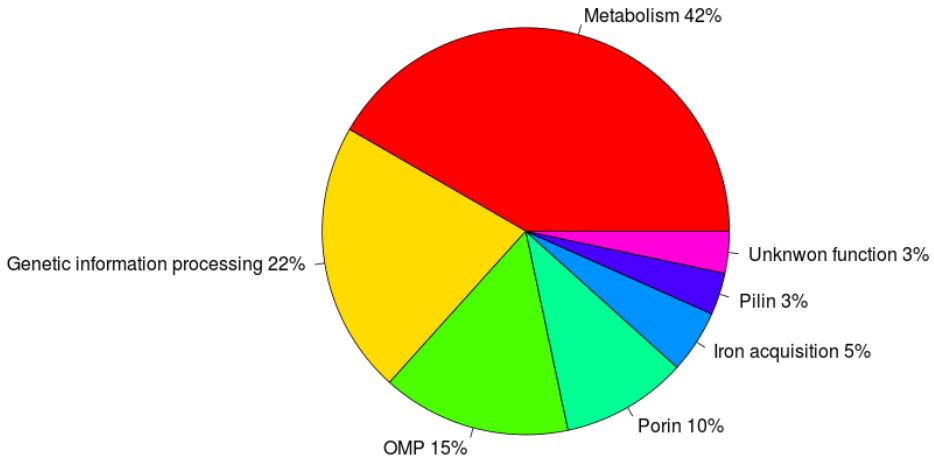


Figure 4.5: The percentage of variable genes with amino acid changes in conserved domains. The total number of variable genes were 96 (Excluding synonymous loci 35 out of 131) belonging into different CCs and these loci were used in CD BLAST against the database of NCBI's conserved domain databases.

Genic variation (mutable genes/genome) and mutation rate (genic variation /month of carriage) were determined for the 25 pair isolates (Appendix 33). In general, the average number of variable genes was 5.2 and the average genic mutation rate per month of carriage was estimated as $6*10^{-4}$ indicating that there was a low rate of genic variation per month for persistent meningococcal isolates. The low rate of genome variation may be due to the action of stabilizing selection. However, different evolutionary forces played a major role to enhance persistence of carriage. These are as follows;

Stabilizing selection may limit accumulation of mutations but, for many mutations, will be weak and will only operate to purge mutations over longer time-scales. In the mutant SNPs, non-synonymous mutations exceeded synonymous mutations indicating that this pressure is indeed weak. However, we also observed mutations exhibited trends toward surface antigens especially pilin, porins, iron acquisition and capsule genes suggesting that diversifying selection may be acting on these genes resulting in bias in the ratio and played a role in escaping the immune system of the host.

In the recombinant SNPs (Detected using sliding window approach see section 4.10), the number of synonymous polymorphisms was higher than the number of non-synonymous polymorphisms suggestive of the action of purifying selection acting on the recombination blocks in the donor genomes. This may suggest that the genes exhibited one or more non-synonymous polymorphisms, as a consequences of transformation by HGT and have been subjected to purifying selection in the donor genomes. These events are indicative of diversifying selection acting on this variation.

In addition, some variable loci occurred in multiple isolates and these loci may play an important role in the physiological activities of *N. meningitidis* as they have been more frequently selected than others. These loci coded mainly for an antigenic gene (PorB), a mismatch repair protein, D-alanine--D-alanine, ligase pyrroline-5-carboxylate reductase and a fatty acid efflux system protein

4.7 Analysis of variability of the IGRs of isolates belonging to different CCs

As mentioned in chapter three, two Perl scripts, extract_IGR.pl and Identify_var_IGR.pl (See previous appendix 3, 4), were used to extract IGRs from whole genome sequences of *N. meningitidis* and to compare the DNA sequences for each IGR from a pair of isolates. Then, the variable IGRs of each pair of isolates were aligned and printed in separate files.

Filtration processes were carried out to detect errors caused by both the detection method and inaccurate genome assembly. The filtration processes were mainly for the loci with variation in the different repeat tracts, the truncated loci and the loci with more than one copy in the genome. The number of spurious variable IGRs was similar for each CC (Table 4.8) and was mainly due to high variation in CEs and SSR. The number of filtered IGRs due to errors in the assembly were 7, 3, 3, 13, 10 for the CC-174, CC-167, CC-23, CC-60 and CC-1157-32-269 respectively.

Table 4.8: Number of variable IGRs filtered using different filtration processes.

Volunteers	More than one copy	CE	SSR	NIME	REP1,2,3,4,5	Truncated sequence
CC-174	0-1	0-17	1-7	0-1		0-1
CC-167	0-2	0-9	0-4			
CC-23	0-1	1-6	0-3			0-1
CC-60	0-1	0-21	1-4			0-8
CC-1157-32-269	0-1	0-23	1-3	0-1	0-1	0-1

The number of real variable IGRs after the filtration processes was 32, 28, 28, 36, and 18 for complexes CC-174, CC-167, CC-23, CC-60, and CC-1157-32-269 respectively. As mentioned previously, in the genic regions, the number of variable loci was higher in two isolates pairs (V54 and V128) compared with other 23. Conversely, in the IGRs, the number of variable loci in isolates pairs (V54 and V128) was not high as compared with other isolates pairs of different CCs (Figure 4.6). Recombination in isolates pairs (V54 and V128) mainly drives gene content variation rather than IGRs variation and this may happen by chance. However, in all CCs, a round half of the IGRs 71 (50%) varied due to one SNP, while variation in two, three and more than three SNPs occurred in 16 (11.2%), 6 (4.2%) and 29 (20.4%) regions respectively. In regard to other types of variation, there were 20 (14%) indels. In addition, the mean number of variants per IGRs was also detected and normalized by the length of the IGRs. The mean variants in IGRs after normalization and the nucleotide diversity scores indicate that recombination may also affect the IGRs resulting in the presence of multiple SNPs within some IGRs (Table 4.9).

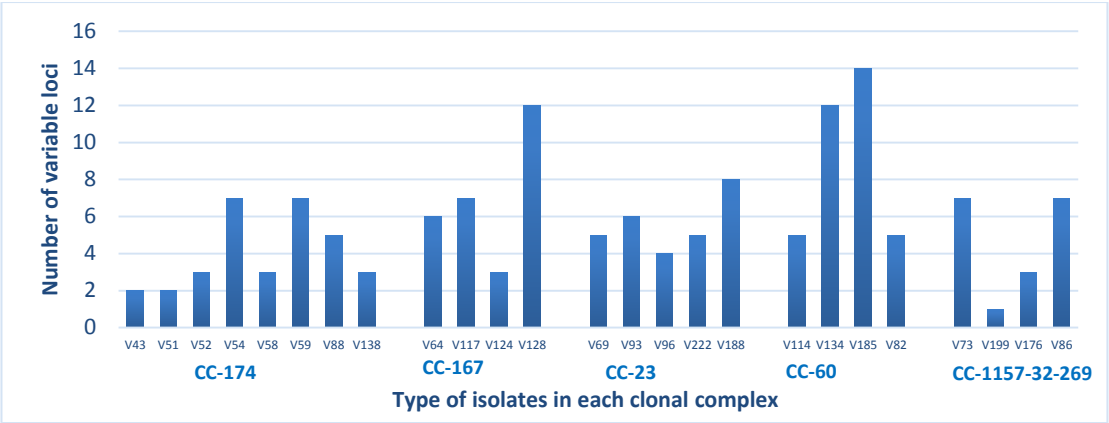


Figure 4.6: The number of variable IGRs for each pair isolates of each CC after filtration of spurious variation.

Table 4.9 : Statistical parameters for variation within the IGRs regions of 25 pairs isolates.

Volunteer	Isolate IDs	Size of total IGRs analyzed	Total number of variants found	*Mean number of variants per IGRs	Nucleotide diversity (pi)
V43	16975_17001	390878	2	0.00056	0
V51	16976_17002	406658	2	0.00159	0
V52	16977_17003	395502	4	0.00127	0
V54	16978_17004	401632	7	0.00083	0
V58	16979_17005	390272	87	0.01521	0.00003
V59	16980_17006	402272	64	0.00681	0.00002
V88	16981_17007	404625	122	0.01226	0.00005
V138	16982_17008	398182	35	0.00983	0.00001
V64	16983_17009	355148	7	0.00086	0
V117	16984_17010	358319	18	0.00282	0.00001
V124	16985_17011	357936	10	0.00231	0
V128	16986_17012	358571	52	0.00496	0.00002
V69	16987_17013	338683	4	0.00126	0
V93	16988_17014	311750	14	0.00198	0
V96	16989_17015	346850	7	0.00148	0
V222	16990_17016	279070	16	0.00612	0.00001
V114	16991_17017	330292	35	0.00265	0.00001
V134	16992_17018	315505	83	0.01118	0.00003
V185	16993_16994_17019_17020	333138	3	0.00057	0
V82	16995_17021	311750	35	0.00818	0.00001
V73	16996_17022	320264	7	0.00085	0
V199	16998_17024	328299	1	0.00416	0
V176	16999_17025	331233	128	0.0354	0.00005
V188	16997_17023	348857	36	0.00503	0.00001
V86	17000_17026	407607	172	0.0345	0.00006

*** means total number of variants in each time was divided by the length of variants conding sequences in each time point**

The total number of variable IGRs was 142 for the 25 pairs of isolates. The intergenic variation in the isolates recovered after a long period of carriage (five/six month's) was 91 for 15 pairs of isolates while for a short period of carriage (two/three month's), it was 51 for 10 isolates. After normalization for the number of isolates, the IGR variation was 61 for a long period and 51 for the short period.

Non-parametric analysis (Mann-Whitney test) for overall intergenic variation in two/three versus five/six months carriage showed that there is no statistically significant difference between variation with time point (P value 0.36). However, there is a possibility that IGRs that varied in more than one pair of isolates among 25 pair isolates may be more important than other loci and these loci affect different functions, (Appendix 34). As an example, the variation in the IGR of *NMB1497* which codes for TonB-dependent receptor may affect the level of expression of this receptor hence affecting the iron acquisition process.

4.8 Variation in the IGRs of 25 pair isolates

Since 92 of the variable IGRs were located upstream of genes, this variation may be in the core promoter region or in the region that functions as a pattern for a transcription factor binding site that can control gene expression. This variation was examined to see if it was located in core promoter region using a program for detection of promoters for procaryote analysis (BPROM"prediction of bacteria promoters"). The promoter prediction algorithm was designed by searching on a set of known promoters from the *E. coli* genome (Gordon *et al.*, 2003). The prediction showed that the variation in 10 of the variable IGRs was located in putative promoters, and half of the nearest genic region (5 out of 10) coded for an OMP. However, most (7 out of 10) IGRs with variation in the core promoter were located tail to tail between two adjacent genes. In addition, one of these IGR was located a long distance from the start codon and may be not a real promoter. This suggests that the variation in the IGRs may have a less signifincant effect than the variation in the genic regions on the persistence of *N. meningitidis* in their host (Table 4.10).

Table 4.10: Variable intergenic sequences predicted to be located within promoter regions. The score indicates the strength of the prediction and LDF value is given as an calculated as a log-likelihood score for each variable SNPs of IGR to be in the core promoter. The highest score indicates a promoter prediction that is more accurate. -10 and -35 are elements of promoters regulated by sigma 70 factors (Solovyev and Salamov, 2011). Tail to tail : means IGRs located tail to tail of two adjacent genes.

Gene	Number of varied copies	Function of nearby genes	Classification	Score -10/-35	Distance from start codon
CC-174					
NMB1540	1	lactoferrin-binding protein A	Outer membrane	68/41	Tail to tail
NMB0196	1	ribonuclease E	repair	26/37	Tail to tail
CC-167					
NMB1794	1	Citrate transporter	Outer membrane	49/12	less than 50 bp
NMB1926	1	lacto-N-neotetraose biosynthesis glycosyl	Outer membrane	33/35	Tail to tail
CC-23					
NMB0928	1	hypothetical protein	hypothetical protein	74/5	Tail to tail
NMB0215	1	hypothetical protein	hypothetical protein	68/41	less than 50 pb
CC-60					
NMB0195	2	4-hydroxythreonine-4-phosphate dehydrogenase	metabolism	60/27	Tail to tail
NMB0964	1	TonB-dependent receptor	Outer membrane	28/35	Tail to tail
NMB0702	1	competence protein	Outer membrane	33/41	around 350 pb
NMB1574	1	ketol-acid reductoisomerase	metabolism	32/31	Tail to tail

In addition, the variation in IGRs was examined for an association with small RNA (i.e. sRNA). An sRNA can control the expression of genes through their binding to mRNA, and therefore changes in the level of expression of an sRNA means a change in the level of mRNA for a particular gene. The expressed sRNA were identified using a customized tiling oligonucleotide microarray on MC58 isolate. It has been observed that 91 loci were identified as intergenic sRNA and the target of sRNA was predicted using TargetRNA program (Del Tordello *et al.*, 2012). The IGRs containing sRNA in our isolates were detected using a BLAST search against the data that has been taken from this paper. There were 6 IGRs predicted to contain sRNA and where the variation was not located within the core promoter. Most of these variable sRNA have an effect on genes that code for hypothetical proteins (Table 4.11).

Table 4.11: Variable intergenic sequences predicted to be located within sRNA regions for the 25 pair isolates.

IGRs	Number of varied copies	Function	Classification
CC-174			
NMB1815	3	hypothetical protein	hypothetical protein
NMB0660	1	hypothetical protein	hypothetical protein
CC-167			
NMB1815	3	hypothetical protein	hypothetical protein
NMB1881	2	hypothetical protein	hypothetical protein
NMB1107	1	hypothetical protein	hypothetical protein
CC-60			
NMB1452	1	hypothetical protein	hypothetical protein
NMB1815	1	hypothetical protein	hypothetical protein
CC-1157-32-269			
NMB1497	1	TonB-dependent receptor	membrane

In summary, these findings indicate that the observed variation with the time of carriage is mainly due to random point mutation and may introduce no additional benefits for bacterial cells or the population. However, the variation of IGRs have to be checked practically.

4.9 The functions of genes that carried variable IGRs in different CCs

The analysis started with exclusion of IGRs between two adjacent genes in a tail to tail orientation and this left 92 IGRs where the variation is located either upstream of the nearest gene or between two adjacent genes in a head to head organization. In head to head regions, the function of the nearest gene to the SNPs was considered as the affected gene. The functions of genes that have varied IGRs were identified (Table 4.12). Most of these code for hypothetical proteins (n = 29, 31.5%), membrane proteins (n=27, 29.3%). metabolic pathway enzymes (n = 23, 25%) and genetic information processing (n = 11, 11.9%). In addition there were two pseudogenes. The proportional effects for each scheme showed that the genes connected with hypothetical proteins and membrane functions were the highest with 13.9% and 13%, then genetic information processing and metabolism with 3.5% and 2.8%. Therefore, the genes connected with membrane functions (Especially in iron acquisition with 7 IGRs) could participate in adaptation as a mechanism for *N. meningitidis* to change the surface proteins and possibly to resist the immune system of the host.

Table 4.12: Functions of all variable loci in all CCs for the genes that carried varied IGRs. The function was grouped using KEGG for 92 variables IGRs.

Volunteers	Metabolism	Outer membrane	Hypothetical protein	Information processing	Pseudo
CC-174	2	16	5	2	
CC-167	2	5	8	1	
CC-23	5	1	8	4	
CC-60	9	1	6	2	2
CC-1157-32-269	5	4	2	2	

4.10 Identification of recombination in genic regions and IGRs of persistent isolates using a sliding window approach

The analysis of variation revealed 131 variable genes and 142 variable IGRs among all the CCs for the 25 paired isolates. To determine if point mutations or HR caused this variation, a method for estimating HR was utilized based on the work of Kong *et al.* (2013).

This is an example and this was done for one isolate of every pair. The contigs of ID: 17012 isolate were aligned with MC58 as a reference genome using PROmer alignment. The MapView showed a round 5 Kbp was not matched between reference genome and ID: 17012 isolate. Percent identity of alignment in most contigs was more than 50% (Figure 4.7). The output of PROmer alignment was used as input for ABACAS script.

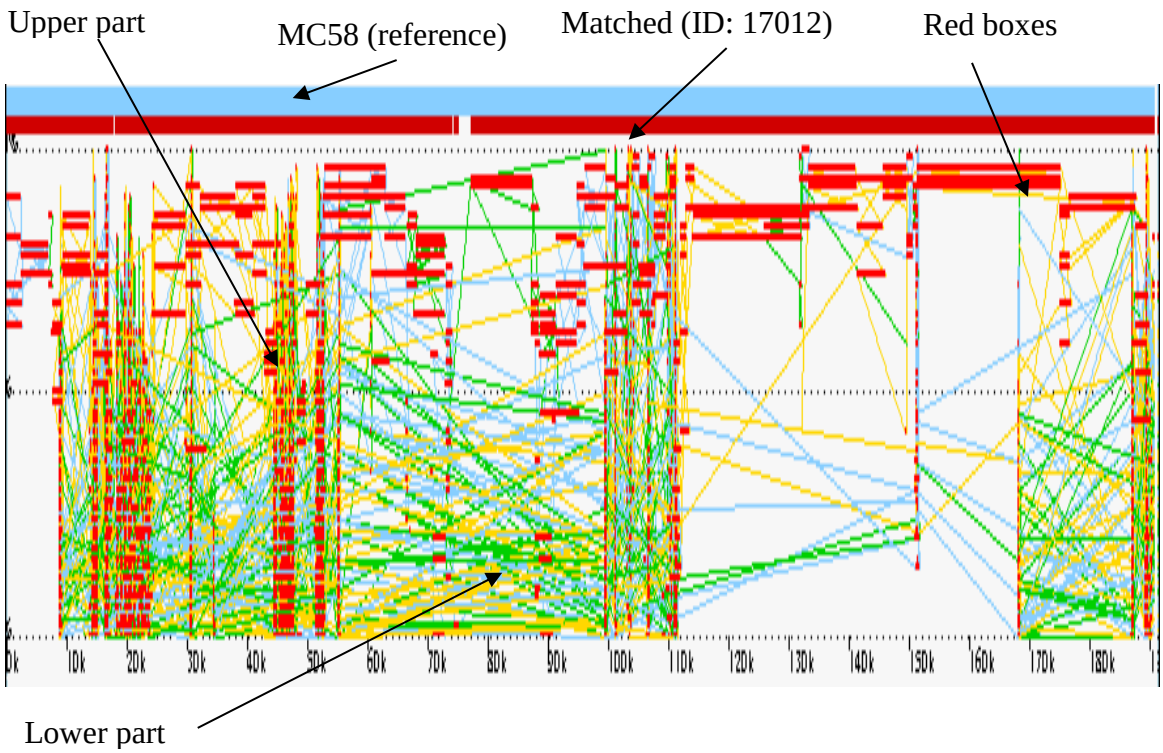


Figure 4.7: PROmer alignment for ID: 17012 isolate with MC58 as a reference genome. This is the first step to order the contigs depending on a reference genome. The blue rectangle spanning the figure refers to MC58 (reference). PROmer matches are showed two times. Once just below the reference with a red rectangle indicates if matched sequences are found or not found. The other red boxes refer to percent identity. The boxes in the upper part have more than 50% identity while those in the lower part have less than 50% identity.

An ABACAS script was then used to rapidly determine contiguity (align, order, orientate) and to estimate gaps between contigs based on the reference genome. The ordered contigs of ID: 17012 isolate can be visualized using ACT program. Variation in genic regions and IGRs were mapped onto the ordered contigs of one isolates of each pair using ACT second bar in Artemis visualization (Figure 4.8). In this case, the positions of all SNPs of varied genic and IGRs among the different contigs of one isolates of each pair were detected. The final result for each contig belonging into each pair of isolates was two files. The first file was the DNA sequence of contig with original sequence (before variation) and the second file was the same DNA sequence of contig with mapped varied SNPs of genic regions and

IGRs (after variation). The index_var.pl script was written to detect the index of variation between two compared DNA sequence and print the position of variation within the compared sequences. The previous two files (original sequence) and (varied sequence) of each contig belonging to each pair of the 25 pair isolates were the input for the (index_var.pl) script (See chapter two 2.6.2.1).

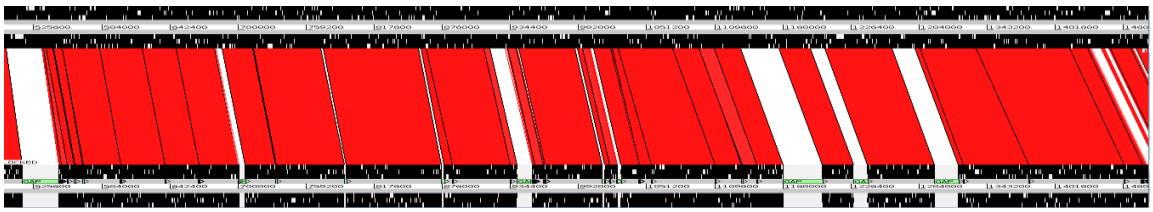


Figure 4.8: ACT program used to visualize the output of ABACAS script for the alignment between ID: 17012 isolate with MC58 as a reference genome. The red block means the contigs were aligned between isolates and reference. Green par is the gap in ID: 17012 isolate which means the pieces of sequence found in the reference but are not present in ID: 17012. The gaps were filled with Ns.

Recombination was detected using a sliding window approach and a specific set of statistics. In brief, the identification of non-vertical homologous genes depends on the upper bound of genome wide divergence (μ) being estimated from multi locus sequence typing (MLST) loci with no observed substitution when nucleotide substitution follows a binomial distribution. Then, this μ value was used to calculate the P-value for the observed nucleotide changes for each of the variable genes in the genomes of a specific sequence type. At a significance level of 0.001, the genes that have more than the expected number of nucleotide changes were considered to have been subject to HR. Depending on this idea, the script (sliding_win.pl) was written to detect recombination patterns for overlapping sections of a DNA sequence (for each contig) using a sliding window approach (See chapter two 2.6.2.1). The output of index_var.pl that is the index of variation for compared contigs was used as input for the sliding_win.pl script.

The total number of SNPs among the 25 pair isolates in genic regions and IGRs was 2303. The program predicts the presence of 2020 recombination SNPs and 283 point mutation SNPs. In addition, 499 indels were detected (Table 4.13). There were 796 recombinant

SNPs out of 2020 located in 17 IGRs of which 6 loci were predicted to be within promoter regions while 2 loci were predicted within sRNA. There were 1224 out of 1352 (Total SNPs in the CDs) (90.5%) recombination SNPs located within CDs. There were 283 SNPs introduced by point mutation, 155 point mutation SNPs were in the IGRs and 128 point mutation SNPs located within CDs.

Table 4.13: Distribution of recombination and point mutation SNPs with indel for different CCs for 25 persistent isolates.

Volunteers	Mutation SNPs/ intergenic	Mutation SNPs/ genic	Recombination SNPs/ intergenic	Recombination SNPs/genic	Indel
CC-174	26	42	297	15	71
CC-167	46	44	41	522	99
CC-23	26	11	51	295	100
CC-60	41	24	115	392	82
ST-1157	16	7	292	0	147
Total	155	128	796	1224	499

To further elucidate the role of recombination, the number and size of the recombination blocks was investigated using the sliding window approach. There were 2020 SNPs introduced by 32 recombination blocks. The size of the identified recombination blocks varied between 502 bp and 2819 bp with a mean length of 1660 bp. The recombination blocks were present in 14 of the 25 pairs of isolates. A signal of recombination was also reported when 3 adjacent loci were present in the same block. There were five

recombination blocks of this type in the isolate from V128 (16986-17012) and one recombination block in V82 (16995-17021). The recombination patches were visualised by plotting the number of SNPs for each window against genome positions (Figure 4.9). The total length of the recombination fragments was 63188 bps indicating that approximately 3.29% of the total genome was subject to recombination.

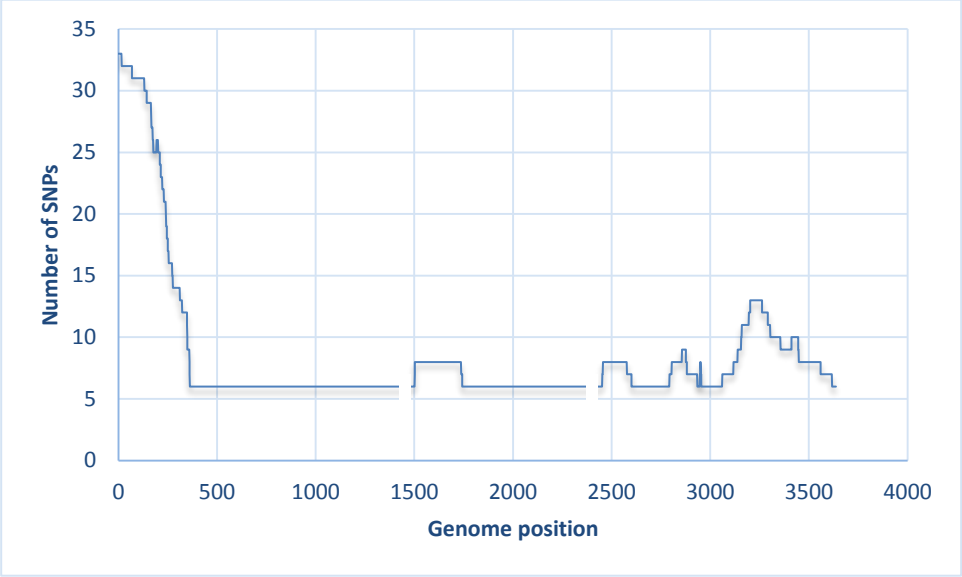


Figure 4.9: Plotting the number of SNPs against genome position for the recombination fragments detected using sliding window approach. An example of signals of recombination that were found with 3 or more adjacent loci on the same block (*NMB0437*, *NMB0438*, *NMB0439*, *NMB0440* and *NMB0441)* in the V128 (16978-17012) isolates.

The number of synonymous and non-synonymous in the genic recombination SNPs was detected using SNAP program. They were 728, and 496 respectively. The Sd/Sn (number of synonymous/number of non-synonymous) in recombination SNPs was (1.4). This indicates that the synonymous was higher than non-synonymous variation and some recombination blocks may be under purifying selection in the donor genomes.

4.11 Functional categories of all the mutation and recombination patterns within the 25 paired isolates

The results showed that recombination has a strong effect on different functional categories of proteins (Figure 4.10). The proportional effect of hypothetical proteins and environmental information processing was the highest with 8% and 6.7%. Furthermore, functional classification of the genes affected by point mutation also showed that effect on environmental information processing was the highest with 7.2%. Then genetic information processing and metabolism with 5.1% and 2.6% respectively (Figure 4.11).

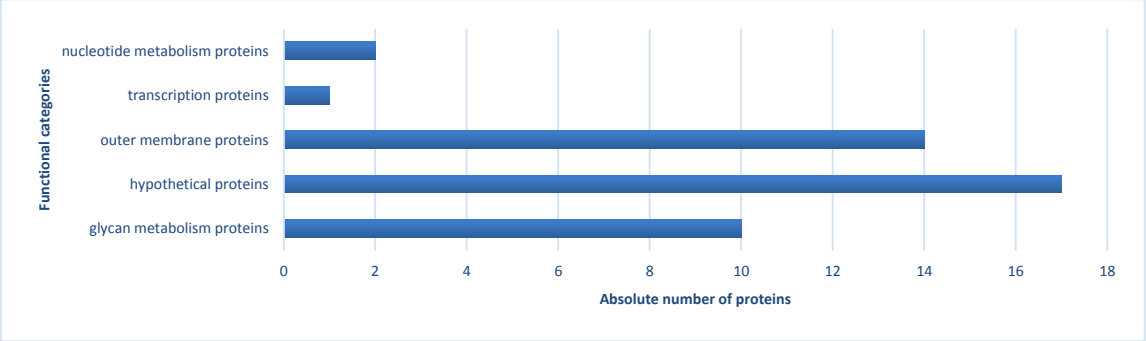


Figure 4.10: Functional categories of proteins affected by recombination

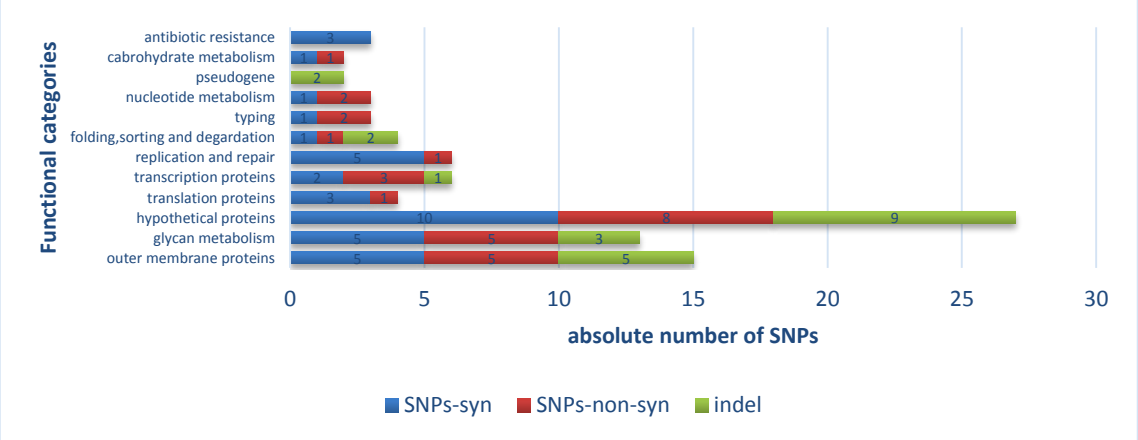


Figure 4.11: Functional categories of proteins affected by point mutation. Blue bars synonymous SNPs, red bars non-synonymous SNPs and green bar is indel.

4.12 Characterization recombination patches containing three or more adjacent loci

Five recombination blocks have recombination signals with three or more adjacent loci within isolates of carrier V128 (16978-17012). The first recombination block contains genes that code mainly for pilus proteins and glycan biosynthesis (metabolism proteins). The second, third and fourth recombination blocks mainly code for hypothetical protein, glycan biosynthesis and an RNA methylase. The last block encodes mainly for genetic information processing (Figure 4.12). There was also one recombination block in V82 (16995-17021). This recombination block contains genes coded for para-aminobenzoate synthatase and a hypothetical protein.

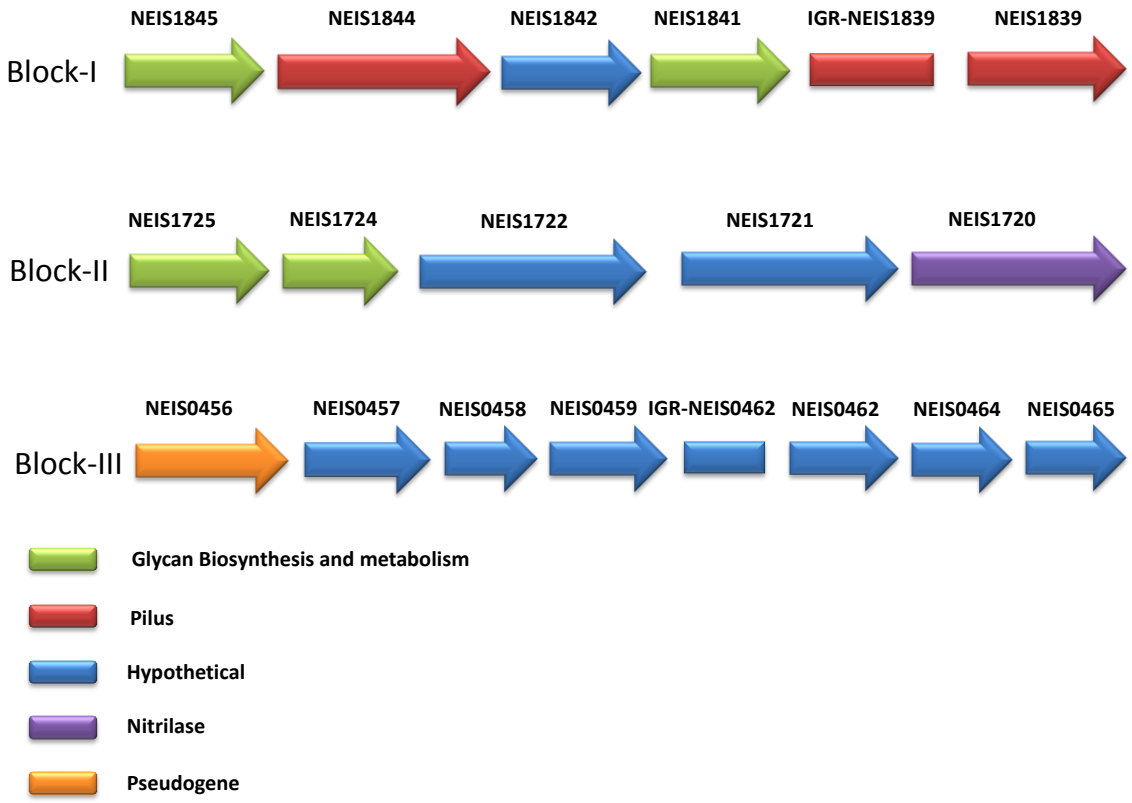


Figure 4.16: Schematic representation of recombinant genes with different blocks detected using sliding window approach. The bigest three recombination blocks from five in the V128 (16978-17012) containing three or more adjacent loci for each block were shown.

The varied genes present in the recombination blocks were likely to contribute to persistence of *N. meningitidis* in all recombination blocks, were as follows. The *NEIS1841* codes for a Dephospho-CoA kinase which is one of the recombinant proteins that are found in many isolates (recombinant *N. meningitidis* serogroup A), this protein has a role in the formation of coenzyme A and variation in this protein can effect acylation and acyl-transfer reactions (ID: Q4W580 UniProt). Interestingly, variation was often located within two pilus modification genes *NEIS1844* and *NEIS1839* (*pilF* and *pilD*) and in the IGR upstream of *pilD*. Pilus genes play a significant role in adhesion to host cells and in transfer of nucleic acids. It has been reported that variation in meningococcal pilus gene is used as a mechanism to escape the immune system of the host (Gault *et al.*, 2015).

The variation was also in both the coding and intergenic sequences of the *NMB1988* gene code for iron-regulated outer membrane protein (FrpB). This gene is involved in iron acquisition by *N. meningitidis* and may overcome the nutritional immunity exerted by the innate immune system on iron availability.

4.13 Summary of main findings

An analysis was conducted of the variation in the genes and IGRs of 25 pairs of persistent meningococcal carriage isolates collected from two different time periods (0 and 2/3 months and 0 and 5/6 months) from different volunteers and different CCs. In the genic regions, the number of variable loci was highest in CC-174 and CC-167 due to there being a high number of variable loci within two pairs of isolate V54 (16978-17004) and V128 (16986-17012).

The comparative analysis of variation between the time periods revealed that there are no statistically significant results with regard to the overall number of variable loci, the number of variable loci coding for hypothetical proteins, synonymous change, and the loci that code for non-surface related function and surface related function. However, the variation was seen in genes related with specific functions in *N. meningitidis* such as pilin, porin, and iron acquisition. These three classes of genes varied in more than one pair of isolates among the 25 paired isolates and in multiple CCs. Therefore, the variation in the genic regions seems to be important in helping *N. meningitidis* to adapt to the stress condition of the host environment.

Some variation in IGRs was predicted to be in the core promoter. However, most of the IGRs variation was located between two adjacent loci with a tail to tail organisation. In addition, variation in the IGRs showed no statistically significant increase with the time period or with any specific classes of functional genes or elements. Therefore, variation in the IGRs seems to be less likely to contribute in the persistence isolates from multiple carriers. However, most patches of variation in the IGRs were seen in the CEs pattern (See chapter six).

The proportional effect of function of genic regions showed that environmental information processing was the highest with (15.4%) therefore most variable genes either encoded OMPs or for enzymes which facilitate modification of outer membrane structures. Moreover, the estimation of nucleotide diversity was significantly high in the environmental information processing genes with 0.027 comparable with 0.009 and 0.005 in the metabolism and genetic information processing respectively. In addition, the proportion effect of function of genic regions with varied IGRs also showed that

environmental information processing was the highest with 13% followed by metabolism and genetic information processing with 3.5% and 2.8%.

In the environmental information processing scheme, the variable genes were associated with membrane transport that include ABC transporter, bacterial secretion system and phosphotransferase system and signal transduction. Bacterial secretion system includes pili and flagella, nutrient acquisition, and efflux of drugs and other toxins. However, variation was mostly found within genes coding for pilin, porin and iron acquisition proteins. This finding suggests that the variation of genic regions in the 25 pair isolates may enhance the persistence of *N. meningitidis* for long time in their host due to the fact that the pilin and iron acquisition systems play an important role in host colonisation and in preventing immune system from clearing the bacteria (Stephens, 2009).

While the analysis of 40 isolates (See chapter three) showed variation in the capsule region A of serogroup Y isolates, in the 25 paired isolates the variation was also seen in the capsule region A of serogroup Y isolates. Moreover, the Locus variable in multiple isolates in the environmental information processing scheme was *NMB2039*, this gene encodes antigenic gene (PorB) and it has been reported that modification of the antigenic structure of *N. meningitidis* by genetic variation may help in evading the immune response especially for LPS, pili and capsules (Uberos *et al.,* 2015).

In general, variation was seen in genes coding for OMPs or modifiers of outer membrane determinants such as the capsule, indicating a role in the persistence of *N. meningitidis* possibly due to these factors having a direct interaction with the immune response of the host so that any change in these factors could serve to establish immune resistance. As example of variation in the OMP, variation in genes coded for capsule polysaccharide. These proteins may be recognized as a receptor by immune system of host therefore the **c**hange in these proteins may help escaping the immune system attack strategies.

The results of the current study are compatible with other studies as follows, a study of variation in nasopharyngeal isolates was carried out with time points over a period of 7 months after colonization of a laboratory worker with disease isolates. The study showed

there was a variation in proteins belonging to the outer membrane (Woods and Cannon, 1990).

Other genome comparative longitudinal studies achieved on different serogroups of disease or commensal isolates showed that there are waves of clonal replacement with the same serogroup due to immunity to non-capsular antigens. A study carried out on meningococcal meningitis in the African "meningitis belt" showed that there was a replacement of ST-7 by the ST-2859; these strains differed due to HR with more than 20% of the recombinant loci acquired from other species. Most changes were in *pgl*, regulation of pilus expression and the synthesis of Maf3 adhesions that reflect the importance of changes in surface proteins for escape responses of adaptive immune system (Lamelas *et al.*, 2014).

A Locus variable in multiple isolates was *NMB2060*, this gene encodes for a mismatch repair protein and was seen in V144 (16991-17017), V59 (16980-17006) and V96 (16989-17015). This finding was compatible with analysis of variation in 40 isolates (See chapter three) where this gene varied in multiple isolates. As mentioned in the chapter three, the variation in this gene may enhance variation in other genes due to a defect in the repair system. A study showed that mutability due to defects in mismatch repair gene lead to resistence to antibiotics (Davidsen *et al.*, 2007).

The analysis of the types of variation showed that most variation was due to one SNP in the CC-174 isolates. This included the isolates from V54 (16978 and 17004), which had a high number of variable loci. The strain in this carrier is likely to have recombined with a multiple closely related strain.

On the other hand, most of the variation in the CC-167 isolates was due to a high number of SNPs within the same loci (Accompanied with high number of variable loci) for the isolates from V128 (16986 and 17012) (This carrier was colonised by another strain but we have not sequenced it yet). In this case, recombination is likely to have occurred with isolates from an evolutionary distant strain resulting in a high amount of variation. Analysis of recombination depending on the sliding window approach (density of SNPs) showed that most of varied genes (26/35 loci 74.2%) in this paired isolates were under

recombination. A more in depth analysis of recombination found that, for genic and intergenic variation, there were 2020 SNPs introduced by 32 recombination blocks with 1224 recombination SNPs located within CDs. There were 283 SNPs introduced by point mutation while 128 point mutation SNPs were located within CDs. In total 1352 SNPs were detected in the 131 variable genes among 25 paired isolates. 1224 out of 1352 (90.5%) were associated with recombination rather than point mutation. Thus, recombination has a major role in driving sequence diversity during evolution of *N. meningitidis* with their host.

The Sd/Sn in recombination SNPs was 1.4. This may indicate that some recombination blocks acquired from other strains where these blocks were under purifying selection for a long time to diminish the effect of deleterious mutations and that 3.2% of the total length of genome was subject to recombination. However, the non-synonymous change was 1.8 higher than synonymous change in mutational SNPs of different functional genes (genes with multiple functions carrying one nucleotide change) suggestive of diversifying selection especially for OMPs, insufficient time for purifying selection to remove the effect of deleterious mutations or the combination of both.

These results are compatible with a study, carried out to visualize the impact of recombination on dN/dS. This study showed a highly significant increase in synonymous changes in the genes under recombination patterns for *Staphylococcus aureus*, and *Clostridium difficile* (Castillo-Ramirez *et al.*, 2011).

The proportion of varied genic regions subject to recombination was the highest for the environmental information processing and metabolism categories. While for varied genic regions under mutation the highest were environmental information processing and DNA information processing. In addition, the variation was also seen in the methylation genes but the percentage was very low. This may indicate that proteins belonging to those schemes are in direct contact with immune effectors and the variation may help in escaping the immune system. In addition the variation in the methylation genes may reduce barriers to DNA exchange. In comparison with our study, Kong *et al.* (2013) suggests that recombination has a major effect on antigenic OMPs that helps in rapidly altering antigenicity of *N. meningitidis* as a mechanism for adaptation of *N. meningitidis*

against stress conditions and genes that coded for restriction and modification systems were also under the effect of recombination, which may reduce barriers to DNA exchange. Therefore, we can also assume recombination is a powerful mechanism for driving evolution by which adaptation of *N. meningitidis* against stress condition within the host can be inferred.

Finally, a recent study investigating the variation within hosts during short – term asymptomatic carriage (a period of two months only) showed the average number of allelic differences between paired isolates was 35 (Bårnes *et al.,* 2017). This variation was significantly high comparable with our study that was carried out on paired isolates for a longer period (1-6 months). In our study, average genic variation per month was low and some carriers even after six months did not show any mutation events, this indicates the role of stabilizing selection in limiting the accumulation of mutations. However, a trend toward mutations in environmental information processing genes was observed which may indicate the action of diversifying selection for the purpose of evolution of meningococcal isolates within the host as mechanism for immune evasion.

Chapter 5: Inferring rate of recombination in 25 CC-174

5.1 Introduction

Several studies have inferred a role for recombination as a driving force for variation and in shaping the evolution of different isolates of *N. meningitidis.* The penicillin-binding protein 2 genes (*penA*) were shown to have been transferred into *N. meningitidis* genomes from other species of Neisseria, especially *Neisseria flavescens,* through recombination (Bowler *et al.*, 1994). This suggests that interspecies recombination may help to establish resistance to antibiotics in *N. meningitidis*.

A study carried out on 107 isolates of *N. meningitidis* showed that recombination events strongly support phylogenetic incongruence among the 30 most diverse isolates (Holmes *et al.*, 1999). The variation in branch lengths of some of these isolates could be explained by recombination events that occurred between *N. meningitidis* isolates and isolates belonging to other species. Thus the importation of divergent genes, or large fragments, led to longer branching, while the accumulation of de novo mutations led to short branching in the phylogenetic constructions. This study also showed that a prolonged timescale of recombination led to formation of high reticulation within a network tree, with the lack of extensive treelike structures strongly suggesting frequent recombination events (Holmes *et al.*, 1999).

A series of studies have detected recombination in specific genes of *N. meningitidis* and found a key role for repetitive sequences. Harrison *et al.* (2013) showed that variation in the HpuAB proteins in some *N. meningitidis* isolates was due to recombination and inferred a role for selection driven by responses of the host immune system. Achaz *et al.* (2003) showed that recombination correlated with the presence of inverted repeats, such that in genomes with high numbers of inverted repeats recombination was high. Aho *et al.* (1997) showed that the *pilS* loci in class II pilin might have arisen through recombination events with class I pilin in other Neisseria species. The *porA* gene was transferred between isolates belonging to an ST-269 clone and other isolates belonging to different CCs. It has been reported that transformation between different isolates in the presence of polysaccharide vaccines may lead to the capsule switching changing the C serogroup into B serogroup for an ST-269 clone in Canada between “2003-2010” (Zhou *et al.*, 2012).

Two studies highlight the key methods used to infer and understand recombination events. Firstly, the *pilE*/*pilS* genes have been shown to be chimeric sequences consisting of highly divergent sequences flanked by conserved regions. This structure is the result of HR. The method used to infer recombination patterns within *pilE*/*pilS* was identification of the recombination breakpoints. This method relies on finding a specific pattern of clustering of variant sites between a pair of sequences, which is statistically significant. The results showed that there were a high number of breakpoints between conserved, highly variable and semi variable regions with the evolutionary pattern being driven by recombination and positive selection. The semi variable region was subject to recombination with small blocks between conserved sequence elements while hyper variable regions were subject to recombination accompanied with selection between silent and expressed loci (Anisimova *et al.*, 2003; Smith, 1992; Andrews and Gojobori, 2004).

Secondly, an isolate M16917 of *N. meningitidis*, was detected as belonging to ST-11 CCs, but was associated with serogroup C using multilocus sequence typing and whole genome sequence comparison. The capsule polymerase gene (*synD*) of M16917, however, was detected as belonging to serogroup B, using sequence similarity comparisons and phylogenetic analysis; this suggests that HR of the capsule of M16917 led to a capsule-switching event (Rishishwar *et al.*, 2012).

Another example is the variation in the FadL-like membrane protein. This protein was divided into four variants depending on the level of similarity between different species. This protein was less divergent than other major OMPs. The mosaic structure of the FadL-like membrane protein was mainly due to recombination, which was detected by ClonalFrame version 1.1 and relying on maximim-likelihood analyses. The level of recombination events was intermediate, with a relative effect of recombination and mutation (r/m) value of 1.8 and a rate of recombination (R) of 3.2 for 68 Neisseria strains. Phylogenetic analysis of the two major variable loops containing VR1 and VR2 showed changes in the topology of trees, indicating the mosaic structure of proteins inferred by recombination (Yero *et al.*, 2010).

It has been reported that recombination as a driving force for evolution forms different population structures within the same organisms (Bart *et al.*, 2001). An example of how

recombination affects the population structures of lineages of *N. meningitidis* illustrated as follows; Budroni *et al*. (2011) proposed the phylogenetic clade (PC) concept and it referred to CCs that are present in one cluster of a phylogenetic tree. This organization can be affacted by recombination. Budroni *et al.* (2011) showed that gene conversion events within the same PC were more than fivefold longer than between different PCs, and was correlated with the number of restriction and modification systems (RMSs) shared between isolates. This study also showed that strains belonging to the same PC only differ in one or very few RMSs while between PCs there were multiple different RMSs. The author concluded that these different RMSs acted as barriers for DNA transformation (Budroni *et al.*, 2011).

The previous studies focussed on macro evolution between CCs and species and that the aim of this study was to focus on microevolution. Therefore, the aim of this chapter was to investigate how recombination within lineages of *N. meningitidis* can result in microevolution. To achieve this aim, a set of 25 highly related isolates belonging to one CC, CC-174 were chosen. These 25 isolates consisted of 18 carriage and 7 disease isolates, mostly collected between 2008 and 2011, except isolate 27497, which was collected in 2000 (Bidmos *et al.*, 2011; Oldfield *et al.*, 2016) (Table 5.1).

Table 5.1: Listing the name, ID, Strain designation and infectivity of isolation of 25 CC-174 isolates of *N. meningitidis*.

Isolate	Isolate id	Strain designation	Infectivity
M10 240694	20072	Y: P1.21,16: F3-7: ST-1466	Invasive
M10 240759	20115	Y: P1.21,16: F3-7: ST-1466	Invasive
M11 240073	20232	Y: P1.22,9: F3-7: ST-1466	Invasive
M11 240161	20282	Y: P1.21,16: F3-7: ST-1466	Invasive
M11 240165	20285	Y: P1.21,16: F3-7: ST-1466	Invasive
M11 240209	20308	Y: P1.21,16: F3-7: ST-1466	Invasive
M11 240211	20310	Y: P1.21,16: F3-7: ST-1466	Invasive
20132	27506	Y: P1.5,2: F3-7: ST-1466	Carriage
21292	27517	Y: P1.21,16: F3-7: ST-1466	Carriage
21789	27523	Y: P1.22,9: F3-7: ST-1466	Carriage
20951	27526	Y: P1.21,16: F3-7: ST-1466	Carriage
21092	27530	Y: P1.21,16: F3-7: ST-1466	Carriage
21888	27538	Y: P1.21,16: F3-7: ST-1466	Carriage
22007	27539	Y: P1.ND,16: F3-7: ST-1466	Carriage
22008	27540	Y: P1.21,16: F3-7: ST-1466	Carriage
22014	27541	Y: P1.21,16: F3-7: ST-1466	Carriage
NO0011039	27497	Y: P1.21,10-46: F3-7: ST-9893	Carriage
22933	27556	Y: P1.21,16: F3-7: ST-1466	Carriage
21570	27572	Y: P1.21,16: F3-7: ST-7850	Carriage
23214	27578	Y: P1.21,16: F3-7: ST-1466	Carriage
23283	27580	Y: P1.21,16: F3-7: ST-1466	Carriage
23326	27581	Y: P1.21,16: F3-7: ST-1466	Carriage
N54.1	28250	Y: P1.21,16: F3-7: ST-8510	Carriage
N59.1	28252	Y: P1.21,16: F3-7: ST-1466	Carriage
N88.1	28253	Y: P1.21,16: F3-7: ST-1466	Carriage

5.2 Analysis of variation in the entire assembled genome of isolates belonging to the CC-174

The variation in the genic regions was detected using GC and AC software options in BIGSdb, and then a series of Perl scripts were used to distinguish between real and spurious variation as described in chapter two (Figure 2.5 and 2.6). In the first instance, GC software detected 698 variable loci, 144 loci missing in all isolates, 128 identical loci in all isolates, 973 identical loci in all isolates except the reference, 120 incomplete loci and 64 potentially paralogous loci among the 25 isolates of CC-174.

A series of Perl scripts (Figure 2.5 in 2.5.4.1 chapter two) were used in pipelines to show that 558 out of the 698 variable loci exhibited real variation. The first filtration process detected around 54 variable genes with more than one copy. The second filtration process detected 12 phase variable genes (6 out of 12 phase variable genes were detected within the 46-omitted core genes (Appendix 9). Variable loci missing in any one of the 25 isolates (40 loci) were also removed from the analysis. Mismatches in the alignment were detected in 22 loci and genes with indel only were found in 12.

The first run of AC software detected the presence of 308 varied loci among the 25 isolates of CC-174. Then, Perl scripts (Figure 2.6 in 2.5.4.1 chapter two) were also used in pipelines to show the presence of 6 out of 308 loci with real variation. The first run of filtration was to exclude the variable genes that were already detected using GC software, 298 variable genes were filtered. The variable genes with more than one copy were 4, therefore the overall real variable genes detected using both GC and AC software was 564 loci (Table 5.2).

Table 5.2: List the real variable genes detected using AC and GC software and Perl scripts.

Name of software and filtration processes	Putative variable loci	Real variable loci after each filtration processes
Genome comparator (genic reigns)	698	First run
Genes with more than one copy	54	644
Phase variable genes	12	632
Missing loci	40	592
Genes with indels only	12	580
Mismatch in alignment	22	558
Allele comparator (genic regions)	308	First run
Genes already detected by GC	298	10
Genes with more than one copy	4	6

Total real: adding the number of variable loci after filtration from both GC and AC methods revealed 564 real variable loci

Perl scripts were used to extract and manipulate the DNA sequences of the IGRs of the 25 isolates of CC-174 (See chapter two 2.5.4.2 figure 2.7). The intergenic sequences of isolate ID: 27497 were used as the reference genome. The filtering process removed mainly regions with mismatches in the alignment and repeat sequences. The filtration process also detected putative variation in CEs and SSR (phase variation), therefore the overall real variable IGRs detected using Perl scripts was 223 loci (Table 5.3).

Table 5.3: List the real variable IGRs detected using using Perl scripts in paired comparison between isolates.

Isolates	Putative variable IGRs	Number of filtered loci with SSR or indel	Number of filtered loci with CE repeat	Real variable IGRs
20115_20232	135	6	18	111
20282_20285	122	7	13	102
20308_20310	122	4	17	101
27497_27506	134	7	17	110
27523_27526	119	8	20	91
27530_27538	106	5	23	78
27539_27540	91	8	19	64
27541_27556	111	4	38	69
27572_27578	109	8	18	83
27580_27581	89	8	17	64
28250_28252	132	10	35	87
28253_20072	90	10	17	63

Note: total real IGRs among 25 isolates was1023 however overlapping IGRs among 25 isolates was 800, this lift 223 real IGRs.

After extracting the variable genes/ IGRs from 25 CC-174 isolates, the entire assembled genome of ID: 27497 was used as a reference genome for building entire assembled genomes of 25 CC-174 isolates carrying the real variation (SNPs) of genic and IGRs (For detail information see section 2.5.4.3). Therefore, all the analysis which include detection of recombination, construction of phylogenetic trees and phylogenetic network, detection of SNPs density and comparison of recombination events between disease and carriage isolates were carried out using entire assembled genomes of 25 CC-174 isolates carrying the real variation (SNPs) of genic and IGRs.

5.3 Interpretation the recombination events in the disease and carriage isolates of CC-174

Each contig containing the real variation in multifasta sequences of the 25 CC-174 isolates, among the 67 contigs was submitted to the ClonalFrameML program to estimate recombination events. The output of ClonalFrameML program indicated that recombination occurred with a similar frequency as mutation, however more of the divergence was caused by recombination than by mutation, and recombination introduced more substitutions than mutational processes confirming the importance of recombination events in the microevolution of these CC-174 isolates (Table 5.4).

Table 5.4: List the parameters obtained from ClonalFrameML program.

Parameters	Current study
Average recombination fragment length (I)	382.6 [352.2-429.5]
Average divergence (D) of donor and recipient	0.0442 [0.0432-.0452]
Ratio of the rate of recombination to mutation (R/theta)	0.849 [0.735-0.966]
Relative effect of recombination and mutation to r/m = (R/theta*I*D) (Number of substitution in recombination relative to mutation)	14.3
average range of substitution (DI)	17

The unit of average recombination fragment length is bps

Two types of recombination events were recognized, recombination events occurred in single isolates with 177 out of 284 and recombination events occurred on node of branch in multiple isolates and hence on a terminal branch or root of tree into terminal branch with 107 out of 284.

The role of recombination in microevolution of CC-174 isolates can also be observed in the difference in branches lengths between the ClonalFrameML tree and the PhyML tree. It has been reported that trees constructed by ClonalFrameML are informative about true topology regardless of whether they include recombination or not. The reason for this is that the model used for estimating recombination in ClonalFrameML takes account of external origin in addition to comparing species, so the substitutions determined on a given

branch are shared by all the genomes that descend from that branch (Didelot and Wilson, 2015). In this case, the ClonalFrameML tree showed differences in the branches lengths compared with the tree constructed using maximum likelihood algorithms (i.e. the PhyML program). In both trees, all isolates form different clusters and aggregate in two groups with the same topology. Group1 constitutes C1 cluster and group2 consists of three clusters (C2, C3 and C4) while the rest of the isolates are located in outlier positions corresponding with different clusters. The scale of branch length in the PhyML tree was significantly greater than in the ClonalFrameML tree because the former accounts for the substitutions introduced by both recombination and mutation. The significant finding from the tree construction was that although the isolates in both trees did not show clustering with year of isolation, the historical isolate ID: 27497 tended to be in an outlier position among 18 carriages isolates that were collected from same place (Nottingham university) suggestive of within-cluster evolution over time (Figure 5.1). The C1, C2, and C3 clusters contained disease and carriage isolates whereas the C4 cluster contained only carriage isolates, the recombination events that occurred in single isolates from clusters C1, C2, and C3 were estimated to be 100/177 (56.4%) while the number of events was 16/177 (9%) in the C4 cluster and 61/177 (34.6%) in the three outlier isolates (IDs: 20308, 27497 and 20115). Many recombination events occurred on the branch nodes of multiple isolates with 107 out of 284 events. 72 out of 107 (67.2%) of the recombination events in the nodes occured in cluster (C4) due to the fact that all carriages isolates in this cluster were collected at the same time and from same place (Nottingham University).

The phylogenetic tree alongside the recombination events were also constructed using ParSnp package from the whole genome sequence of 25 CC-174. The percentage of core genome alignment among all sequences of 25 CC-174 using ParSnp was 99.9% (2112164 bps out of 2114279 bps). The tree also confirmed that the historical isolate ID: 27497 tended to be in an outlier position among 25 isolates and the isolates within the C4 cluster were same as in ClonalFrame tree while the other carriage isolates in C1, C2, and C3 were mixed with disease isolates of CC-174 (Appendix 35).

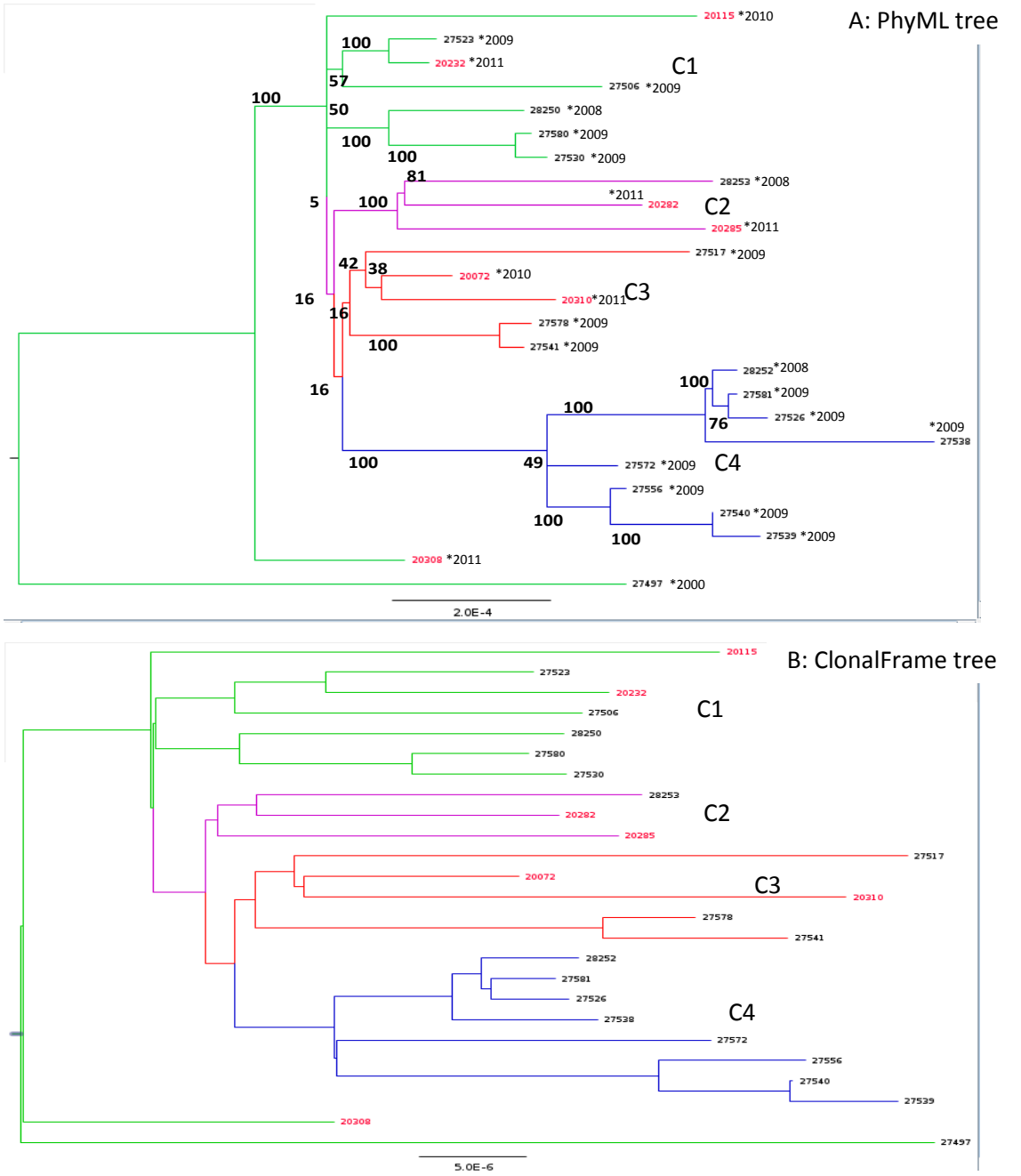


Figure 5.1: Phylogenetic trees constructed using ClonalFrameML and PhyML software. Panel A: PhyML tree was constructed using maximum likelihood algorithm from entire assembled genomes of 25 CC-174 isolates. **Panel B:** ClonalFrameML tree was constructed the ClonalFrameML program from entire assembled genomes of 25 CC-174 isolates. Each color depicts a particular cluster. Red color: disease isolates. Black color: carriage isolates. The C1-C4 are the number of clusters in each tree. Asterik : the year of isolation of each isolate.

Confirmation of changes in the topology of mutation and recombination trees was achieved as follows (Lamelas *et al.*, 2014); DNA sequences from the start to the end of each recombination region were extracted and concatenated into one file. In same way, the DNA sequences for the mutation regions were also extracted and concatenated. These sequences were used to construct trees with PhyML for both recombination and mutation regions (Figure 5.2). A comparison between these trees showed that five out of the 25 isolates exhibited differences in their tree topologies, three disease isolates with IDs: 20115, 20310 and 20285 and two carriage isolates with IDs: 27580 and 27530. In addition, in the mutation tree the isolates with IDs 20072 and 27517 cluster together. Many isolates had differences in the length of the branches of their nodes, and in the length of their terminal branches. The separation of different clusters was not associated with the year of isolation or disease versus carriages. Surprisingly in both the mutation and the recombination tree the apparent outlier position of the historical isolate ID: 27497 lost (Compare figure 5.1 and 5.2). Overall, therefore, it can be concluded that recombination has a major role in shaping the evolution of isolates of CC-174.

An example of an alignment of DNA sequences showing patches of recombination among different isolates can be viewed in (Appendix 36).

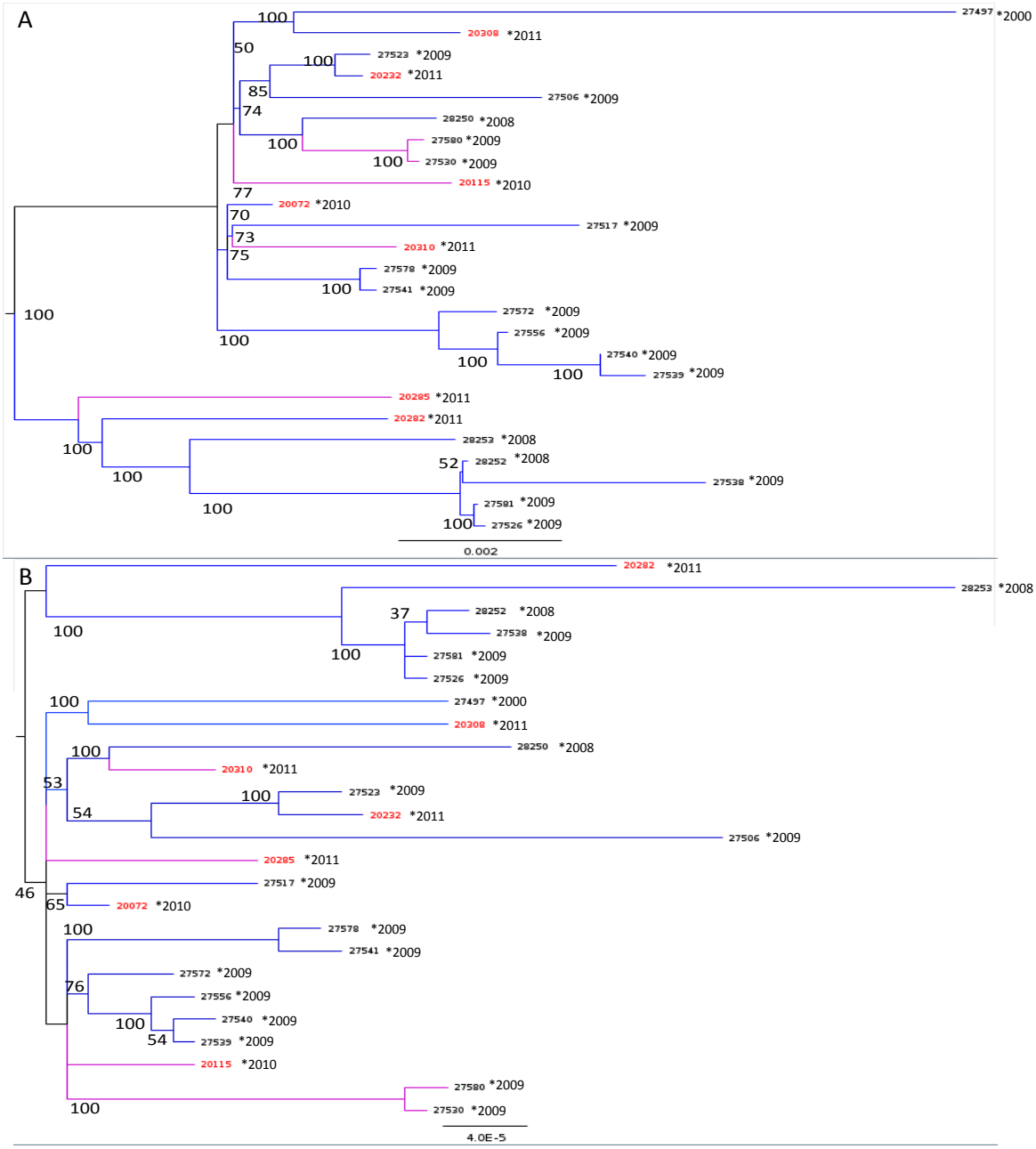


Figure 5.2: Phylogenetic trees for different sources of genetic variations. Panel A: Recombination tree was constructed using maximum likelihood algorithm in the PhyML program from recombinant SNPs of entire assembled genomes of 25 CC-174 isolates. **Panel B:** Mutation tree was constructed using maximum likelihood algorithm in the PhyML program from mutant SNPs of entire assembled genomes of 25 CC-174 isolates. Red color: disease isolates. Black color: carriage isolates. Changes in the positions of the isolates are indicated by purple coloured lines. Asterik : the year of isolation of each isolate.

As mentioned previously, the total number of recombination events detected by ClonalFrameML was 284. Isolate ID: 27497 had the highest number of events (40). Regions that share more than three recombination signals were considered as recombination hot spots, and 24 such regions were detected using ClonalFrameML with multiple imports on several branches. Large importations of more than 3 kbp were detected in 10 regions and isolate ID: 27497 had the highest number of hot spots (n=11) while the largest importation events were in disease isolate ID: 20282. Twenty three of the 25 isolates had recombination events, while 14 out of these 23 showed recombination events with both hot spots and large importation regions. Eleven out of 25 isolates (5 disease and 6 carriage) constitutes around 161/177 (90.9%) from total recombination events that occured on single isolates (Appendix 37).

The size of the recombination blocks ranged between 9 and 5508 bp, with a mean length of 2758 bp. The total length of recombination fragments was 180171 indicating that 8.37% of the total genome was subject to recombination events. The total length of recombination fragments in the core genome was 172598 and total length of core genes was (1663890) indicating that 10.3% of the total core genome was subject to recombination (Figure 5.3).

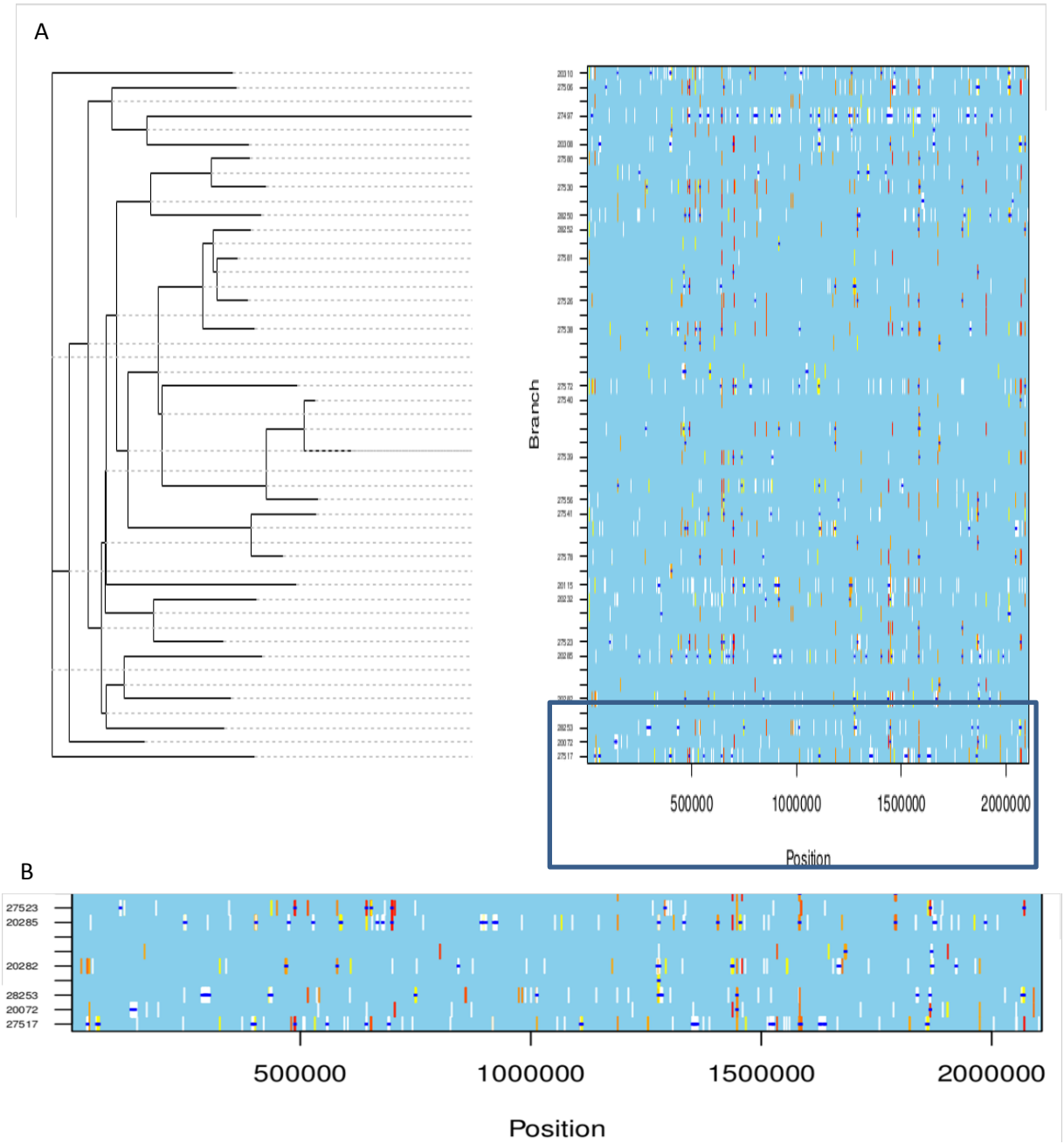


Figure 5.3: The recombination events detected using ClonalFrameML program for the 25 CC-174. The blue dots refer to recombination with different events as follows; blue dots in white bar: showing the recombination events mapped by position in the genome across all the isolates of CC-174. Blue dots in red bar: showing the recombination events on terminal branches of shared isolates. Blue dots in yellow bar: showing the recombination events from terminal branch till the root of tree. The ClonalFrameML tree: showing the relationship among recombinant isolates and their branches. **Panel A**: complete figure, **panel B**: zoom in to the area under the square of panel A.

5.4 The correlation between recombination, synonymous and non-synonymous variation in the CC-174

Although the number of genes subject to mutation events (n = 377; 66.8%) was higher than those subject to recombination events (n = 187; 33.1%) in the genomes of CC-174 isolates, nevertheless, the genome variation in the recombinant genes comprised 5582 (66 %) of the SNPs, compared to 2869 (34%) attributed to the mutant genes. Substantial variation was therefore associated with the recombination events. As the number of substitutions was higher in recombinant as compared to mutant genes, it was important to check the SNPs density, which is the number of SNPs, divided by the total number of nucleotides. The density was higher for the recombinant genes (0.026) than for the mutant genes (0.006).

In this context, the SNAP script was used to detect the dN/dS ratio for the recombinant genes. The dS was 13.9 times higher than dN. Therefore, the dN/dS ratio was 0.15. The nucleotide diversity was checked for the synonymous and non-synonymous polymorphisms in the recombinant and mutants genes. In general, nucleotide diversity was quite high in recombinant genes with (0.01) comparable with mutant genes with (0.0028). The diversity for the synonymous polymorphisms was slightly higher in both recombinant and mutant genes with 0.006 and 0.001 than observed for non-synonymous polymorphisms with 0.0041 and 0.0009 respectively. This may relate to the fact that recombinant genes were experiencing purifying selection.

The type of selection acting on recombinant variable genes was estimated using the method described in chapter two (2.6.3). The number of recombinant genes showing purifying selection was (77.5%; 145/187) while two genes showed positive selection. These genes were *NMB0071*, which codes for a part of the transporter capsule (CtrA) and *NMB1605*, which codes for topoisomerase IV subunit A.

A phylogenetic network was then constructed for the concatenated sequence of recombinant genes using SplitsTree4. Reticulation events were observed indicating the presence of recombination. All the branches were supported by bootstrap analysis and the disease isolates showed reticulation as well carriage isolates (Figure 5.4).

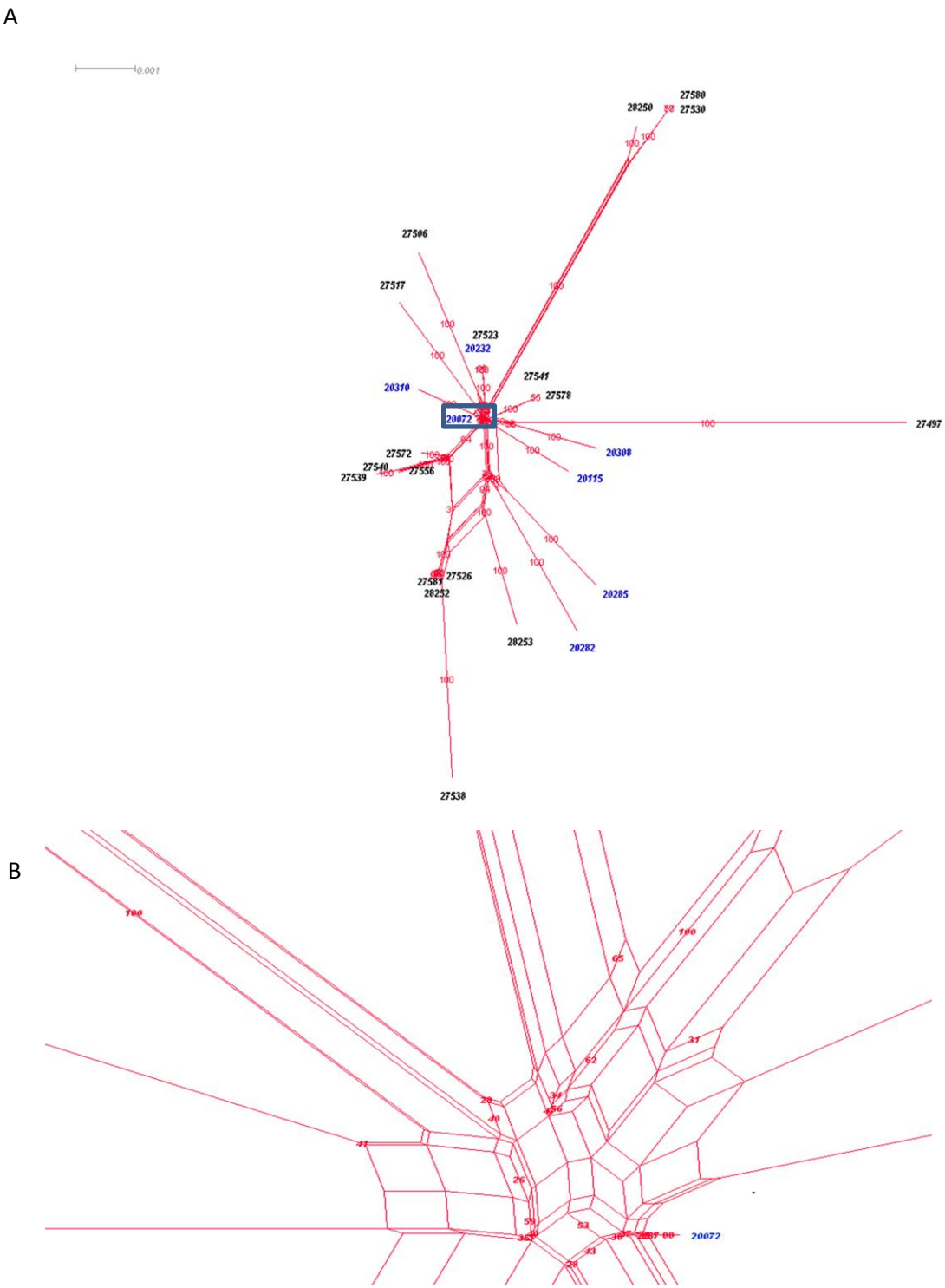


Figure 5.4: A phylogenetic network for the concatenated sequences of recombinant genes of CC-174 isolates was constructed using NeighbourNet algorithms in SplitsTree4. Panel A: the network with reticulation for all isolates. **Panel B**: zoom in to the area under the square of panel A. Disease isolates depicted by blue color and carriage isolates by black color.

5.5 Interrogation of recombination events within the IGRs of CC-174 isolates

Recombination was examined in the IGRs and 21 (9.4%) were observed to exhibit variation which was lower than 187 (33.1%) of genes that experienced recombination. Eleven isolates showed recombination in their IGRs, with the highest number of events being in isolate ID: 27517. Three disease isolates had recombination events in the IGR (Table 5.5).

For inferring a complete picture of the evolution of CC-174 isolates, the number of genic regions with indels was compared with the number of IGRs with indels. The results showed 38 indels in the IGRs compared to 18 in the genic regions (excluding indels in phase variable loci). Chi-square analysis showed that there was a significant difference with a P value of 0.001. The low number of indels in the CDs regions may relate to selection against frame shift mutation in these sequences.

Table 5.5: Recombination events detected in the IGRs of 25 CC-174 isolates using ClonalFrameML program.

Isolates ID or node number	Starting point	Ending point	Size	Number of SNPs	Intergenic name
27541	5277	5322	45	2	INT NEIS0034
20232	3319	3449	130	14	INT NEIS0606
20308	1936	2070	134	8	INT NEIS0402
20310	16881	16975	94	9	INT NEIS0816
27497	699	1874	1175	37	INT NEIS1706
27497	8721	9730	1009	65	INTNEIS0487
27506	18513	18592	79	9	INT NEIS0814
27541	21600	21630	30	15	INT NEIS1782
27517	7059	7521	462	6	INT NEIS0062
27517	2583	3077	494	12	INT NEIS0978
27517	13758	13865	107	16	INT NEIS2133
27572	41911	42156	245	4	INT NEIS0612
27580	26875	26906	30	9	INT NEIS1428
28250	29813	29961	148	14	INT NEIS0366
28250	4813	4965	152	18	INT NEIS0897
28253	16305	16579	274	14	INT NEIS0354
28253	9935	10022	87	3	INT NEIS0068
NODE_26	2037	2187	150	14	INT NEIS0402
NODE_28	9779	9801	22	6	INT NIES0486
NODE_35	28436	28526	90	12	INT NEIS1777
NODE_39	8190	8306	116	21	INTNIES1691

5.6 The function of recombinant and mutant variable genes

The functional protein categories for recombinant and mutant variable genes were identified using the KEGG classification system. As mentioned previously, the proportional effect of variation on gene function was calculated by dividing the number of variable genes by the total number of genes in each class. For recombinant genes and mutant genes, signaling and cellular processing genes had the highest number of variable genes at 16.2% (39/240) and 28.3% (68/240) respectively, while metabolism genes for both variation types were the second with 13.2% (99/752) and 24.5% (184/752) respectively. Each functional scheme was also tested for genes under purifying selection. This also showed that signaling and cellular processing genes was the highest with 12.5% (30/240), while metabolism genes had the next highest levels with 11.6% (87/752) (Figure 5.5).

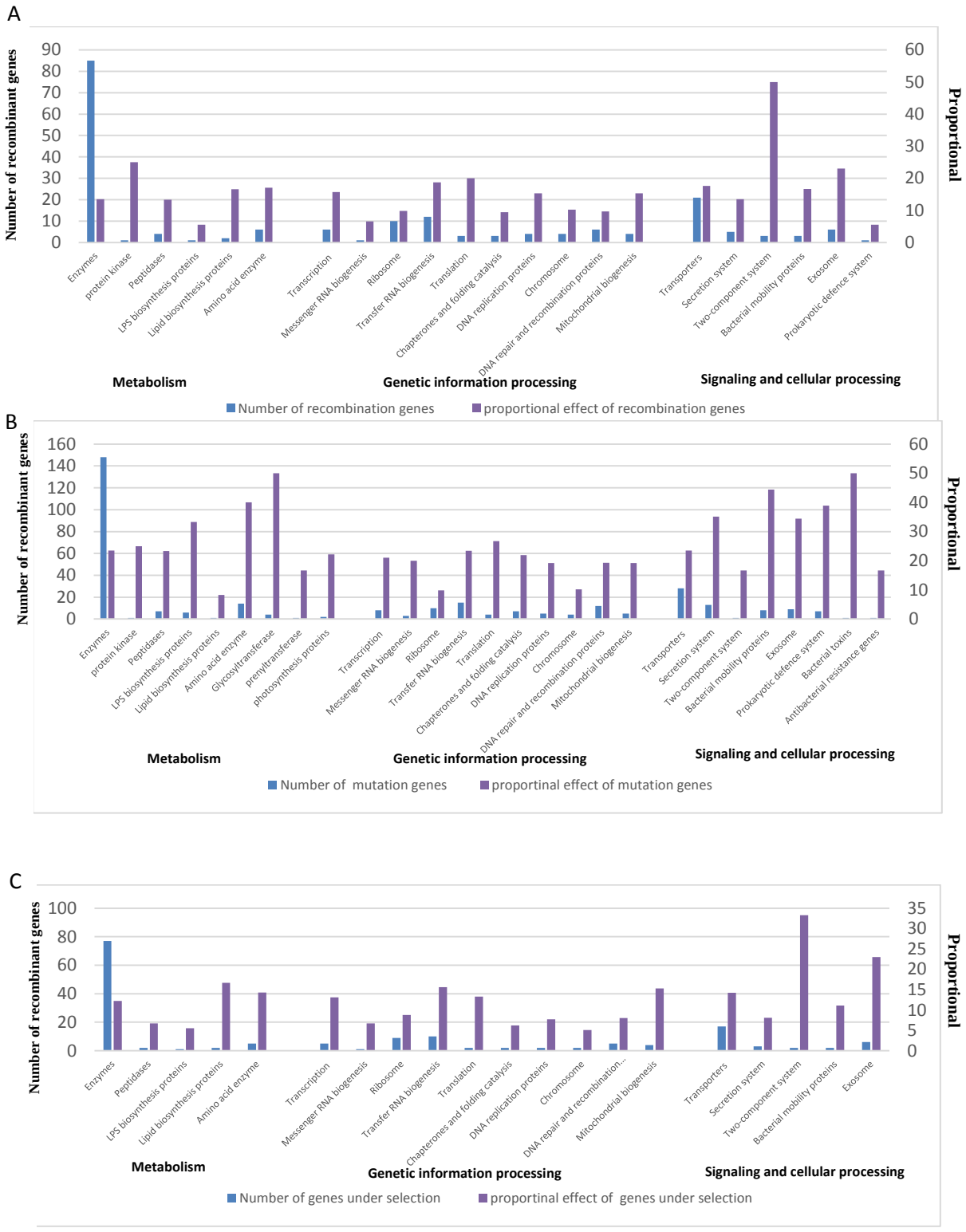


Figure 5.5: The proportional effect of the functional protein categories of different schemes for recombinant, recombinant with purifying selection and mutant genes for CC-174. Panel A: recombinant genes. **Panel B**: mutant genes **Panel C**: recombinant genes under purifying selection. All the schemes were detected using the KEGG classification system for 7 disease and 18 carriage isolates.

Analysis of the proportion of recombinant genes in each functional group for carriage and disease isolates showed that 9.2% (22/240) and 5% (12/240) of variable genes respectively were again in the scheme for signaling and cellular processing and 7.6% (57/752) and 4% (30/752) in the metabolism scheme (Figure 5.6).

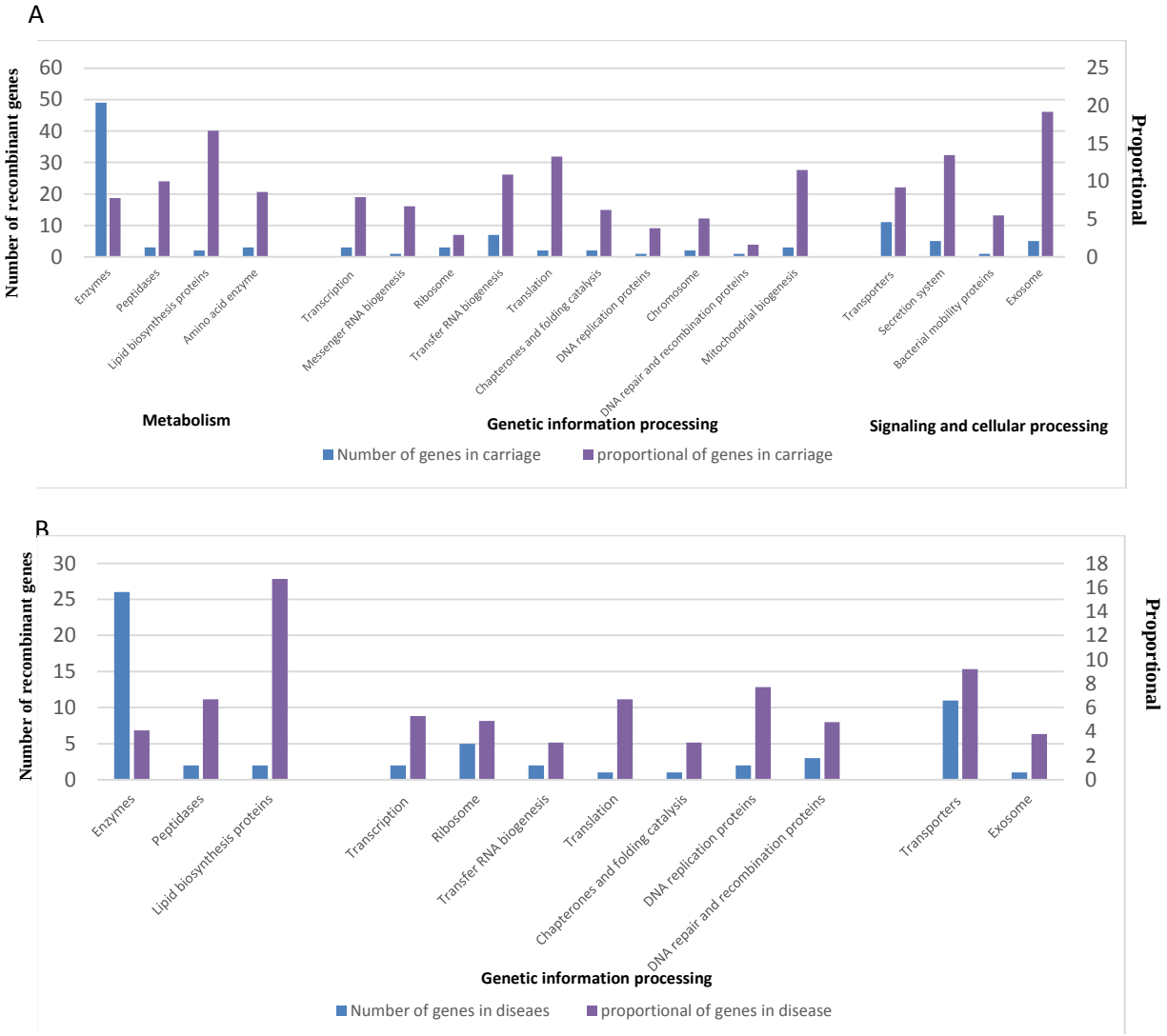


Figure 5.6: Classification of recombinant genes by functional protein categories for disease and carriage isolates of CC-174. Panel A: the total and proportional effect of all genes in 18 carriage isolates for different schemes. **Panel B**: the total and proportional effect of all genes in 7 disease isolates for different schemes. All the schemes were detected using the KEGG classification system.

The proportion of the functional categories of all hot spot recombinant and large importation recombinant genes was the highest in the metabolism with 5% (38/752) then signaling and cellular processing schemes with 4.2% (10/240) (Figure 5.7).

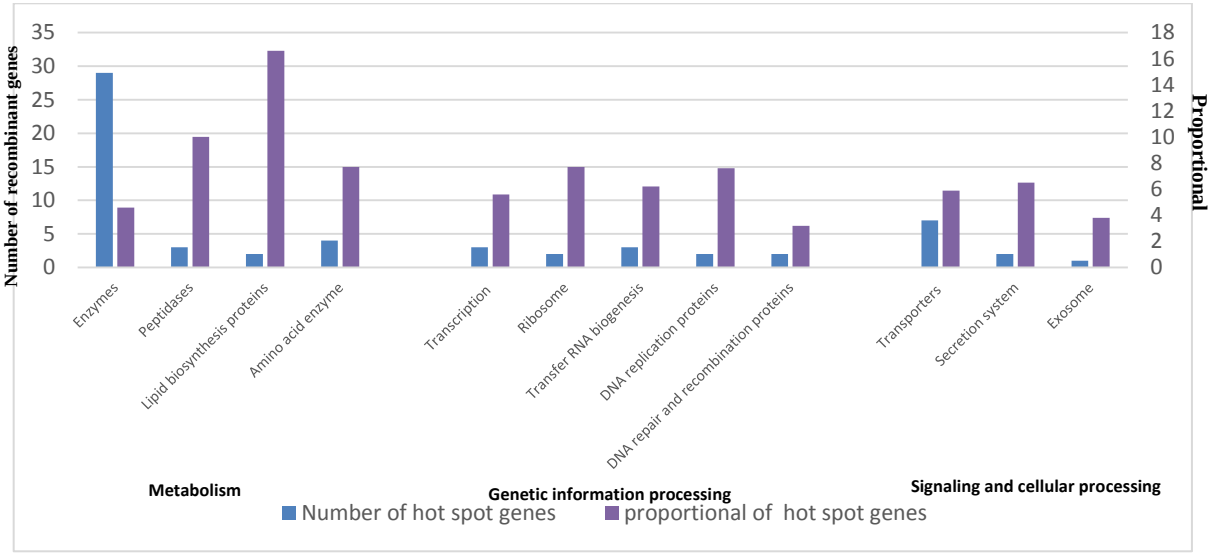


Figure 5.7: The proportional effect of the functional protein categories of different schemes for hot spot and large import fragments of recombination for CC-174 isolates. The proportion effect of host spot and large import fragments of recombination were detected in 18 carriages and 7 disease isolates of different schemes.

In general, the proportion effect of all recombinant, mutant, recombinant under purifying selection and hot spot recombinant and large importation recombinant genes in both disease and carriage isolates showed that there was no a big difference among all the schemes. However, there was a slight increase in the signaling and cellular process scheme for recombinant, mutant, recombinant under purifying selection and in the metabolism scheme for hot spot recombinant and large importation recombinant genes.

Comparison of recombination events between disease and carriage isolates was carried out. The amount of recombination in each of these groups was estimated by counting the number of recombinant SNPs in disease versus carriage isolates. The length of the concatenated DNA sequence of non-hot spot recombinant genes of all isolates belonging to disease isolates was adjusted to be the same as that for all the carriage isolates. The length of concatenated DNA sequence of hot spot and large importation recombinant genes was

also adjusted in a similar way. The number of isolates was also adjusted to be same in both disease and carriage isolates. Counting of the SNPs in the hot spot recombinant genes and large importation events indicated that disease isolates contained (n=1,220, 72%), while carriage isolates had (n=470, 28%). Moreover, the SNPs in other recombinant genes was (n=819, 75%) for disease isolates and (n=27, 25%) for carriage isolates. Therefore, the number of SNPs introduced by recombination was higher in disease isolates than carriage isolates. The profile distribution of the number of recombination events between disease and carriage isolates was investigated. In the first instance, the distribution was higher in the carriage isolates with 126 recombination events comparable with 51 events in disease isolates, however normalization was carried out to adjust for the number of isolates to be same in both disease and carriage isolates, therefore the number of events was 49 in carriage isolates after normalization. This indicates that both events are approximately equal.

5.7 Summary of main findings

In general, *N. meningitidis* has a high rate of recombination with a high level of diversification of clones, however, Bart *et al.* (2001) showed that serogroup A of *N. meningitidis* had a truly clonal population structure with a very low level of diversification of clones. Relying on the study carried out by Oldfield *et al.* (2016) which showed that there was an extensive similarity between 7 disease and 18 carriage isolates of *N. meningitidis* CC-174 the aim was to investigate how recombination and genomic plasticity within a single lineage of *N. meningitidis* affects microevolution of the isolates belonging to one CC and if there were differences between disease and carriage isolates.

Recombination was detected to have an average fragment length (I) of 382.6 bp, an average divergence of donor from recipient (D) of 0.04, and finally a ratio recombination to mutation (R/theta) of 0.84 with an average range of substitution of 17. From these parameters, the relative effect of recombination and mutation r/m = (R/theta*I*D) was calculated to be 14.3. This gives strong evidence for the fact that recombination is driving a substantial amount of the variation among these isolates. This contrasts with other species as can be observed clearly, by comparing our result with other studies. The study carried out by Didelot and Wilson (2015) on 86 genomes of *Clostridium difficile* showed that the average divergence of donor from recipient of 0.03 and (R/theta) of 0.3 and r/m of 5.6. Another study carried out by the Didelot and Wilson (2015) on 110 genomes of *Staphylococcus aureus,* showed that the average divergence of donor from recipient of 0.007 and (R/theta) of 0.2 and r/m of 0.27. However, many studies on *N. meningitidis* and based on MLST genes showed the r/m was higher comparable with what has been observed in this study and some studies showed values higher than 100 (Didelot *et al.,* 2009). The low value in the r/m in our study comparable with others accounts for the high similarity among the isolates under the current study, which all belong to the same CC, i.e. CC-174 and all these carriage isolates were collected in the same year and place suggestive of intra group transmission.

Another important piece of evidence for recombination being the source of a substantial amount of variation is that there were changes in the branch lengths of most isolates within the ClonalFrameML tree as compared with the PhyML tree. In addition, there was changes

in the topology and branch lengths of most isolates within the recombination tree as compared with the mutation tree. While in the study carried out by (Didelot and Wilson, 2015), this comparison between trees showed differences in branch length of only a few isolates. Moreover, In spite of having only 33.1 % (187/564) of variable genes being subject to recombination, and similarly only 9.4% (21/223) IGRs, the SNPs density was highest in the recombinant genes with 0.026 as compared to 0.006 in the mutant genes.

To elucidate the role of recombination events in forming different population structure, the trees comparison showed that recombination enhances changes the topology of three disease isolates with IDs: 20115, 20310 and 20285 and two carriage isolates with IDs: 27580 and 27530.

Another crucial phenomenon observed from phylogenetic comparison was that the population structure of disease and carriage isolates did not cluster regarding the year of isolation, however long-term temporal stability was inferred due to the presence of the isolate with ID: 27497. This isolate was ancestral of other isolates and indeed, this isolate showed the highest number of recombination events of (40) and recombination hot spot of (11) suggesting diversification between this clone and other isolates over the 8 year time difference between when it was isolated at 2000 and the others in (2008-2011).

The recombination events were detected in single isolates or on node of branch in multiple isolates. The key finding was that the number of recombination events in single isolates highest with 56.4% among isolates of the C1, C2 and C3 clusters which contain temporally overlapping carriage and disease isolates whereas only 9% of events were found for the carriage isolates of C4 cluster. However, the isolates in C4 had 72 out of 107 (67.2%) of recombination events contained in the nodes in multiple isolates as they are collected at the same time, place and are spatially and temporal related.

Around 10.3% of the core genome was subject to recombination events. Comparing our result with other studies, Yu *et al.* (2014) showed that around 62.7% to 98.4% of the core genome of Neisseria was subject to intraspecies and interspecies recombination events for (14 *N. meningitidis*, *3 N. gonorrhoeae* and 1 commensal *N. lactamica*. On the other hand, Joseph *et al.* (2011) found that 30 *N. meningitidis* isolates from different CCs showed

recombination in 39.6% of core genome. The current study detected a low percentage of recombinant genes in the core genome comparable with these studies. As mentioned previously, this is because the high similarity among the isolates under the current study forms tightly cluster and may have high levels of one specific highly transmissible clone.

The dN/dS ratio in recombination SNPs was 0.15 and there were 13.9 times more dS than dN. In addition, nucleotide diversity in synonymous polymorphism 0.006 of recombinant genes was 1.5 times more than nucleotide diversity in non-synonymous polymorphism (0.004) which may reflect the action of purifying selection. Lamelas *et al.,* (2014) found 2.3 times more synonymous than non-synonymous recombination associated SNPs due to purifying selection. The study carried out by Castillo-Ramirez *et al.* (2011) showed that the dN/dS ratio for recombinant fragments was 0.12 comparable with the whole genome with 0.33 in *S. aureus.* Therefore, the author argues that recombination diminishes the dN/dS ratio and this is compatible with the finding in the current study.

The selection test showed that 77% of recombinant genes were under purifying selection diminishing the effect of deleterious mutation in the population. While there were two genes under positive selection, which are *NMB0071* coded for capsule polysaccharide export outer membrane (CtrA) and *NMB1605* coded topoisomeraseIV subunit A. Changing the CtrA protein may establish resistance against the immune system as it is on the outer surface. While topoisomerase IV may lead to establishment of fluoroquinolone resistance and increase antibiotic resistance (Shultz *et al.*, 2005). The action of purifying selection on recombinant genes was also shown in the study carried out by Yu *et al.* (2014) who showed that there were 635 recombinant genes and only 10 genes that showed positive selection. These genes mainly coded for DNA processing and iron uptake functions. Indels were found to be significantly higher P-value 0.001 in IGRs with 38 as compared with genic regions with 18. This is presumably because of selection against frame shifts mutation.

In general, the proportional effect of each scheme of function for recombinant genes, recombinant genes under purifying selection, mutant genes in both disease and carriage isolates showed that signaling and cellular processing scheme was the highest comparable with other schemes but with no significant difference. In general, the genes in the signaling and cellular processing scheme mainly code for membrane proteins (transporters and

secretion system), bacterial mobility proteins, two-component system, toxin, antimicrobial resistance and defense proteins. Changing OMPs may lead to the establishment of resistance against stress conditions. Similarly, the genes coding for bacterial mobility proteins, two-component system and toxin, have major roles in virulence and the establishment of host colonization. While antimicrobial resistance and defense proteins may have essential physiological activities in respect to defense against host immune system and antibiotics. In comparison, a study was carried out by Yu *et al.* (2014) to infer recombination and selection patterns for some Neisseria species. Their study showed that two schemes (1) replication, recombination, and repair (2) translation, ribosomal structure and biogenesis -exhibited high diversity among isolates.

As there was no significant difference in all functional schemes, therefore the function of recombinant gene with hot spot or large importation may be more important than other schemes in shaping the evolution of isolates through recombination.

The variable genes detected within the hot spot recombination and large importation events have different roles in the physiological activities of *N. meningitidis*. Firstly, variation was detected in one group, coded for systems involved in uptake of iron, sulfur and other compounds. In this group, the *NMB0293*, *NMB0461*, *NMB1381*,*NMB0879* and (*NMB0880-NMB2026*) code for TonB-dependent receptors, transferrin-binding proteins, iron-sulfur cluster assembly protein (IscA), sulfate ABC transporter- ATP-binding protein and sulfate ABC transporter- permease respectively. Genes in this class are subject to variation and are discussed as examples of how variation may affect host colonization by *N. meningitidis*. TonB-dependent receptors that belong to iron acquisition genes and are outer-membrane proteins (Turner *et al.*, 2001). The transferrin-binding proteins are iron acquisition genes which are located on the outer-membrane. These proteins support growth of *N. meningitidis* by the acquiring iron through two transferrin-binding proteins, TbpA and TbpB (Stokes *et al.*, 2005). The iron-starvation protein PigA coded by *NMB1669* is located upstream of *hmbr* gene and it codes for the heme oxygenase. This protein has a major role in release of iron from imported heme. Therefore, it is important for heme, hemoglobin (Hb), and haptoglobin-Hb utilization (Kahler *et al.*, 2001; Zhu *et al.*, 2000). The sulfate ABC transporter, ATP-binding protein and sulfate ABC transporter, permease proteins have a

major role in catalysis of the transfer of sulfur from one side of a membrane to the other. These proteins may be involved in the import of sulfur to provide essential nutrients to this bacterium (ID: Q9JZW0 UniProt).

Secondly, variable genes coded for DNA processing enzymes or proteins that are essential for DNA stability. In this class, variation was seen in *NMB1384*, *NMB1605*, *NMB0453*, *NMB0697*, *NMB1537* and *NMB1873* that code for the DNA gyrase subunit A, topoisomerase IV subunit A, MutT protein, dimethyladenosine transferase, DNA primase and DNA polymerase respectively. The DNA gyrase subunit A has major role in DNA stability therefore, it is important for controlling variation in genomic DNA (Yu *et al.*, 2014). It has been reported that mutation in DNA gyrase or topoisomerase IV may lead to antibiotics resistance (Shultz *et al.*, 2005). The MutT protein controls the mutability of genes and can influence establishment of resistance to some antibiotics (Davidsen *et al.*, 2007). The methylation genes that code for dimethyladenosine transferase cause variation in RMSs and are important because this variation may reduce barriers to DNA exchange and increased rates of recombination in other genes (Kong *et al.*, 2013).

Thirdly, the variable genes code for some enzyme activities that are essential for physiological activities of *N. meningitidis*. *NMB1428* gene code for the aminopeptidases is discussed as example for this class. Generally, aminopeptidases help in the maturation, activation, or degradation of proteins in *N. meningitidis* (Nocek *et al.*, 2008). In addition, it has been reported that variation was seen in *NMB0877* coded for penicillin-binding proteins. It has been reported that the penicillin-binding proteins reduce binding affinity of penicillin G into penicillin-binding protein and establish to one class of antibiotics resistance (Mendelman *et al.*, 1988).

The conclusion is that most of the variable genes belonging to hot spot recombination or large importation events, mainly coded for DNA processing systems, proteins and enzymes that are in direct contact with the external environmental, as well as host niches. This result was compatible with the study carried out by Yu *et al.*, (2014). Thus, recombination and selection may play an important role as driving forces in the microevolution of isolates to specific host niches.

Chapter 6: Analysis of genome variation in CEs and SSR of meningococcal isolates

6.1 Introduction

N. meningitidis has high abundance of different types of repeats. The repeats have many functions some of which may aid survival of *N. meningitidis* in the host. There are many different types of repeat patterns such as SSR, CEs, NIME and different types of repetitive extragenic palindromes (REP2, 3, 4, 5) (Bentley *et al.*, 2007). CEs have a major role in controlling expression of different genes through sequence motifs that include the Snyder promoter, Black promoter and binding site for integration host factor (IHF) (Siddique *et al.*, 2011; Snyder *et al.*, 2009). In addition, they can terminate the transcriptional process when inserted inside genes or can be considered as a target for intergenic recombination (Snyder *et al.*, 2009) (Figure 6.1). It has also been reported that CE may control expression of some small non-coding RNAs (Siddique *et al.*, 2011).

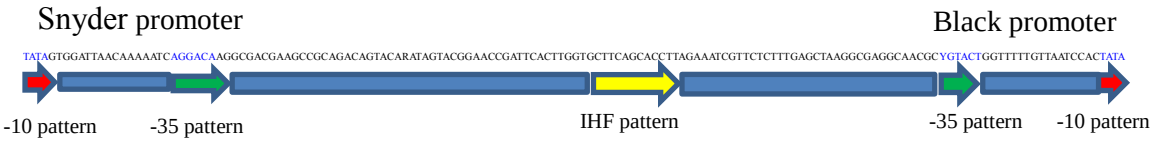


Figure 6.1: The complete CE with two promoters (Snyder promoter and Black promoter) and IHF. The Snyder promoter is on the left arm and Black promoter on the right arm while IHF pattern is in the middle. -10 pattern showed in red arrows, -35 pattern showed in green arrows and IHF showed in the yellow arrow.

CEs have an inverted repeat of 26 bp at both ends. CEs are distributed randomly across the non-coding RNA (Roberts *et al.*, 2016) but some researchers suggested that they are mainly found in IGRs of virulence, metabolic and transporter genes (Liu *et al.*, 2002; Snyder *et al.*, 2009). Some studies have shown that they can transfer between species by horizontal gene transfer that are a source of variation (Buisine *et al.*, 2002).

It has been reported that there are eight types of CEs. The classification of CE depends on two aspects. Firstly, a CE that has complete sequence is around 153 to 157 bp and contains two promoters and IHF site. A partial sequence CE, around 104-108 bp and contains only two promoters as there is a deletion of 50 bp in the middle of the sequence covering the IHF. This first aspect divided the CEs into alpha or beta with complete sequence or alpha prime and beta prime with partial sequence. Secondly, CEs contain three regions: left, middle and right arm. CEs divide into alpha and beta sequences, which are conserved patterns, located within the left arm or right arm of CEs. These conserved can be combined to give the eight types of CEs (Siddique *et al.*, 2011) (Figure 6.2).

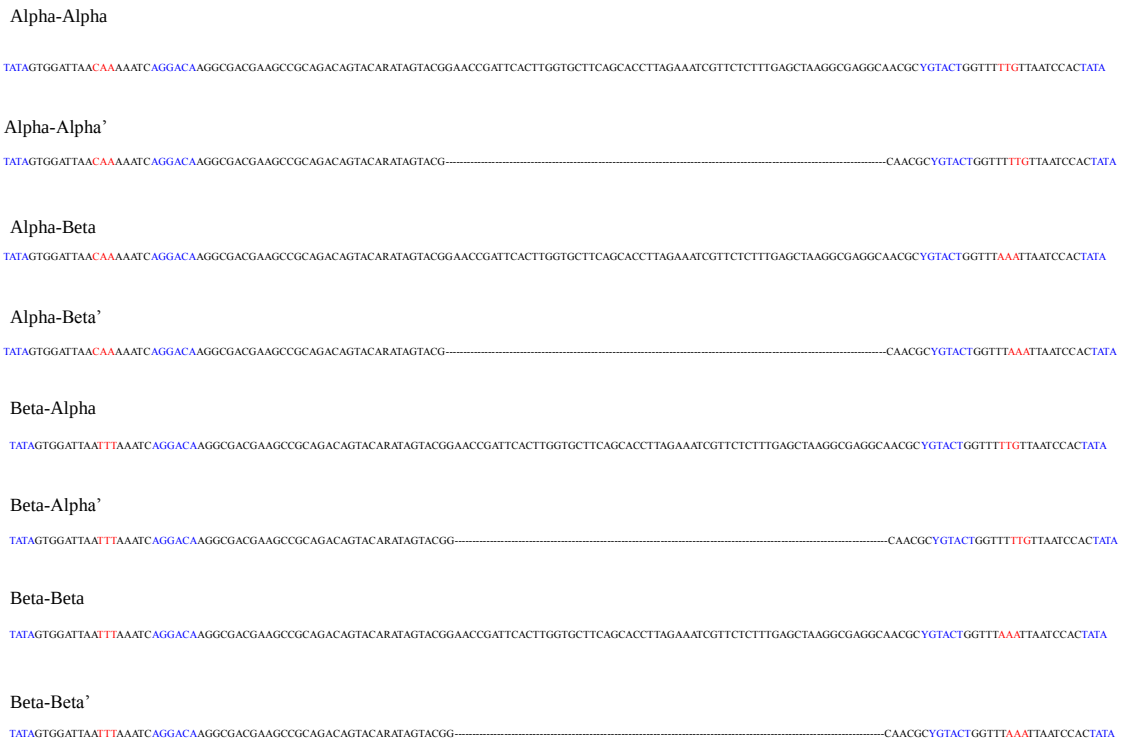


Figure 6.2: An overview of the different types of CEs and their DNA sequence. The complete CEs are around 157 bp while the partial CEs are around 104 bp. The red colour sequences show the difference in the alpha and beta patterns. The blue colour sequences show the patterns of Snyder and Black promoters.

The general aim was to analyse the variation in CE elements within two sets of persistence isolates. CEs are a source of variation between different strains therefore, it is necessary to analyse variation in CEs within our strains in their host that represents evolution over a small-time scale (for period of months).

6.2 Analysing the dynamics of variation in CE of IGRs of a pool of 40 isolates from four different time points of one carrier

The CEs were identified in the IGRs of the 40 carriage isolates using a series of scripts as set out in section (2.9).

The majority of variation was located within different repeat tracts especially CE with 71 intergenic loci. Some of the variation in CEs in the 40 isolates was not real but was due to poor assembly in the sequences of isolates. The spurious variation in CEs was observed through a BLAST search between variable CE patterns in old assembly and new assembly. In cases where there is no identity between old and new assembly in the position of variation, the presence of spurious variation in CEs was confirmed (See section 2.3). There were 32 loci out of 71 (45%) that carried CEs with poor assembly.

In first instance, most variation of CEs was found in partial CEs (14 out of 21) 66.6%, (13 out of 22) 59% and (13 out of 25) 52% in first, second and third time periods. Conversely, (15 out of 28) 54.5% variable CEs detected within the complete CEs in fourth time period. Variation in the CEs increased with time point with the fourth time period having a higher number of variable CEs in IGRs compared with other time periods. In addition, the variation of CEs patterns conducted with 3 out of 8 types of CEs in first time period while the panel of variation was shown in 5 out of 8 types of CEs in fourth time period (Figure 6.3). The Poisson mean test ($P < 0.001$) revealed that there was a significant difference in the number of variable CEs for the first against fourth time points.

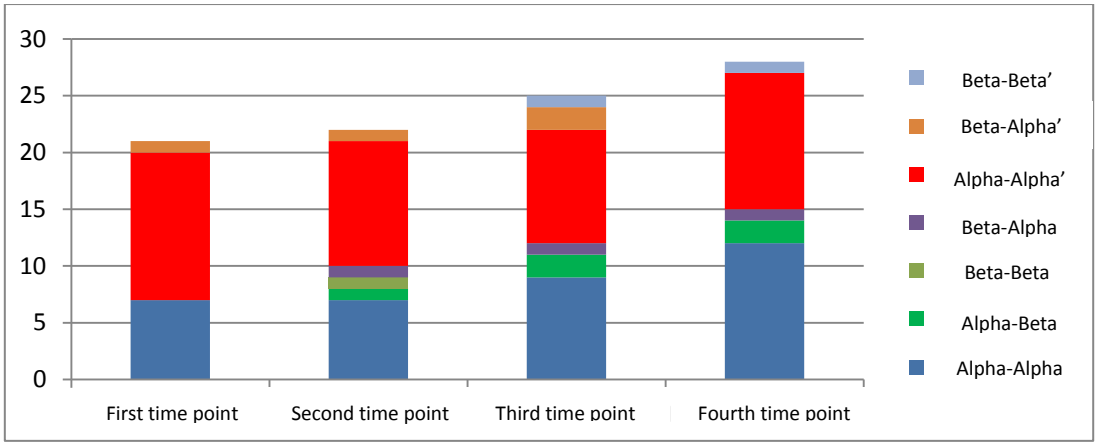


Figure 6.3: Variation in CEs for 40 persistent isolates of one carrier across four time points. Blue color: complete alpha-alpha, green color: complete alpha-beta, light green color: complete beta-beta, purple color: complete beta-alpha, red color: partial alpha-alpha´, orange color: partial beta-alpha´ and light blue color: partial beta-beta´. The number of CEs in IGRs of isolates were 197, 201, 203 and 200 while the number of variable CEs in IGRs were 21, 22, 25 and 28 in first, second, third and fourth time point respectively.

Although, there were no significant trends for increasing variation in CEs with time point, it would be that selection was only acting on specific CEs due to the fact that multiple isolates are affected for the same variable CE patterns with time periods. 20% or more of the population was affected by the same variable CEs with 9 out of 39 CEs being variable in multiple isolates (Figure 6.4).

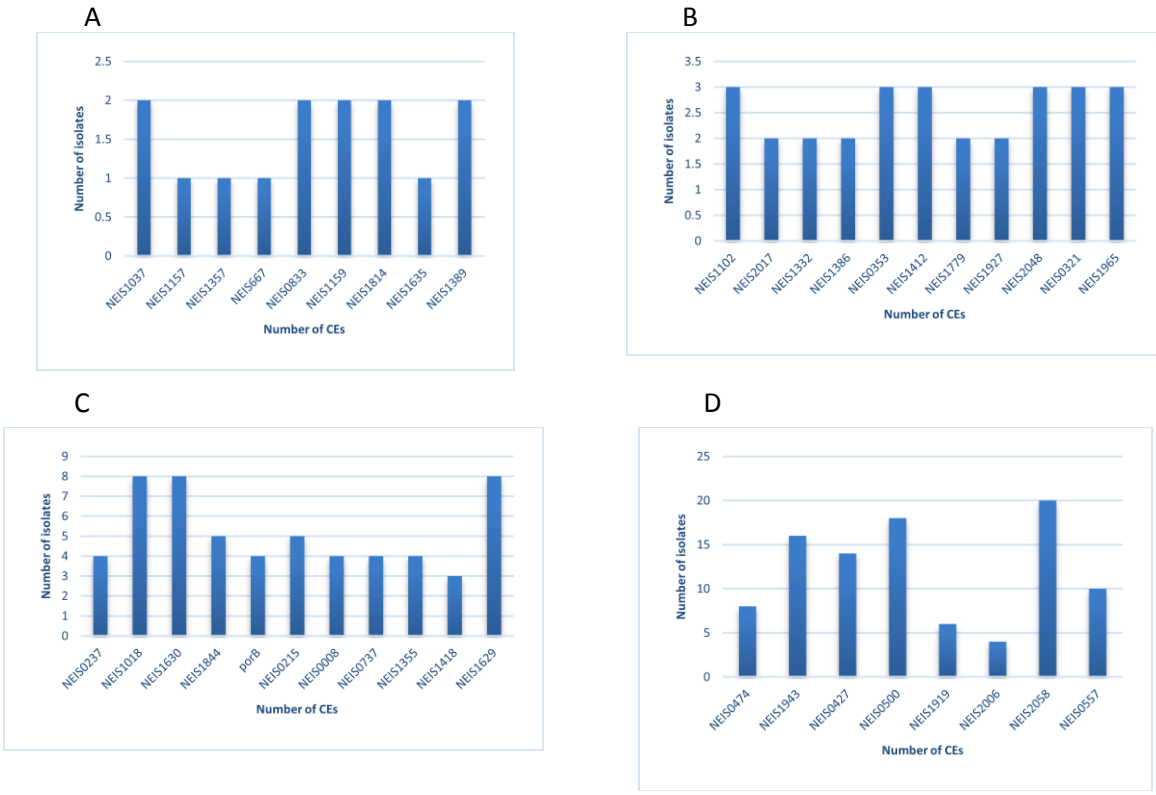


Figure 6.4: The level of variability for each variable CE in IGRs. Panel A, B, C, D represents the CEs having variation in one, two, three and four time point's respectively. The number of isolates is 10 for each time point. The CEs having variation in eight or more isolates affected 20% or more of population (40 isolates).

The location of variation within CEs is crucial as may be within the conserved functional patterns of CE. Variation was found within the -35 position of the first promoter (Snyder promoter) in 3 out of 39 (*NEIS0474, NEIS1629, NEIS2006*). Variation was also detected within the -35 position of the second promoter (Black promoter) in 1 out of 39 (*NEIS1630*). The changes in these positions may lead to a change in expression of genes. These genes mainly are transporter genes being variable in multiple isolates with time points. Interestingly, variation was also found within the IHF pattern of CE in 2 out of 39 (*NEIS0237* and *NEIS0008*) which may change the expression of genes through changing the strength of binding between transcriptional binding factors. These genes are metabolism genes being variable in multiple isolates with time points. However, the distance between

CEs that have variation within promoters regions or IHF binding site and the starting codon of nearby gene may indicate some variable CEs were not enrolled in changing the expression of their genes. Nevertheless, practical work has to be carried to confirm their role in effecting the level of transcription. High frequency of SNPs was seen in other positions within the CE patterns may indicate the change in these positions lead into effecting the gene expression. These positions are two nucleotides after the -10 pattern of Snyder promoter changing T to C, two nucleotides after the -35 pattern of Snyder promoter changing G to A, two nucleotides before the -10 pattern of Black promoter changing A to G and two nucleotides before the -35 pattern of Black promoter changing C to T (Figures 6.5, 6.6). From the above figure, the dynamic of variation showed that the overall variation in CEs in the IGRs was due to SNPs rather than indels.

CEs were conserved in the IGRs of the four different isolates; each one selected arbitrarily from a particular time point (isolates: 20879, 20896, 20905 and 20911). Some IGRs that carried CEs were missing particularly those at the end of the contig and due to the sequence not being complete in seven loci (*NEIS0060, NEIS1354, NEIS1457, NEIS1943, NEIS1944, NEIS1945,* and *NEIS1963)*. Therefore, there was no evidence of movement of CE between the different positions. The CEs can move into genes in the same way transposon like elements and can terminate the transcription of gene by forming loop like structure. The inserted variable CEs within the genic regions were detected in two genes that are *NEIS1702* and *NEIS0500* (Table 6.1). Both genes encode for hypothetical proteins, which may not have a role in persistence of *N. meningitidis.*However, inserted CEs within the genic regions of *NEIS1702* and *NEIS0500* found in all the 40 isolates with different time periods therefore there was no movement of CEs within the time points among 40 isolates representing at least 1-6 months host persistence.

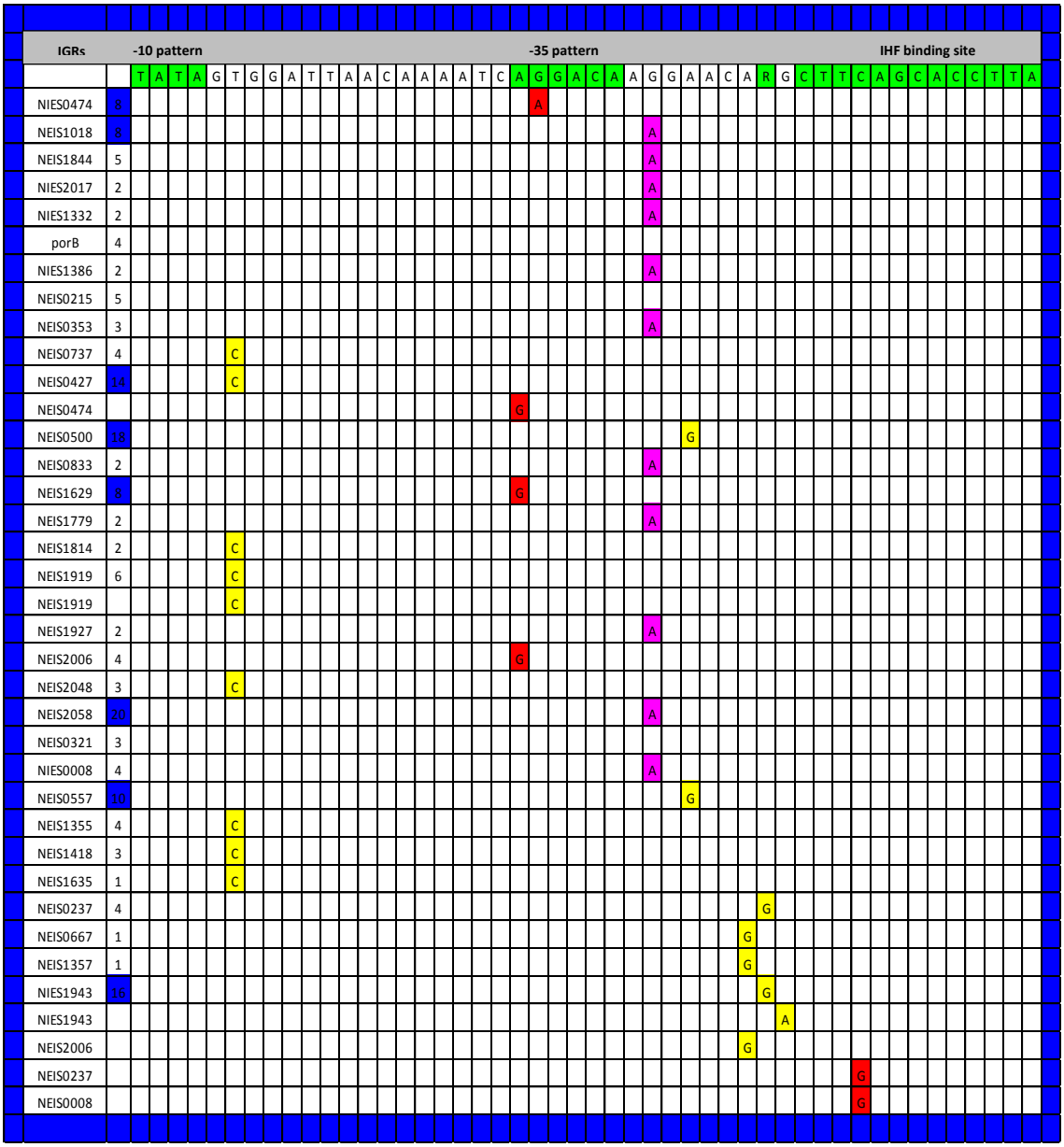


Figure 6.5: The variation in different parts of CEs and in the Snyder promoter within the 40 isolates. Green color: -10, -35 patterns of Snyder promoter, IHF binding site and R (high variable position detected by (Siddique *et al.*, 2011)). Red color: variable nucleotides in IGRs having variation within -10,-35 and IHF binding site of CE. Yellow color: varied nucleotides in other positions of CEs. Pink color: varied nucleotides in position of two nucleotides after the -35 pattern. The number of isolates is the numbers that have a varied nucleotide out of 40. The number of isolates in blue color: same varied CEs with more than 20% of population.

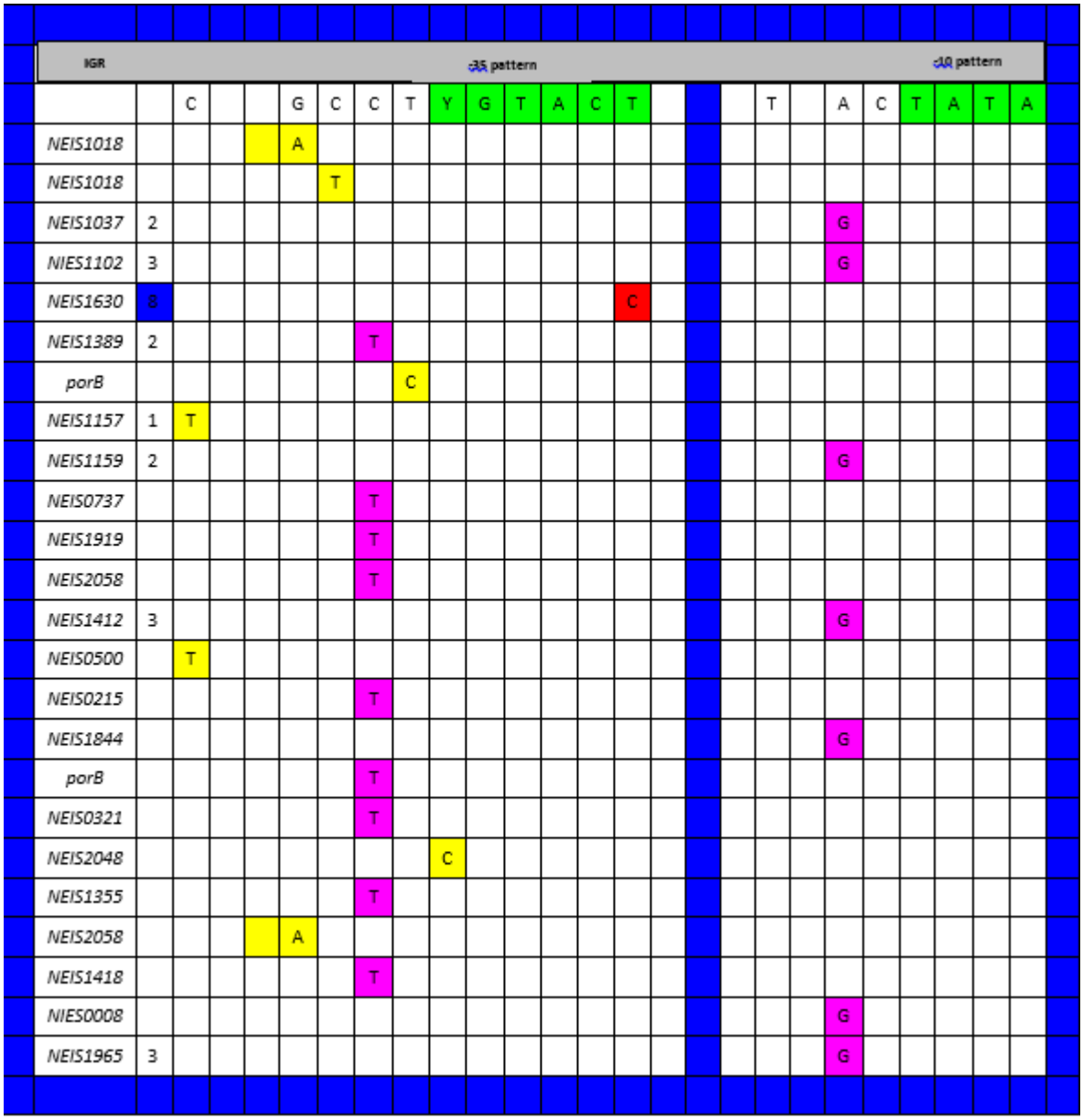


Figure 6.6: The variation in different parts of CEs and in the Black promoter within the 40 isolates. Green colour: -10, -35 patterns of Black promoter and IHF binding site. Red colour: variable nucleotides in IGRs having variation within -10,-35 and IHF binding site of CE. Yellow colour: varied nucleotides in other positions of CEs. Pink colour: varied nucleotides in position of two nucleotides before the -35 and -10 patterns. The number of isolates is the numbers that have a varied nucleotide out of 40. The number of isolates in blue colour: same varied CEs with more than 20% of population.

Table 6.1: List of genes carried IGRs with variation located in promoters and IHF of CE and genes carried variable CEs within their sequence in the 40 isolates from one carrier. The variation in these CEs may have an important role in persistence of *N. meningitidis* for months.

Snyder promoter	Gene(NMBsuffix)	Function	Number of variable isolates	Distance of CE from gene (bp)
NEIS0474	*NMB0535*	galactose transporter	16	480
NEIS1629	*NMB1711*	GntR family transcriptional regulator (transcriptional repressor for pyruvate dehydrogenase)	8	120
NEIS2006	*NMB2027*	gluconate permease (transpoter)	8	212
IHF site	**Gene(NMBsuffix)**	**Function**		
NEIS0237	*NMB0242*	NADH dehdrogenase subunit B (metabolism)	5	169
NEIS0008	*NMB0031*	glucosamine- fructose-6-phosphate B (metabolism)	3	135
Black promoter	**Gene(NMBsuffix)**	**Function**		
NEIS1630	*NMB1712*	L-lactate permease-like protein (transpoter)	8	343
Genic region	**Gene(NMBsuffix)**	**Function**		
NEIS1702	*NMB0451*	hypothetical protein	2	
NEIS0500	*NMB0558*	hypothetical protein	12	

6.3 Analysing the dynamics of variation in CE of IGRs of the pairs isolates from 25 carriers representing 2/3 to 5/6 months carriages

The CEs were identified in the IGRs of each pair of 25 isolates. There were 174 variable CEs in the combined analysis of the 25 pair isolates. However, some of this variation was due to poor assembly of the sequences of isolates. Again, no identity between old and new assembly in the position of variation in a BLAST search of variable CE patterns was used to detect the spurious variation (See section 2.3). There were 37 loci out of 174 (21.2%) that carried CEs with poor assembly.

The types of variable CEs were analysed for each CC for two different periods of carriage (2/3 and 5/6 months). In the first period, two to three months, variation was found mainly complete with alpha-alpha type for all CCs. Variation in partial CEs with alpha-alpha' type

was found in CC-167 and CC-60 while no variable CEs was observed in CC-1157-32-269. In the second time period, five to six months, the variable CEs were also mainly complete with alpha-alpha type in all CC. The partial CEs with alpha-alpha' type was found in CC-174 and CC-60 (Figure 6.7).

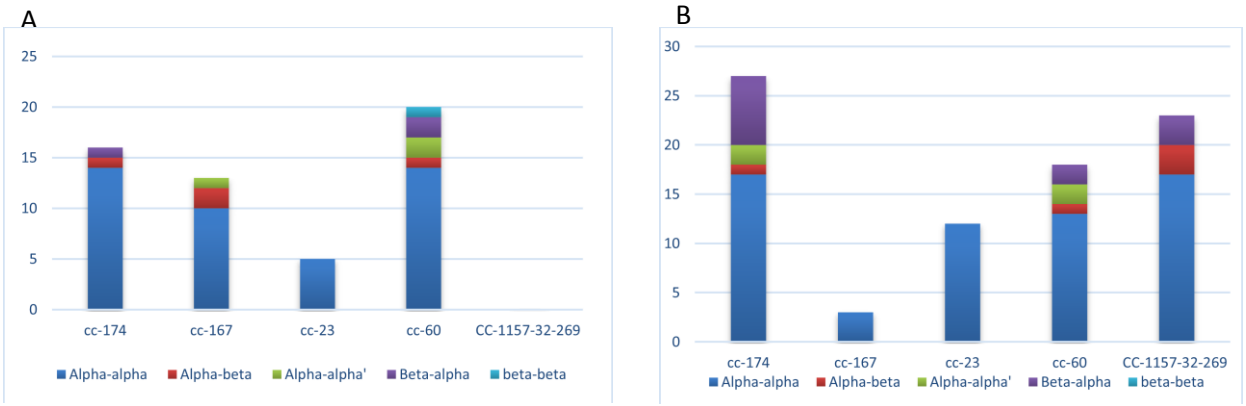


Figure 6.7: Variation in CEs for three versus six months carriage. Panel A: variation within different CCs for three months carriage, **panel B**: variation within different CCs for six months carriages.

After normalization between numbers of isolates in two different time periods, the number of variable CEs was 20.25 and 27.1 for two different periods of carriage (2/3 and 5/6 months). There was significant different with (P value 0.008) for the compared carriage 2/3 against 5/6 months of carriage using Poisson mean test.

The same variable CEs found with multiple isolates within the 25 pair isolates may introduce more advantage for the persistence of *N. meningitidis* than other CEs. 10 % or more of 25 pair isolates affected by the same variable CEs recorded in 12 out of 137 CEs. However, location of variable nucleotides within the conserved functional patterns of CE (promoters and IHF binding site) may give more evidence on contributing CEs in controlling the gene expression. Variation was found within the -35 and -10 positions of the first promoter (Snyder promoter) in 3 out of 137 and within the -35 and -10 positions of the second promoter (Black promoter) in 9 out of 137. Variation was also found within the IHF

binding site in 3 out of 137. One variable CE was inserted in *NEIS0202* gene, however there was also no evidence on the movement of CEs within the 25 pair isolates in a small-time scale (for period of months). Although, more than 60% of variable CEs that have variation in the conserved functional patterns (two promoters and IHF binding site) located upstream of genes and encode for hypothetical proteins or located in tail to tail between adjacent genes (Table 6.2). However, the variation within these CEs may enroll in physiological activities of *N. meningitidis* mostly with metabolism schemes. Moreover, variation was also detected in one nucleotide before or after conserved functional patterns in ten variable CEs. Furthermore, a high frequency of variation was seen in different positions. These positions are two nucleotides before the -10 pattern of Black promoter changing A to G, six and seven nucleotides before the -35 pattern of Black promoter changing C to T and A to G, twelve and twenty six nucleotides before the -35 pattern of Black promoter changing G to T and T to C. All these positions may indicate that changing of variable CEs correlated with changing gene expression levels (Figures 6.8, 6.9). Again, practical work has to be achieved to confirm these results.

The overall dynamic of variation in CEs of 25 pair isolates showed that variation was caused mostly by one SNP 106 out of 137 (77.3%) with rare occurrence of other types of variation. Finally, it has been shown that there were only 2 variations within NIME and one REP2 in the intergenic belong into *NEIS0510* genic region in overall 25 pair isolates.

Table 6.2: List of genes carried IGRs with variation located in promoters and IHF of CE and genes carried variable CEs within their sequence in the 25 pair isolates.

Snyder promoter	**Gene(NMBsuffix)**	**Function**	**Number of variable isolates**	**Distance of CE from gene (bp)**
NEIS0565	*NMB0621*	hypothetical protein	2	55
NEIS1281	*NMB1345*	hypothetical protein	3	Tail to tail
CE-1155	*NMB0418*	glucokinase	1	Tail to tail
IHF site	**Gene(NMBsuffix)**	**Function**	**Number of variable isolates**	**Distance of CE from gene (bp)**
NEIS1697	*NMB0454*	hypothetical protein	3	Tail to tail
NEIS0407	*NMB1814*	dehydroquinate synthase (metabolism)	4	299
NEIS0269	*NMB0274*	ATP dependent DNA helicase	1	Tail to tail
Black promoter	**Gene(NMBsuffix)**	**Function**	**Number of variable isolates**	**Distance of CE from gene (bp)**
NEIS1475	*NMB1558*	diacylglycerol kinase (metabolism)	1	106
NEIS2043	*NMB2061*	Phosphoenol pyruvate carboxylase	1	Tail to tail
NEIS1530	*NMB1609*	O-succinylhomoserine sulfhydrolase (metabolism)	1	59
NEIS1364	*NMB0312*	ATPase (transporter)	1	570
NEIS1945	*NMB1971*	hypothetical protein	2	196
NEIS0203	*NMB0211*	L-serine dehydratase (metabolism)	1	47
NEIS0612	*NMB0663*	Outer memebrane protein	1	1
NEIS1438	*NMB1451*	DNA polymerase III subunit epsilon	1	1
NEIS1018	*NMB1057*	Gamma-glutamyl transpeptidase (metabolism)	1	1
Genic	**Gene(NMBsuffix)**	**Function**		
NEIS0202	*NMB0210*	Pseudogene	1	

Tail to tail: means the IGR located between tail to tail of adjacent two genes.

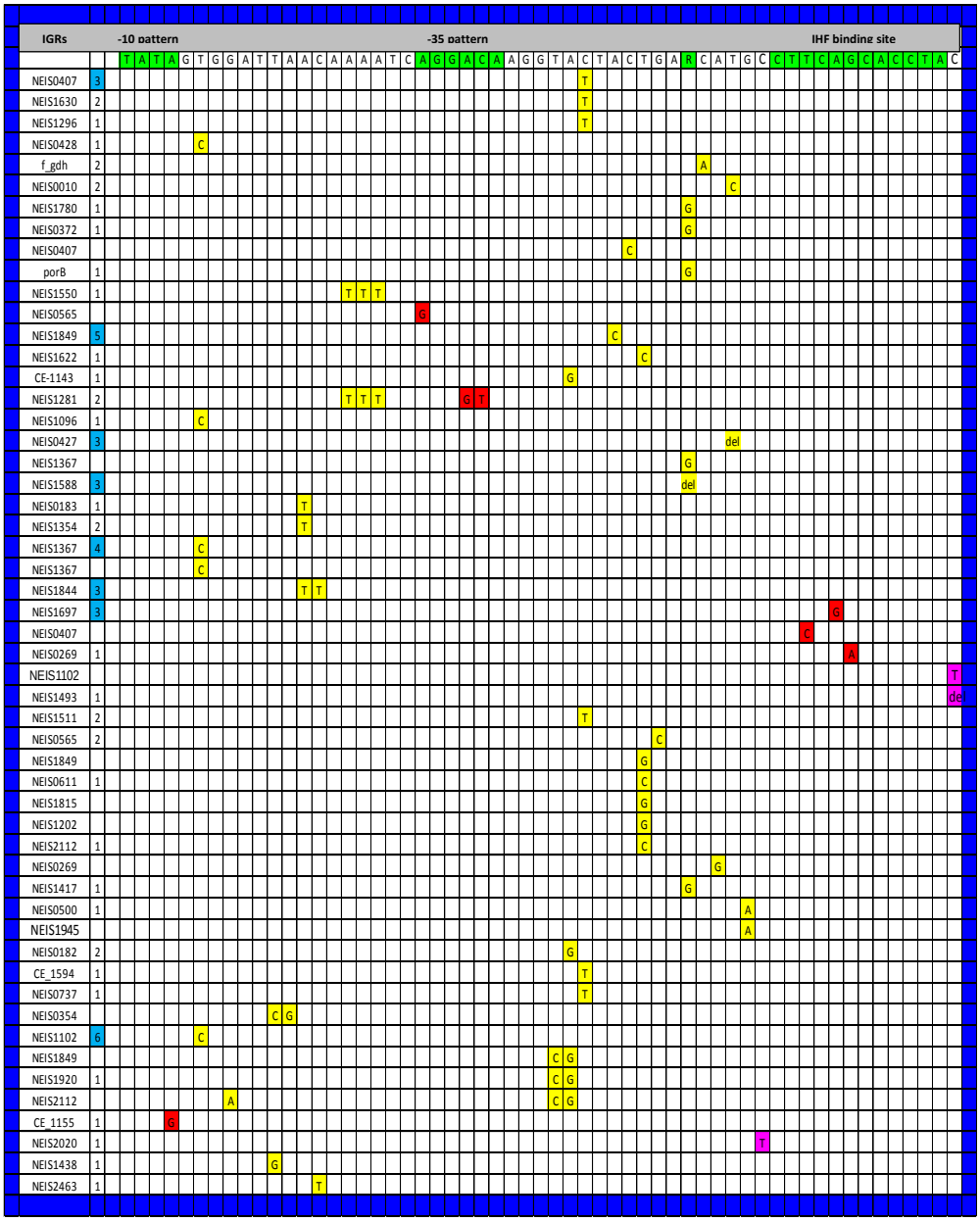


Figure 6.8: The variation in different parts of CEs and Snyder promoter within the 25 pair isolates. Green color: -10, -35 patterns Snyder promoter, IHF binding site and and R (high variable position detected by (Siddique *et al.*, 2011)). Red color: variable nucleotides in IGRs having variation within -10,-35 and IHF of CE. Yellow color: varied nucleotides in other positions of CEs. Pink color: varied nucleotides in position of one nucleotide before and after the IHF binding site. The number of isolates is the numbers that have a varied nucleotide out of 25 pair. The number of isolates in blue color: same varied CEs with more than 10% of population.

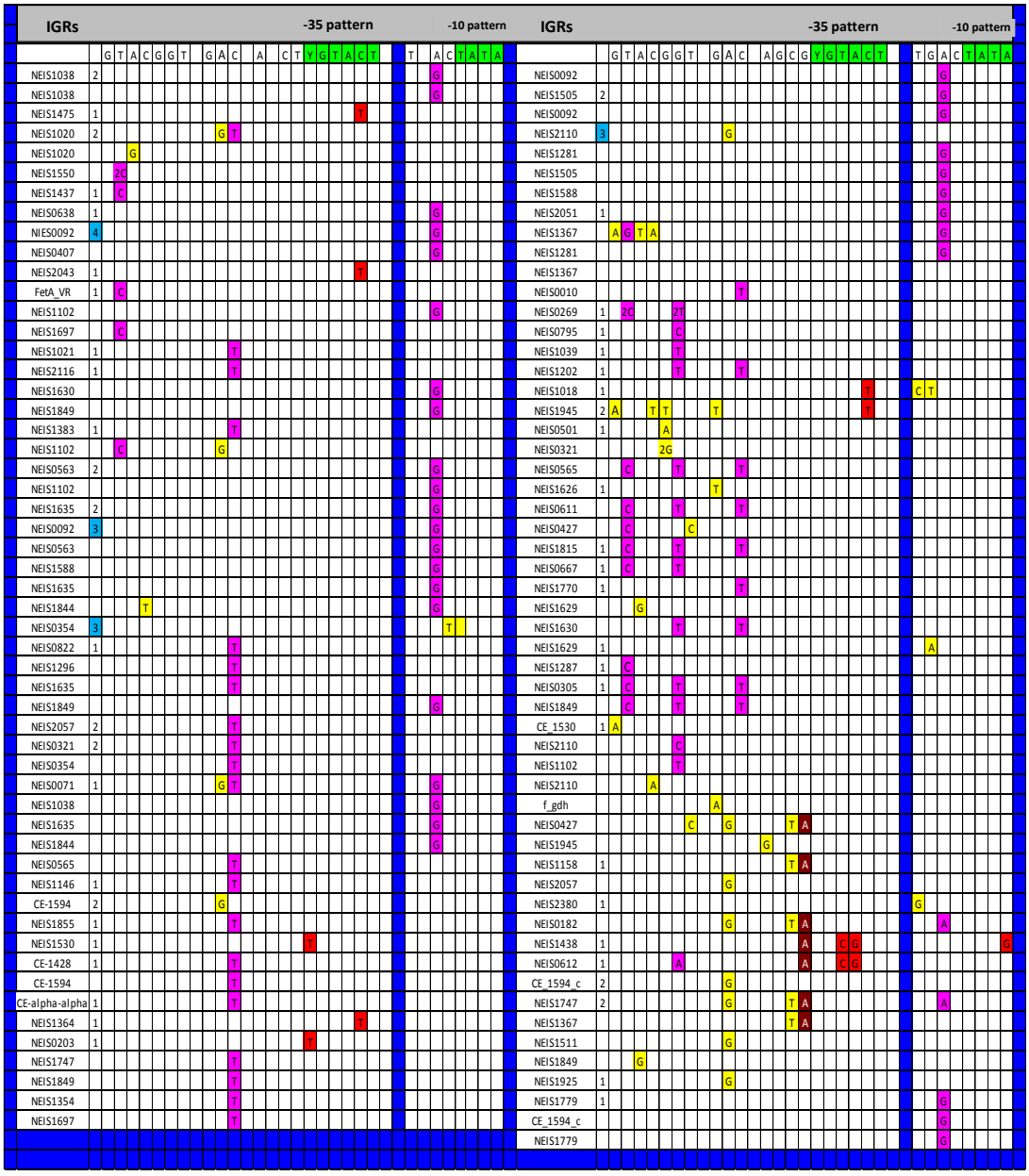


Figure 6.9: The variation in different parts of CEs and Black promoter within the 25 pair isolates. Green colour: -10, -35 patterns of Black promoter and IHF binding site. Red colour: variable nucleotides in IGRs having variation within -10,-35 and IHF of CE. Yellow colour: varied nucleotides in other positions of CEs. Pink colour: varied nucleotides in positions of two nucleotides before the -10 pattern, six and seven nucleotides before the -35 and twelve and twenty six nucleotides before the -35 pattern. The number of isolates is the numbers that have a varied nucleotide out of 25 pair. The number of isolates in blue colour: same varied CEs in 10% of population or more.

6.4 Detection the missing CEs in disease and carriage isolates of CC-174

The third type of data set in our study was 25 carriage and disease isolates (18 carriage and 7 disease isolates) from CC-174. Detection of the absence of CEs in a particular position from whole genome alignment of these isolates indicates the movement of CEs due the action of recombination or the action of TA and CA site. These sites are located in the flanks of the CE pattern and serve in insertion CE within different parts of the genome in the same way transposon like elements (Buisine *et al.,* 2002). Recombination was observed within 25 carriage and disease isolates of CC-174 (See chapter five) therefore the aim of this section was to detect the missing CEs within these isolates due the action of recombination or function of TA, CA pattern within the CE.

The CEs were missing in many isolates particularly if their location was at the end of the contig. However, whole genome alignment of 25 carriage and disease isolates of CC-174 may indicate the absence of CEs at three positions due the action of recombination or function of TA and CA site. In intergenic region of *NEIS2017*, a CE was missing in one carriage isolate, in intergenic region of *NEIS2110*, a CE was missing in a disease isolate and in intergenic region of *NEIS2043*, a CE was missing in two carriages. The missing CEs in *NEIS2110*, *NEIS2017* and *NEIS2043* were visualized by Mauve alignment (Figure 6.10). The missing CE in one isolate compared with others in disease and carriage may suggest the CE had moved from this position into another position.

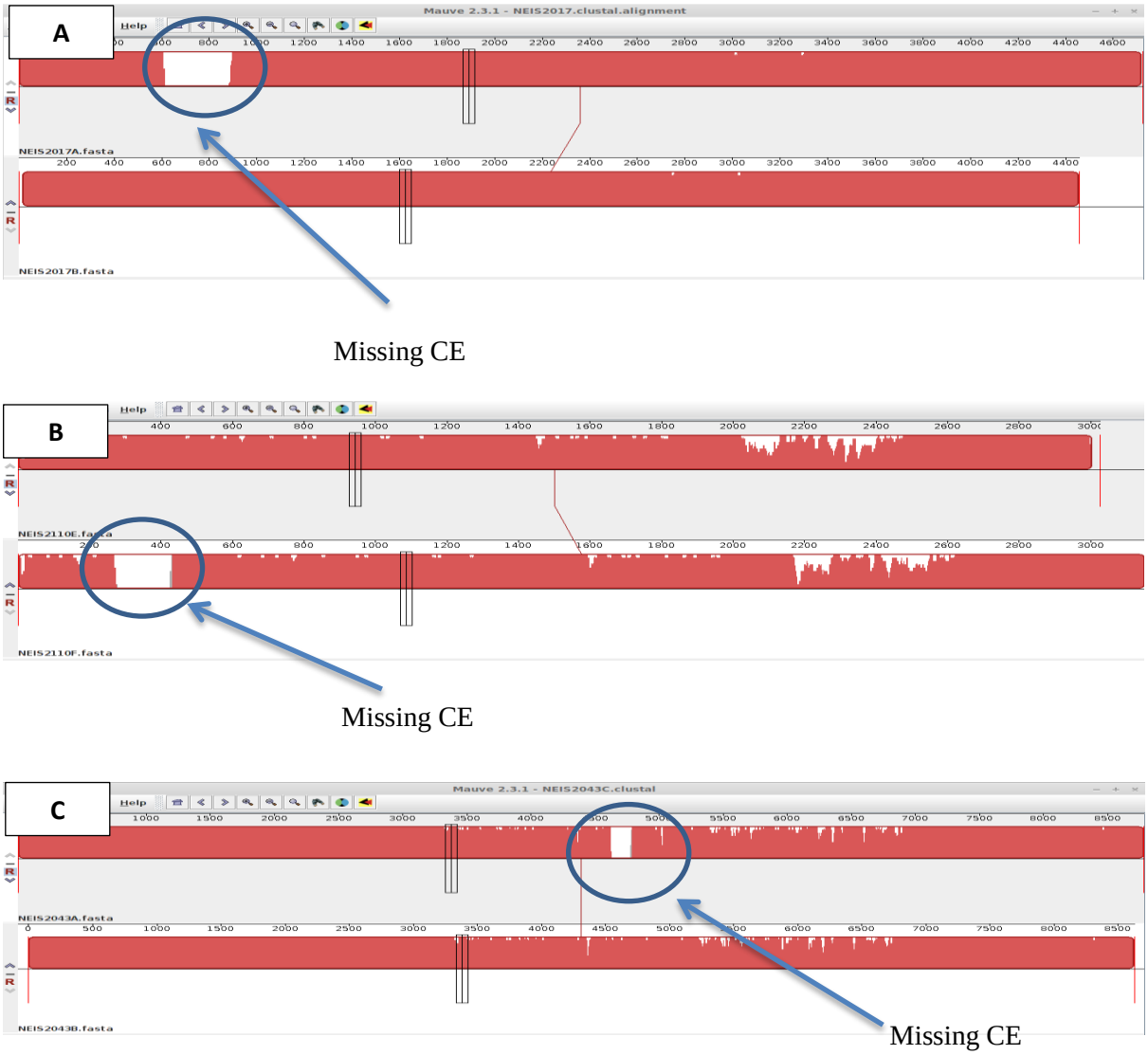


Figure 6.10: Examination of missing CE using Mauve alignment in 25 CC-174 isolates. **Panel A**: Intergenic region of *NEIS2017* in isolate ID 28250. **Panel B**: Intergenic region of *NEIS2110* in isolate ID 20072. **Panel C**: Intergenic region of *NEIS2043* in isolate ID 27530 and 27580.

6.5 Summary of main findings

The analysis of variation in the CEs showed that in the 40 isolates of carrier V59, was highest in the first, second and third time points in partial CEs while in the fourth time point the variation was mainly in complete CEs. In the 25 pair isolates, variation in complete CEs was higher in all the time points. In the 40 isolates, the CEs in the fourth time point may be more functional than other CEs in other time points. This is because the complete CEs have higher possibility of effecting gene expression than partial because they contain both promoters and IHF while the partial CEs contain eiher one promter or two promoters. In 40 isolates and 25 pair isolates, there were an increasing number of variable CEs with increasing persistence of *N. meningitidis,* however, there was no statistically significant difference. This may indicate that the location of variation within the conserved functional pattern of CEs is more important than the number of variable CEs.

Snyder *et al.* (2009) showed that the CEs could move from one position into another within the same genome. There was no evidence of movement of CEs within 40 isolates and 25 paired isolates over small time scale within their host as the missing CEs were located at the end of the contigs or the gene were missing due partial sequencing of these strains by Illumina Hiseq. There were three missing CEs in disease and carriage isolates of CC-174 that were in *NEIS2110*, *NEIS2017* and *NEIS2043*. This may indicate there was a movement in CEs in these diseases and carriage isolates, nevertheless, these differences need to be confirmed by PCR. This result gives evidence that CEs need more evolutionary time than the duration of persistence of *N. meningitidis* with their host in order to alter genome structure or variation.

Important points were concluded from this analysis related with the presence of the same variable CEs within multiple isolates with different time points being under the selection and the location of SNPs within the conserved functional pattern of CEs. In the 40 isolates, 9 out of 39 variable CEs conducted with multiple isolates, each one affected around 20% or more of population. In the same way, the presence same variable CEs within multiple isolates belonging to same CCs or different CCs of 25 pair isolates detected in 12 out of 137 variable CEs, each one affected around 10% or more out of 25 pair isolates. Interestingly, some CEs showed variation within promoters or IHF patterns or were inserted

inside the genes, and were associated with genes that coded mainly for different enzymes within metabolism scheme or transporters within outer membrane scheme in both 40 and 25 pair isolates. This variation may change the level of transcription of genes belonging into previous mentioned schemes and this altered expression may help *N. meningitidis* to resist the immune system. There were 6 out of 39 and 15 out of 137 variable CEs that showed variation within the conserved functional pattern in 40 isolates and 25 pair isolates respectively. Furthermore, some CEs showed variation located within the nucleotides nearby promoters or IHF patterns and other showed variation with high frequency in particular positions within the CE pattern. The change in these positions may also enhance changing the level of transcription of genes that carried these variable CEs within their IGRs.

Some studies have shown that CEs have a strong promoter that affects the level of transcription of genes especially when the variation located at (Y128 that is the first nucleotide within Black promoter) (Siddique *et al.*, 2011). However, another study showed that the promoters regions, IHF of CEs located upstream of *pilH-X*, and *pilGD* gene, did not influence the level of transcription. In *pilF* gene, the activation of Q70 promoter was enhanced by insertion of CEs. Nevertheless, it was found that Q70 promoter was also functional in the absence of CE insertion (Lin *et al.*, 2011). Finally, it was suggested that CEs have a role in controlling expression of the mtrCDE operons (Rouquette-Loughlin *et al.*, 2004), however, a study was carried out by Enrquez *et al.*, (2010) showed that the deletion of CE has no any effect on susceptibility of *N. meningitidis* to antimicrobial drugs.

Another function of CEs is as a source of variation in *N. meningitidis*. A study showed that CEs might correlate with deletion of *fetA* in some isolates. The deletion was mediated by a repeat array that was flanking this gene. This suggests that recombination mediating CEs has a role in generating variation within populations of *N. meningitidis* (Claus *et al.*, 2007). It has also been reported that the CE moved in the same strain due to recombination and are major source of allelic variation in the population (Siddique *et al.*, 2011). Our study has provided some evidence that CEs are source of variation in persistent meningococcal infections within their host. However, the analysis of variation within the CEs carried out in a silico study therefore the genome variation needs to be confirmed by practical work.

6.6 Practical work

6.6.1 Introduction

Some experiments were designed to investigate alteration in SSR in five phase variable genes during a persistent carriage of three meningococcal strains in three carriers (a CC22, in V54, N428), (a CC198, in V64, N436) and (a CC269, in V124, N419). Fifteen isolates from 2008-2009 were investigated in this study. The data for the first, second and third time points were generated by (Alamro *et al.*, 2014). The aim of this study was to visualize how the SSR of phase variable genes in the fourth time point altered as these carriers were colonized by different strains at this time point.

Another aspect was to validate the variation that had been identified in the genomes of 25 pair isolates using the SAP method and Sanger sequencing.

6.6.2 GeneScan

Variation within the repeat tracts of meningococcal surface proteins is crucial for phase variation. The length of each repeat tract was estimated by the GeneScan technique. Specific primers labelled with fluorescent dyes were used to amplify particular SSRs for five phase variable genes: - *opc*, *hpuA*, *nalP*, *hmbr* and *nadA*. The amplification of the SSRs with four genes was successful while the *nadA* was not found in the samples (N419, N436 and N428) (Figure 6.11).

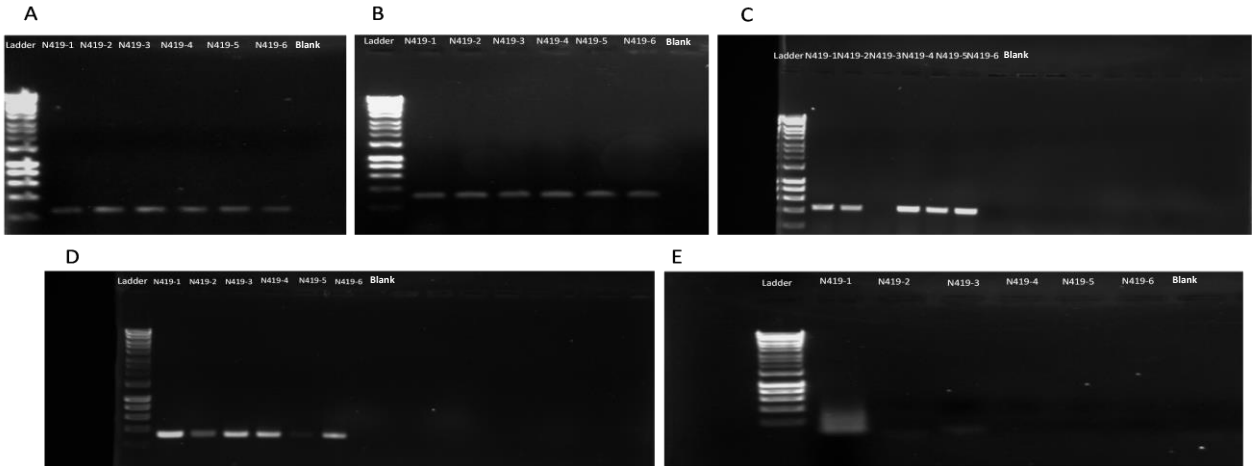


Figure 6.11: Amplification of SSRs for five phase variable genes in six isolates. Panel A: *opc*, **panel B**: *hpuA* gene, **panel C**: *hmbr* gene, **panel D**: *nalP* gene, **panel E**: *nadA* gene.

The GeneScan and Peak Scanner Software were used to estimate the size of fluorescent PCR products. In general, there was one major and one minor peak. The peak with the highest signal was considered the main peak when the ratio between primary and secondary peaks was more than 1.2. On the other hands, if the ratio was between primary and secondary peaks were less than 1.2, the GeneScan technique for particular samples was repeated (Figure 6.12).

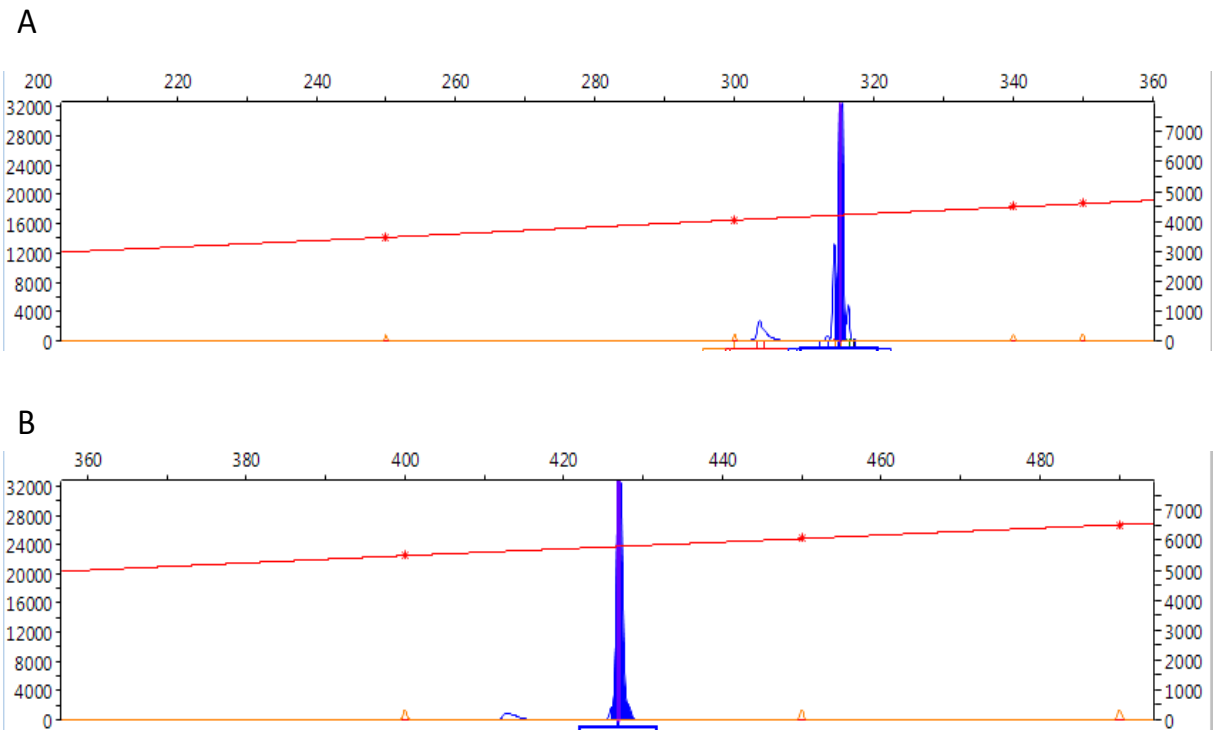


Figure 6.12: GeneScan analysis for two phase variable genes form. Panel A: GeneScan analysis of *nalP* with peak size 315 in blue colour, which correlates to nine repeats, **panel B**: GeneScan analysis of *hmbr* with peak size 426 in blue colour, which correlates to eight repeats.

To be more precise, sequencing of a sub-set of repeat tracts was used to confirm that changes in the PCR product size was due to changes in the length of repeat tract among phase variable genes (Figure 6.13).

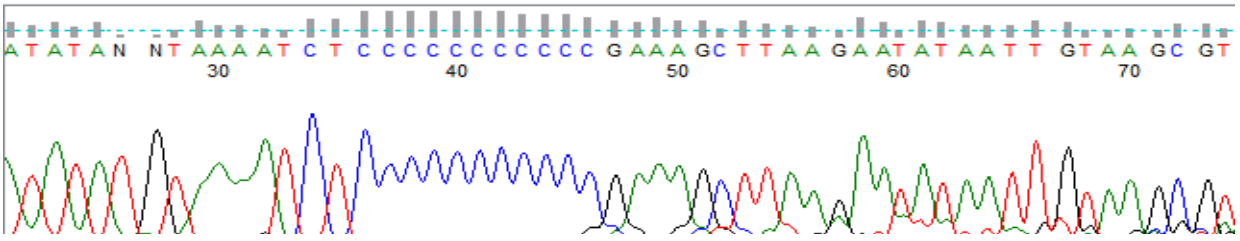


Figure 6.13: Example of sequence of the repeat tract of *hpuA* gene from isolate N419.3. The repeat sizes obtained from GeneScan were confirmed by sequencing.

6.6.3 Distribution of tract lengths during carriage

In general, Alamro *et al.* (2014) determined the repeat tract length for four genes under the study in up to 6 colonies of carriers V64 (for strains of type N64 in first time point, N257 in second time point, N348 in third point), V124 (for strains of type N124 in first time point, N290 in second time point, N336 in third point) and V54 (for strains of type N54 in first time point, N237 in second time point, N343 in third point).

In the current study, the repeat tracts length were determined for four genes in up to 6 colonies from the fourth time point of same carriers V64 (for strain of type N436) and V124 (for strain of type N419) and up to 3 colonies from the fourth time point of volunteer V54 (for strain of type N428). The strains in the fourth time point were from different CCs than those in the earlier time points.

The results showed that high frequencies of phase variation occurred during the fourth time point in all four genes (Appendix 38).

The repeat tract of *opc* is located within the core promoter and changes in the repeat lead to changes in the transcription level of the gene. Sarkari *et al.* (1994) showed that the *opc* gene varied in its expression depending on the length of repeat tract with high expression of *opc* is correlated with 12Cs, 13Cs, 11Cs, and 14Cs while the low or intermediate expression of *opc* is correlated with less or equal to 10Cs or larger and equal to 15Cs. In carrier V54, in the first time point, the *opc* repeat numbers were 11Cs with high level of expression then in the second time point; *opc* repeat numbers were 10Cs with four colonies with low or intermediate expression and two colonies with 11Cs with high expression. In the third time point, *opc* repeat numbers were 11Cs with three colonies (high expression), 10Cs with two colonies (low or intermediate expression) and 12Cs with one colony (high expression) (Alamro *et al.*, 2014). In this study, in carrier V54, in the fourth time point, the *opc* repeat numbers were 10Cs with two colonies (low or intermediate expression) and 14Cs with one colony (high expression). In carrier V64, in the first time point, the *opc* repeat numbers were 13Cs, 14Cs with high expression. In the second time point, the *opc* repeat numbers were 1 colony with 15Cs and correlated with low or intermediate expression and 5 colonies with 14Cs and correlated with high expression. In the third time point, the *opc* repeat numbers were 2 colonies with 15Cs and correlated with low or intermediate expression and

4 colonies with 14Cs and 12Cs correlated with high expression (Alamro *et al.*, 2014). In this study, in the fourth time point, in carrier V64, the *opc* repeat numbers were 9Cs and 10Cs with two colonies with low or intermediate expression and 4 colonies with 13Cs and 14Cs with high expression. In carrier V124, in the first, the *opc* repeat numbers were 3 colonies correlated with low or intermediate expression with 15Cs and 3 colonies correlated with high expression with 13Cs, 14Cs. In the second, the *opc* repeat numbers were 13Cs, 14Cs with five colonies (high expression) and 15Cs with one colony (low or intermediate expression). In the third time point, the *opc* repeat numbers were also 15Cs with one colony (low or intermediate expression) and 14Cs with five colonies (high expression) (Alamro *et al.*, 2014). In this study, in the fourth time point, in carrier V124, the *opc* repeat numbers were less than 10Cs therefore all with intermediate or low expression.

It has been reported that the poly G in the *hpuA* gene is located within the gene therefore the change in the repeat tract leads into frameshift mutations. Tauseef *et al.* (2011) showed that the repeat numbers 7Gs, 10Gs, 13Gs, 16Gs, and 19Gs were associated with ON expression state while other repeat numbers were OFF expression state. In carrier V54, in the first time point, most colonies had a repeat number of 12Gs therefore; *hpuA* was in the OFF state. In the second time point, the *hpuA* repeat numbers were 5 colonies with 13Gs (ON state) and one colony with 11Gs with (OFF state) while in the third time point, the *hpuA* repeat numbers were 13Gs with one colony (ON state) and 11Gs or 12Gs with (OFF state) (Alamro *et al.*, 2014). In this study, in the fourth time point, in carrier V54, the *hpuA* repeat numbers were 10Gs with one colony (ON state) and two colonies with less than 10Gs (5Gs, 6Gs) (OFF state). In carrier V64 and V124, in the first, second and third time point, the *hpuA* repeat numbers were 10Gs therefore all the phase variable genes with ON state (Alamro *et al.*, 2014). In this study, in the fourth time point, in carrier V64 and V124, the *hpuA* repeat numbers were 5Gs, 7Gs, 11Gs, 12Gs and 15Gs so mostly in the OFF state.

The *nalP* repeat tract is also located within the open reading frame. In carrier V54, in the first and second time point, the *nalP* repeat numbers were 10Cs therefore all the phase variable genes were in the ON state. In the third time point, the *nalP* repeat numbers were between 11Cs and 12Cs and in an OFF state (Alamro *et al.*, 2014). In this study, in the fourth time point, in carrier V54, the *nalP* gene was 9Cs and also in the OFF state. In

carrier V64 and V124, in the first, second and third time points, and the *nalP* repeat numbers were 10Cs therefore, all the phase variable genes had an ON state (Alamro *et al.*, 2014). In this study, in the fourth time point, in carrier V64 and V124, the *nalP* repeat numbers were within a range from 8Cs to 10Cs. There were five colonies with 10Cs with ON state while there was one colony with 8Cs and an OFF state. For strain of types N419 the *nalP* repeat numbers were 8Cs and 9Cs hence in the OFF state.

The *hmbr* gene has a repeat tract in the coding sequence. In this study, in carrier V54, in the fourth time point, the *hmbr* gene had 9Gs or 12Gs with an ON state. In carrier V64, the *hmbr* gene was also between 9Gs and 12Gs with ON state. While for V124, there were two colonies with 8Gs with OFF state and three colonies with 9Gs and one colony with 12Gs with ON state and hence 4 colonies within ON state (Figure 6.14).

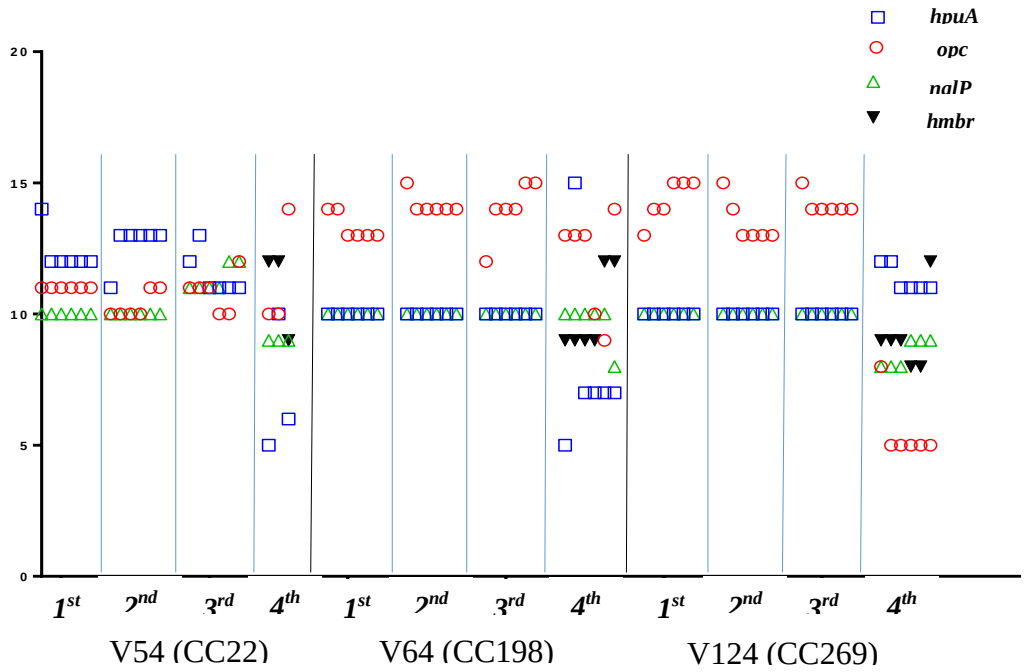


Figure 6.14: SSRs of four phase variable genes in fourth time of three carriers. The *hpuA* gene, blue square, *opc*, red circle, *nalP*, green triangle and *hmbr*, black triangle. 1st, 2nd and 3rd indicate first, second and third time point, these data were taken from (Alamro *et al.*, 2014). The data in this study represented in the fourth time points (4th).

6.6.4 Validation of some variable genes belongs into 25 pair isolates

The validation of variation in the 25 pair isolates was carried out using Simple Allele discriminating PCR (SAP method). The SAP method is applied for a small scale genotyping and gene mapping. This method was used to distinguish between initial and variant alleles. The principle of the method is correlated with designing forward initial type; forward variant type and common reverse primers (See 2.10.8). The generation of PCR products for each primer pair was used as an indication for the presence of initial and variant alleles. Two genes were tested in carrier 96 in up to six colonies in each isolate within the pair isolate. Initial allele of *NMB0329* and *NMB1930* had samples (259.1 to 259.6) with G and C nucleotides while variant allele had samples (445.1 to 445.6) with T nucleotide in both genes.

The result was compatible with initial sequencing of two genes by Illumina Hiseq and showed that the *NMB0329* and *NMB1930* genes had variant (T) in all the samples (445.1 to 445.6) instead of initial (G) and (C). The sequencing of both genes was as a further indication of the reliability of variation in both genes (Figure 6.15).

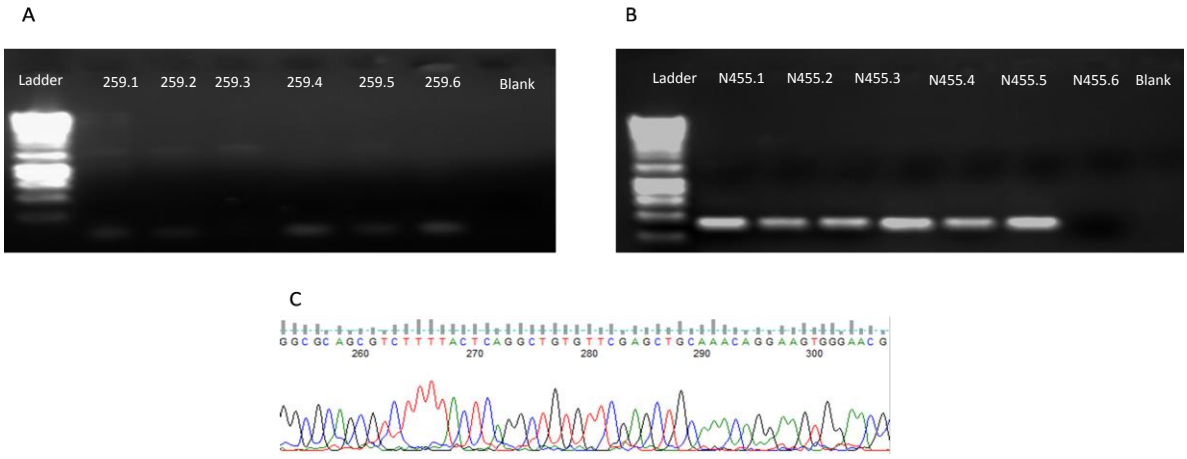


Figure 6.15: Example of a validation of a SNP using Simple allele-discriminating PCR (SAP method). Panel A: primer designed to amplify G nucleotide in gene *NMB0329* of sample N259.1 to N259.6, showed no amplification. **Panel B**: primer designed to amplify T nucleotide in gene *NMB0329* of sample N445.1 to N445.6, showed amplification. **Panel C**: sequencing confirmed nucleotide T at the position 270 in *NMB0329* of sample N445.1.

6.6.5 – Summary of main findings

Alteration of repeat tract length within the SSR of phase variable genes may enhance the persistence of isolates within their host for a long time (a period of months) (Alamro *et al.*, 2014). The switching ON/OFF mechanism mediating phase variable genes has been shown as a driving force for the concept of within host evolution (WHE) through modulating the effect of virulence factors being encouraging fitness rather than leading into state of death (Klughammer *et al.*, 2017). Alamro *et al.* (2014) showed there was trend towards selection for OFF state or low expression for three phase variable genes (*opc, hmbr, nadA, nalP, hpuAB*) in three carriers (V54, V124 and V64) within strains belonging into CC174 and CC167 in first, second and third time points. He concluded that the selection for low or OFF state helped *N. meningitidis* to persist for a long time (Alamro *et al.*, 2014). The current study aimed to detect the alteration in the repeat tracts of the same five variable genes within the previous three carriers (V54, V124, and V64) but for other strains belonging to CC22, CC269, and CC198 at the fourth time point. There was also a trend towards selection for an OFF state or low expression for three genes which are (*opc, hpuAb, nalP*) with 2/3 (66.6%) carriers while *hmbr* gene showed ON state in all carriers (100%). This indicates that antibodies formed against (*opc, HpuAb, nalP*) genes in strains belonging to CC174 and CC167 in first, second and third time points were able to enhance immunization against the isolates in the current study for CC269 and CC22 resulting in the selection for the OFF state or low expression. Conversely, antibodies formed against *hmbr* gene in the previous time points on isolates within the CC167 may not show immunization against *hmbr* gene in the isolates of current study CC22, CC198 and CC269 and so the gene stayed in the ON state. Interestingly, the *nadA* gene was missing from isolates under the current study. These results may reflect important aspects in the vaccination program especially as *nadA* is one of the component of Bexsero vaccine (Vogel *et al.*, 2013).

Finally, validation with SAP method and sequencing confirm the variation within the tested variable genes as previously identified by NGS.

Chapter 7: Conclusion and future work

N. meningitidis is a major causative agent of meningitis and septicaemia throughout the world. In the UK, a conjugate MenACWY vaccine has been recently introduced to combat infectivity of a hypervirulent MenW strain. Similarly, the Bexsero vaccine has been recently introduced into the infant immunization program to counteract MenB strains. However, it has been suggested that Bexsero may not eradicate MenB disease due to the evolution of vaccine escape mutants. The current study aimed to investigate the role of recombination, de novo mutation, and localised hypermutation in the evolution of carriage isolates during long-term persistence within their host and to estimate the effect of recombination on population structure of a single CC. This investigation will aid development of vaccines and immunization programs and so help in reductions in disease caused by pathogenic meningococcal strains.

The evolution of one strain of *N. meningitidis* was examined by analysing the genomes of 40 isolates from one carrier (representing 1-6 months host persistence). This study revealed that mutation was the only force driving genomic variation and in formation of the population structure. This genome variation was produced by introduction of SNPs while no indel events were observed (except within phase variation repeat tracts). Comparing the mean nucleotide diversity of isolates in first and third time points against isolates in first and fourth time points generated a P value of 0.03 for the genic regions. While the same comparison did not show statistically significant for the IGRs. This observation was confirmed by constructing a haplotype network for the genic regions versus IGRs. The number of haplotype found in the IGR network was 26, which was higher than the 15 haplotypes in the genic network. However, all the isolates for the different time points in the IGR haplotype network were dispersed randomly while genic haplotypes showed some clustering with the fourth time points having a nucleotide diversity higher than with first, second and third time points. This observation indicates a major role for mutation in forming different population structures during evolution of *N. meningitidis* within their hosts. An important observation was seen from the haplotype construction was that temporal haplotypes and persistence of identical or near identical isolates in all four-time points were detected.

Contribution of variable genes in the persistence of *N. meningitidis* assumed to be controlled by different factors; these are potential functional effect of the variability in different genes, the variation located within functional important residues and the transient nature of the SNPs.

According to KEGG classification, variation was detected in various functional schemes, however, the genes having a major role in shaping the variation and their SNPs located within important regions of the protein were present in metabolic schemes. These genes code for D-alanine-D-alanine ligase and iron-sulfur cluster insertion. It has been reported that metabolic adaptation is a key feature of *N. meningitidis* in counteracting the effect of nutrient starvation (Schoen *et al*., 2014). In addition there was a SNP in MutS, a DNA mismatch repair protein and it has been reported that epidemic meningococcal strains of serogroup A are associated with the presence of mutations in MutS (Richardson *et al.*, 2002). However, the haplotypes may be temporary and so have no fitness advantage.

The transient nature of the SNPs which means expansion of a variable gene within different time points was detected in genes that code for ATP-binding subunit ClpA and capsule region A of serogroup Y. The change of the latter gene may alter the structure or amount of capsule and facilitate escape the immune system. While contractions nature of the SNPs, which means appearing and disappearing allelic variants within the different time points, detected in the rest of variable genes. However, in the fourth time point, allelic variants were dominated by a new lineage. This critical observation means that the persistent meningococcal carriage population was not in a steady state, but highly dynamic by which allelic variants fluctuated within the different population with no fixation.

The evolution of *N. meningitidis* during different periods of host persistence was further investigated by analyzing genome variation within 25 pairs of isolates belonging to different CCs. Although the genic variation (mutable genes/genome) in the current study was very low as compared with other studies (Bårnes *et al.*, 2017) this was presumably due the action of stabilizing selection. Two pairs of isolates from V54, colonised by isolates (16978-17004) from CC-174, and V128, colonised by isolates (16986-17012) from CC-167, scored the highest for the genic variation (mutable genes/genome) with 0.010 and 0.017 respectively. A high genic variation is a signal of recombination. Therefore, the SNP

density approach was used to infer recombination and predicted recombination events in 26 loci out of 35 (74.2%) of the V128 isolate pair. In the same manner, a high level of nucleotide diversity and evidence of recombination was found in the isolates of two volunteers colonised with CC-60 strains,V82 (16995-17021) and V114 (16991-17017), one colonised with a CC-23 strain, V222 (16990-17016; CC-23), and two with CC174 strains, V54 (16978-17004) and V51 (16976-17002) in the CC-174 (0.00016, 0.00014, 0.00001, 0.00001 and 0.00001). Therefore, SNPs density approach was applied to the genomes of all the 25 pair isolates and this indicated that 3.2% of the total length of genome was subject to recombination with an average fragment size of 1660 nucleotides. The recombination contributed 90.5% of the total variation in the genic regions of the 25 paired isolates. Additionally, the number of synonymous polymorphism in all the 25 pair isolates was 728, which was 1.4 times higher than the number of non-synonymous polymorphisms (496 SNPs) presumably due to the action of purifying selection to diminish the effect of deleterious mutations. However, the non-synonymous polymorphisms was 1.8 times higher than synonymous polymorphism in the variable genes with a single mutation due to either the action of diversifying selection or not enough time for removal by purifying selection.

These results indicate that recombination was the main driving force for generation of genome variation during evolution of *N. meningitidis* strains within the host. The genome variation was mainly due to SNPs while few insertion/deletion events have been observed. The nucleotide diversity for each functional scheme showed that the diversity in the environmental information processing genes at 0.027 was 2.9 and 4.6 times higher than the diversity in metabolism (0.009) and genetic information processing (0.005) respectively but with no significant difference P value 0.9. This result may indicate that the diversifying selection acts on the surface antigens and there is selection for antigenic variation in OMPs during persistence carriages. The environmental information processing involves membrane transport, ABC transporters, bacterial secretion systems, signal transduction, iron acquisition, and surface antigen genes. However, variation was mainly conducted with pilin, porin and iron acquisition genes. Variation was also detected in the capsule of serogroup Y. This was further confirmed as environmental genetic processing had the highest proportion of recombinant (6.7%) and mutant (7.2%) variable genes. Examples of recombinant genes that may contribute to the persistence of the 25 pair isolate are those

coding for pilF and pilD proteins, and iron acquisition proteins. These genes were found in the recombination blocks that have recombination signals on three or more adjacent loci. The variation in the pilin, porin, iron acquisition, and capsule genes may play an important role in colonization and persistence of *N. meningitidis* with their host and indicating the concept of within-host evolution (WHE) rather than evolution of the carriage isolates to cause disease.

Recombination and mutation are major forces that are responsible for formation of the population structures of different bacteria. Mutation can produce high levels of linkage disequilibrium and allows formation of a clonal population structure as seen for *E. coli* and *Salmonella enterica* population (Feil *et al.,* 1995; Zhou *et al.,* 1997). Conversely, recombination enhances high levels of diversification of clones such as seen for naturally tranformable *N.meningitidis* populations (Spratt *et al.,* 2001). However, the rate of recombination in serogroup A strains appears to be significantly lower than other serogroups, so that meningococcal lineages of serogroup A strains have a more clonal population structure (Bart *et al.*, 2001). Recently, it was reported that 25 disease and carriage strains of CC-174 isolates from the UK have genomes of high similarity and formed a tight cluster (Oldfield *et al.*, 2016). The assumption was that these strains have a low rate of recombination and form a clonal population structure in the same manner as observed for the serogroup A strains. This hypothesis was studied by estimating the rate of recombination and inferring whether the population structure was clonal or highly divergent using the 25 CC174 disease and carriage strains.

A series of new Perl scripts were used to construct contigs containing the real SNPs for each of these disease and carriage isolates. These contigs were examined using software for estimating the rate of recombination among these strains. Initial results showed that the ratio of recombination to mutation and the relative effect of recombination and mutation were 0.84 and 14.3 respectively. The SNPs density in recombination fragments was 0.026, which was significantly higher than the 0.006 observed for mutational SNPs. In these genomes, 10.3% of the core genome was subject to recombination. Comparison of phylogenetic trees derived using ClonalFrameML and PhyML showed the presence of changes in the branch length. Comparison of these changes with other studies from other

species that form clonal population such as *S. aureus* provides evidence of a role for recombination in generating variation (Didelot and Wilson, 2015). However, these parameters indicate a low rate of recombination within the 25 disease and carriage isolates in comparison to other studies on *N. meningitidis* explaining the high genome similarity and tight cluster (Didelot *et al.*, 2009).

The changing in the population structure clearly can be seen from the trees comparison of firstly; ClonalFrameML and PhyML and secondly; recombination and mutation tree. There was a change in clustering of some disease and carriage isolates.

The strains did not however cluster by year of isolation, thus isolate ID: 27497 was isolated in 2000 but was ancestral of other isolates that isolated between "2008-2011". This suggest that there has been a long period of clone diversification.

The key finding was that the recombination events were higher in clusters containing disease and carriage isolates as compared with the C4 cluster containing only carriage isolates due to two main reasons; - firstly, a temporally overlapping carriage and disease population was detected within clusters containing disease and carriage isolates, secondly high genome similarities may be observed due to one specific highly transmissible clone within the carriage isolates of C4 cluster as all isolates were collected from the same time and the environmental niche and were temporal related.

The dS was 13.9 times higher than dN and nucleotide diversity for synonymous polymorphisms was 1.5 fold higher than for non-synonymous polymorphism. This indicates that purifying selection is acting on a high proportion of the variable genes in this group of strains and indeed purifying selection found for 77% of recombinant genes for diminishing the effect of deleterious mutations on meningococcal genomes. However, two genes exhibited evidence of positive selection, *NMB0071* codes for a capsule polysaccharide export protein (CtrA) while *NMB1605* codes for topoisomeraseIV subunit A. The first gene may contribute to escape of the immune system by *N. meningitidis* and the latter gene may play a key role in DNA processing and antibiotics resistence that is important for the life cycle of *N. meningitidis* (Shultz *et al.*, 2005). This result introduces a

good evidence for the role of recombination in the adaptation of meningococcal lineages through evolution over mutation.

The signaling and cellular processing scheme contained the highest number of genes subject to variation by recombination and mutation within the CC174 disease and carriage strains, however there was not a significant difference between the different schemes. This indicates that the selection did not act on a particular scheme and this was compatible with a previous finding that positive selection only occurred in two genes *NMB0071* and *NMB1605*. Therefore, the recombinant genes with a hot spot for recombination may have a greater contribution to adaptation of these isolates than other recombinant genes because they were more frequently selected than other loci. These loci were mainly coding for acquisition systems for iron, sulfur, and other compounds, DNA processing enzymes or proteins that are essential for DNA stability. These genes may affect host colonization by *N. meningitidis* and have an essential role in physiological activities of *N. meningitidis.*

The analysis of genomic variation in the 40 isolates and 25 pair isolates was extended to look for variation in CE patterns. Many variable CEs were not real due to poor assembly in the sequences of isolates with 32/72 (45%) in the 40 isolates and 37/174 (21.2%) in the all 25 pair isolates. There was no significant increase in variability of CEs with time periods, however the same variable CE patterns were detected in multiple isolates with time periods presumably due to the action of selection in 9 out of 39 variable CEs in the 40 isolates. Similarly the same variable CE patterns were detected in multiple isolates detected in 12 out of 137 variable CEs in 25 pair isolates. Therefore, the same variable CEs in multiple isolates seem to be more likely to contribute to the diversity and evolution of the isolates within the host. Alternatively they may be assembled poorly more often. However, many researchers have shown that variation in CEs in the conserved functional patterns (Snyder promoter, Black promoter and IHF binding site) may play a significant role in changing the level of expression for specific gene (Siddique *et al.*, 2011). However another study showed variation in CEs in the conserved functional patterns did not influence the level of transcription (Lin *et al.*, 2011). Therefore, practical work has to be achieved to provide good evidence on the effect of variable CEs on level of expression for specific genes. The results also showed that CE serves as a powerful mechanism for driving the variation

within the IGRs of many genes and especially within the metabolism and transporter genes. This was compatible with other studies (Liu *et al.*, 2002; Snyder *et al.*, 2009). The correlation of variable CEs with these schemes may also indicate their role in adaptation of *N. meningitidis* within thier host against stress condition. Finally, the result has proven that CEs are a source of variation in *N. meningitidis* within their host but this study is a silico study and need to be confirmed by practical work.

Finally, alteration of SSR in five phase variable genes was investigated within persistent carriage isolates of *N. meningitidis* from the fourth time point (a period of 6 months) of three carriers (V54, V124, and V64). These carriers exhibit displacement of one strain by another strain at this time point. Alterations in SSR of the same five phase variable genes have been already detected for the same three carriers within the first, second and third time points but for the starting strains (Alamro *et al.*, 2014). The replacement strains were examined and three phase variable genes (*opc, hpuAb, nalP*) showed switches into an OFF state or low expression. This suggests that there was selection against expression of the antigens by the immune system. Similarly, *nadA* has been not detected within these isolates. Again this suggests that the antibodies formed against these phase variable genes in the starting strains in first, second and third time points were able to stimulate immunization against replacement strains belonging to CC-269 and CC-22. Therefore, they showed OFF state or they were not present.

In summary, the examination of evolution for 40 isolates from one host was accomplished and showed only an effect of mutation on genome variation during persistence of isolates for six months within a carrier. Variable genes included capsule region A in serogroup Y, D-alanine-D-alanine ligase, iron-sulfur cluster insertion, and DNA mismatch repair which may help *N. meningitidis* to persist for a long time. On the other hand, the evolution of the 25 paired isolates within their hosts were mainly affected by recombination but with no significant increase in genome variation with persistence of isolates for 5-6 versus 1-3 months. The variable genes encoded membrane transporters, ABC transporters, bacterial secretion systems, signal transduction, iron acquisition, and surface antigen genes. However, pili, porin genes, iron acquisition, and capsule serogroup Y may enhance persistence of *N. meningitidis* for a long time and encouraging the concept of within-host

evolution rather than evolution to cause disease. In the 25 CC174 disease and carriage strains, the rate of recombination was low as compared with other meningococcal lineages, however it was high and enhanced the diversification of clones in this population comparable with other clonal population species. In this case, the variable genes mainly encoded transporters, secretion systems, bacterial mobility proteins, two-component systems, antimicrobial resistance, defense, and toxin genes. These findings have provided new insights into within host evolution and microevolution of *N. meningitidis* strains.

Future work should focus on validation of genomic variation in the variable genes and IGRs and particularly in the CEs that are located upstream of the genes. The validation is required prior to studies of whether these variable genes and CEs play important roles in the physiological activities of *N. meningitidis*. The second suggestion is to achieve the same design but with a larger data set that may give more comprehensive ideas on genetic variability that may be relevant for vaccination and immunization programs.

Appendix

1 Parameters of genome sequences assembled using SPAdes. Isolate name-T: refers to paired and single end data; isolate name-P: refers to paired end data after reassembly. The N50 and L50 values in most isolates indicate that paired end data assembly (new assembly) is better than (old assembly) paired end and single end data assembly.

Isolates	Paired-reverse and forward	Single-reverse	Single-forward	Sum of single reverse and forward	% removed	N50	L50
N59.1-T	1751991	1692	31767	33459	1.873526989	19518	32
N59.1-P						19958	31
N59.3-T	1606150	1601	26704	28305	1.731366327	20747	30
N59.3-P						20747	29
N59.4-T	1845747	1858	34182	36040	1.914705298	20084	29
N59.4-P						21063	28
N59.6-T	1841955	1672	32771	34443	1.835194298	19682	32
N59.6-P						19682	31

N59.7-T	1800747	1715	33425	35140	1.913611913	611	6
N59.7-P						841	3
N59.9-T	1847534	1742	32309	34051	1.809255358	21526	29
N59.9-P						21526	29
N59.10-T	1627337	1577	29187	30764	1.854912376	20228	32
N59.10-P						20232	32
N59.11-T	1783445	1684	30707	32391	1.783421465	21680	30
N59.11-P						21760	30
253.1-T	1640642	1574	29644	31218	1.866802689	20468	32
253.1-P						690	6
253.3-T	1987340	1967	39060	41027	2.022126232	20229	31
253.3-P						20510	31
253.5-T	1867337	1781	33801	35582	1.869395544	650	6

253.5-P						579	5
253.7-T	2059438	1977	38800	40777	1.941016924	658	6
253.7-P						745	2
253.8-T	1901078	1774	32775	34549	1.784478359	21555	28
253.8-P						21555	28
253.9-T	1741430	1667	32098	33765	1.90157324	21584	29
253.9-P						21584	29
253.10-T	1829336	1754	32769	34523	1.851807152	20142	30
253.10-P						20747	29
352.1-T	1882585	1816	34157	35973	1.874569112	20218	33
352.1-P						20217	32
352.2-T	2031489	1918	39239	41157	1.985230276	568	6
352.2-P						626	8
352.3-T	2045922	1954	37764	39718	1.903866213	623	8

352.3-P						648	4
352.7-T	2059438	1977	38800	40777	1.941016924	551	9
352.7-P						542	3
352.8-T	1933954	1843	36089	37932	1.923161748	20522	32
352.8-P						20747	31
352.9-T	1940632	1841	36933	38774	1.958348107	20147	32
352.9-P						20504	31
352.10-T	1850995	1775	34503	36278	1.92177805	20576	30
352.10-P						20576	30
438.1-T	1887967	1760	32400	34160	1.776815163	20406	32
438.1-P						20747	30
438.4-T	1573557	1541	28162	29703	1.852191351	600	3
438.4-P						603	6
438.5-T	1846607	1761	33793	35554	1.888502201	711	6

438.5-P						598	5
438.6-T	1671105	1536	32011	33547	1.967475975	657	9
438.6-P						633	5
438.7-T	1991464	1857	37840	39697	1.953892527	20471	31
438.7-P						20471	31
438.8-T	1855231	1717	33778	35495	1.876835791	20228	32
438.8-P						20747	31
438.9-T	1881752	1821	32212	34033	1.776049606	22572	29
438.9-P						22572	29

9 List of the genes filtered from analysis of variation. The variable gene depicts with yellow colour showed significant match with the variable genes in the study.

Locus	Gene	Function
BACT000060	-	-
BACT000065	-	-
NEIS0045	rfbC	-
NEIS0048	galE	UDP-glucose epimerase
NEIS0116	tuf	tuf
NEIS0128	tuf	tuf
NEIS0213	pglA	pilin glycosyltransferase
NEIS0222	-	Hypothetical protein
NEIS0273	-	putative thiol:disulphide interchange protein
NEIS0276	-	putative rotamase
NEIS0361	-	Hypothetical protein
NEIS0380	pgll	O-acetyltransferase
NEIS0439		Secretion protein
NEIS0524		Putative peptidase
NEIS0568	pglE	Glycosyltransferase
NEIS0581	galU	glucose 1-phosphate uridylyltransferase
NEIS0595	-	Hypothetical protein
NEIS0599	-	alternative toxic C-terminal extremity
NEIS0795	-	putative periplasmic protein
NEIS0901	-	putative oxidoreductase
NEIS0932	sucB	dihydrolipoamide succinyltransferase E2 component (EC 2.3.1.61)
NEIS0950	-	phage replication initiation protein
NEIS0953	-	Hypothetical protein
NEIS0957	-	Hypothetical protein
NEIS0967	-	amidase
NEIS1156	-	Hypothetical protein
NEIS1357	-	Hypothetical protein
NEIS1447	-	acetate kinase
NEIS1516	-	putative polyamine permease substrate-binding protein
NEIS1574	-	DNA transport competence protein
NEIS1653	-	Hypothetical protein
NEIS1661	-	Hypothetical protein
NEIS1664	-	Hypothetical protein
NEIS1689	-	putative polyamine permease substrate-binding protein
NEIS1750	-	Hypothetical protein
NEIS1778	galM	aldose 1-epimerase

NEIS1789	$mafA_{MGI-1}$	MafA adhesin
NEIS1795	$mafI_{o2MGI-2}$	MafI immunity protein
NEIS1796	-	Hypothetical protein
NEIS1805	-	Hypothetical protein
NEIS1859	autA	autotransporter A
NEIS1880	-	DNA transport competence protein
NEIS1902	lgtA	lacto-N-neotetraose biosynthesis glycosyl transferase
NEIS1943	-	outer membrane peptidase
NEIS2083	$mafA_{MGI-3}$	MafA3 lipoprotein
NEIS2148	pgk	phosphoglycerate kinase (EC 2.7.2.3)

27 Genes excluded from analysis of variation by GC method. Filtered genes mostly found with more than one copy in the genomes and encode transposes.

Number	Gene	Function	Filtration
1	*NMB0141*	transposase, truncation	More than one copy
2	*NMB0487*	hypothetical protein	More than one copy
3	*NMB0522*	transposase, truncation	More than one copy
4	*NMB0583*	IS1016C2 transposase	More than one copy
5	*NMB0701*	hypothetical protein	More than one copy
6	*NMB0891*	hypothetical protein	More than one copy
7	*NMB0919*	IS1106 transposase	More than one copy
8	*NMB0991*	IS1106 transposase	More than one copy
9	*NMB1157*	hypothetical protein	More than one copy
10	*NMB1195*	hypothetical protein	More than one copy
11	*NMB1411*	IS1016C2 transposase	More than one copy
12	*NMB1539*	IS1106 transposase	More than one copy
13	*NMB1553*	transposase, truncation	More than one copy
14	*NMB1601*	IS1106 transposase	More than one copy
15	*NMB1401*	pseudogene	More than one copy
16	*NMB1399*	IS1106 transposase	More than one copy
17	*NMB0583*	IS1016C2 transposase	More than one copy
18	*NMB0225*	IS30 family transposase	More than one copy
19	*NMB0443*	IS30 family transposase	More than one copy

20	***NMB0635***	**IS30 family transposase**	**More than one copy**
21	***NMB0805***	**IS30 family transposase**	**More than one copy**
22	***NMB0834***	**IS30 family transposase**	**More than one copy**
23	***NMB0911***	**IS30 family transposase**	**More than one copy**
24	***NMB1022***	**IS30 family transposase**	**More than one copy**
25	***NMB1050***	**IS30 family transposase**	**More than one copy**
26	***NMB1099***	**IS30 family transposase**	**More than one copy**
27	***NMB1259***	**IS30 family transposase**	**More than one copy**
28	***NMB2148***	**IS30 family transposase**	**More than one copy**
29	***NMC1156***	**putative glycosyl transferase**	**Phase variable gene**
30	***NMB1969***	**serotype-1-specific antigen**	**Phase variable gene**
31	***NMB0451***	**hypothetical protein**	**Corriea element**
32	***NMB0558***	**hypothetical protein**	**Corriea element**

28 Estimation the nucleotide diversity of each 25 pair isolates. The values of nucleotide diversity with yellow colour refers into the highest values among all the values of nucleotide diversity of 25 pair isolates.

Isolates	Number of SNPs	Nucleotide diversity Pi
17004	17	0.00001
17002	16	0.00001
17003	6	0
17005		
17006	4	0
17007	2	0
17008	4	0
17009	1	0
17010		
17011	3	0
17012	529	0.0026
17013	2	0
17014	5	0
17015	4	0
17016	278	0.00014
17017	14	0.00001
17018	5	0
17019-20	3	0
17021	319	0.00016
17022	3	0
17023		
17024	3	0
17025	4	0
17026	2	0

29 Functions of all variable loci all CCs in detail. The function was grouped using KEGG scheme for all 131 variable genes found in the 25 paired isolates.

Different schemes	Number of genes	Number of genes in all genome	Percentage of proportion effect (number of gene/ number of genes in all genome in each scheme) %
1- Genetic Information Processing		388	4.6
A -Replication and Repair	6		
B-Transcription	7		
C-Translation	5		
D-Folding, sorting and degradation			
2 - Metabolism		695	5
amino acid metabolism	11		
Carbohydrate metabolism	8		
Metabolism of cofactor and vitamins	6		
Energy metabolism	5		
Nucleotide metabolism	4		
Glycan biosynthesis and metabolism	2		
Xenobiotics biodegradation and metabolism	1		
5 - hypothetical protein	39	208	18.7
6- pseudoprotein	5		
10 - Environmental Information Processing (membrane protein)		207	15.4
A - Membrane transport (ABC transporter, Bacterial secretion system and phosphotransferase system)	31		
B - Signal transduction	1		
Total	131	1634	

30 Estimation the diversity scores for each variable gene in different schemes. The diversity score for each scheme that classified using KEGG was estimated relying on estimation of diversity scores of each variable gene listed below (See chapter two 2.7.3).

Gene	Pi	Gene	Pi	Gene	Pi	Gene	Pi	Gene	Pi
N59_0006	0.00226	*NMB0424*	0.001	*NMB0700*	0.00081	*NMB1429*	0.0008	*NMB2160*	0.00039
N59_00721	0.00224	*NMB0424*	0.00109	*NMB0777*	0.00135	*NMB1468*	0.02778	*NMB2160*	0.00039
NEIS0962	0.0008	*NMB0437*	0.0743	*NMB0786*	0.00089	*NMB1485*	0.00064		
NEIS1021	0.0045	*NMB0438*	0.00271	*NMB0841*	0.0013	*NMB1507*	0.01667		
NEIS1623	0.00151	*NMB0439*	0.00866	*NMB0850*	0.00241	*NMB1537*	0.00056		
NEIS1946	0.09333	*NMB0440*	0.0091	*NMB0856*	0.0021	*NMB1568*	0.0022		
NEIS1947	0.02181	*NMB0441*	0.0172	*NMB0866*	0.00126	*NMB1569*	0.00065		
NMB0040	0.0007	*NMB0444*	0.00117	*NMB0884*	0.00085	*NMB1573*	0.001		
NMB0055	0.00084	*NMB0449/450*	0.00054	*NMB0885*	0.00071	*NMB1585*	0.00231		
NMB0055	0.00126	*NMB0460*	0.04809	*NMB0920*	0.00045	*NMB1593*	0.0011		
NMB0133	0.00024	*NMB0475*	0.0007	*NMB0927*	0.0246	*NMB1682*	0.0005		
NMB0185	0.00125	*NMB0480*	0.005	*NMB0930*	0.0459	*NMB1714*	0.00071		
NMB0215	0.00176	*NMB0504*	0.0555	*NMB0962*	0.00035	*NMB1797*	0.00071		
NMB0216	0.02904	*NMB0505*	0.02105	*NMB0976*	0.00505	*NMB1821*	0.00085		
NMB0217	0.0258	*NMB0514*	0.00415	*NMB1029*	0.00072	*NMB1830*	0.00455		
NMB0274	0.00043	*NMB0515*	0.052	*NMB1059*	0.0045	*NMB1880*	0.00207		
NMB0319	0.0013	*NMB0516*	0.0235	*NMB1100*	0.00207	*NMB1883*	0.0034		
NMB0319	0.00131	*NMB0517*	0.00641	*NMB1104*	0.00426	*NMB1923*	0.0026		
NMB0326	0.0051	*NMB0518*	0.0434	*NMB1116*	0.0076	*NMB1926*	0.017		
NMB0329	0.00358	*NMB0524*	0.25	*NMB1240*	0.00061	*NMB1970*	0.0174		
NMB0329	0.006	*NMB0540*	0.00838	*NMB1241*	0.0008	*NMB1971*	0.00988		
NMB0330	0.01429	*NMB0541*	0.02811	*NMB1288*	0.00087	*NMB1985*	0.00023		
NMB0331	0.01422	*NMB0557*	0.00295	*NMB1298*	0.00144	*NMB1988*	0.05448		
NMB0332	0.0197	*NMB0594*	0.00071	*NMB1341*	0.00038	*NMB2040*	0.04522		
NMB0333	0.00081	*NMB0631*	0.00067	*NMB1345*	0.06899	*NMB2041*	0.02169		
NMB0341	0.00038	*NMB0666*	0.0004	*NMB1346*	0.04771	*NMB2057*	0.00231		
NMB0363	0.0652	*NMB0669*	0.001	*NMB1347*	0.0292	*NMB2065*	0.00079		
NMB0368	0.0022	*NMB0670*	0.00161	*NMB1348*	0.05147	*NMB2149*	0.00385		
NMB0383	0.0045	*NMB0682*	0.04058	*NMB1390*	0.00101	*NMB2149*	0.01154		
NMB0409	0.0008	*NMB0683*	0.03052	*NMB1410*	0.0079	*NMB2159*	0.001		

31 The position of allelic variation (amino acid variation) relative to conserved domains for CC-167 isolates. The number of isolates were 4.

Genes	Position	Variation in amino acid	Domains	Identifier
NMB2039	203	deletion DNVKI	outer membrane channels share a beta-barrel structure	cl21487
NMB1429	199	P to S	Gram-negative porin	pfam00267
NMB2039	203	deletion DNVKI	outer membrane channels share a beta-barrel structure	cl21487
NMB0409	101	A to V	SAM-dependent methyltransferase	COG1565
NMB0326	298	A to T	Isoprenoid Biosynthesis enzymes	cl00210
NMB0329	29	N to K	type IV-A pilus assembly ATPase PilB	TIGR02538
NMB0330	2,23,66	T to A, K to E, R to M	Endogenous inhibitor of DNA gyrase	COG3024
NMB0331	124	S to I	Dephospho-CoA kinase	pfam01121
NMB0332	47,56	K to E, K to E	Bacterial Peptidase A24 N-terminal domain (Type IV leader peptidase family)	pfam06750
NMB0437	28,32,68, 77,93,143	S to A, V to S, A to S, S to A, E to P, P to L	DNA-binding transcriptional regulator	COG1959
NMB0439	322	A to S	Uncharacterized protein involved in response to NO [Defense mechanisms]	COG3213
NMB0440	119	R to H	Prephenate dehydrogenase [Amino acid transport and metabolism]	COG0287
NMB0441	32	A to E	Nit1, Nit 2, and related proteins	cd07572
	140	D to E		
NMB0480	68	V to I	*Nm* TspB protein	pfam05616
NMB0516	33,70.86, 89	S to A, K to N, Q to R, DDE to NST	Immunity protein 49; A predicted immunity protein with an all alpha-helical fold	pfam15575
NMB0541	21,153	P to T, Q to R	Domain of unknown function (DUF4375)	pfam14300
NMB0682	58,154,30 7	R to C, E to G, G to D	dihydroorotase	PRK05451
NMB0683	63	RQ to QK	transcription antitermination protein NusB	PRK00202
NMB0976	68	V to I	*Nm* TspB protein	pfam05616
NMB1100	16	stop codon	SEC10/PgrA surface exclusion domain	TIGR04320
NMB1104	10,395	L to P, T to P	Mu-like prophage tail sheath protein gpL	COG4386

NMB1345	59,190,353,370,444,449,475,508	T to M, A to S, A to T, D to N, G to A, D to A, N to D, A to G	Uncharacterized conserved protein YdgA	COG5339
NEIS1282	186,219,229,239,277,296	T to K, G to D, P to L, Q to E, deletion LN, deletion VTARNY	Porin superfamily. These outer membrane channels share a beta-barrel structure	cl21487
NMB1348	10,128,193	A to D, P to S, M to S	SpoU rRNA Methylase family	cl21505

32 The position of allelic variation (amino acid variation) relative to conserved domains for CC-174, CC-32, CC-60 CCs and the CC-1157-32-269. The number of isolates were 8, 5, 4 and 4.

Genes	Position	Variation in amino acid	Domains	Identifier
NMB1715	638	DAVAI deleted	Hydrophobe/Amphiphile Efflux-1 (HAE1) Family; Proteins	TIGR00915
NMB1926	25,40	H to R, A to R	Glycosyltransferase involved in LPS biosynthesis	COG3306
N59_01116	225	Internal stop	Glycosyltransferase_GTB_type super family	Cl10013
NMB1341	33	Q to R	pyruvate dehydrogenase subunit E1	PRK09405
NMB0055	182	A to T	Pyrroline-5-carboxylate reductase dimerization	pfam14748
NMB0631	171	R to H	Phosphotransacetylase	COG0280
NMB2160	195	A to T	MutS domain II	pfam05188
NMB0557	99	N to H	iron-sulfur cluster insertion protein ErpA	PRK13623
NMB1288	23	E to Q	ribonucleotide-diphosphate reductase subunit beta	PRK09101
NMB2057	67	V to A	50S ribosomal protein L13	PRK09216
NMB0777	43	S to F	uroporphyrinogen-III synthase	PRK05928
NMB1390	223	p to L	glucokinase	PRK00292
NMB2065	43	T to A	unknown domain/N5-glutamine S-adenosyl-L-methionine-dependent methyltransferase	PRK14966

			fusion	
NMB0700	802	stop to Q 802 many stop after this position	Immunoglobulin A1 protease	pfam02395
NMB2160	358	PRDLAAL addition	MutS domain III	pfam05192
NMB0329	445	R to L	type IV-A pilus assembly ATPase PilB	TIGR02538
NMB0927	194,204,248,308	D to A, C to R, Q to R, H to R	proline iminopeptidase	TIGR01249
NMB0460	99,137,169, 171, (137-405)	N to D, deletion SINGG, LE, NNLI, many variation	Transferrin binding protein-like solute binding protein	pfam01298
NMB2040	43,125,163, 200,210	R to S, K to Q, SQ to FR, R to H, R to Q	Thiamine biosynthesis protein ThiC	COG0422
NMB2041	39	R to G	Xanthine/uracil permease	COG2233
NMB1880	289	E to stop codon	Siderophore binding protein FatB	cd01140
NMB1241	184	E to K	Probable RNA and SrmB-binding site of polymerase A	pfam12627
NMB0666	92	L to F	NAD+ dependent DNA ligase adenylation domain	cd00114
NMB2160	214	Q to stop codon	MutS domain V	pfam00488
NMB1485	138	I to M	Membrane protein TerC	COG0861
NMB1029	270	S to F	Aspartase	cd01357
NMB0055	209	Q to stop codon	Pyrroline-5-carboxylate reductase dimerization	pfam14748
NMB0884	192	V to A	Superoxide dismutase	pfam02777
NMB0217	105,147,174,182,208	A to S, Q to K, L to S, A to S, H to R	Sigma-54 factor, core binding domain	pfam04963
NMB1970	184, 384-489	N to K, muli variation	Amino-transferase class IV	pfam01063
NMB1971	218	R to S	Porin superfamily	cl21487
NMB1988	269-279	muli variation	Porin superfamily	cl21487
NMB0216	192	S to G	Clade 3 of the heme-binding enzyme catalase	cd08156
NEIS1947	244	addition G	TonB dependent/Ligand-Gated channels	cd01347
NMB1821	251	R to H	dTDP-4-amino-4,6-dideoxygalactose transaminase	COG0399
NMB0215	57	G to S	Putative Mn2+ efflux pump MntP	COG1971

33 Estimation of genic variation (mutable genes/genome) and (mutation rate/month) within the 25 paired isolates. The number with red colour indicates the value of genic variation (mutable genes/genome) and (mutation rate/month) for each CC.

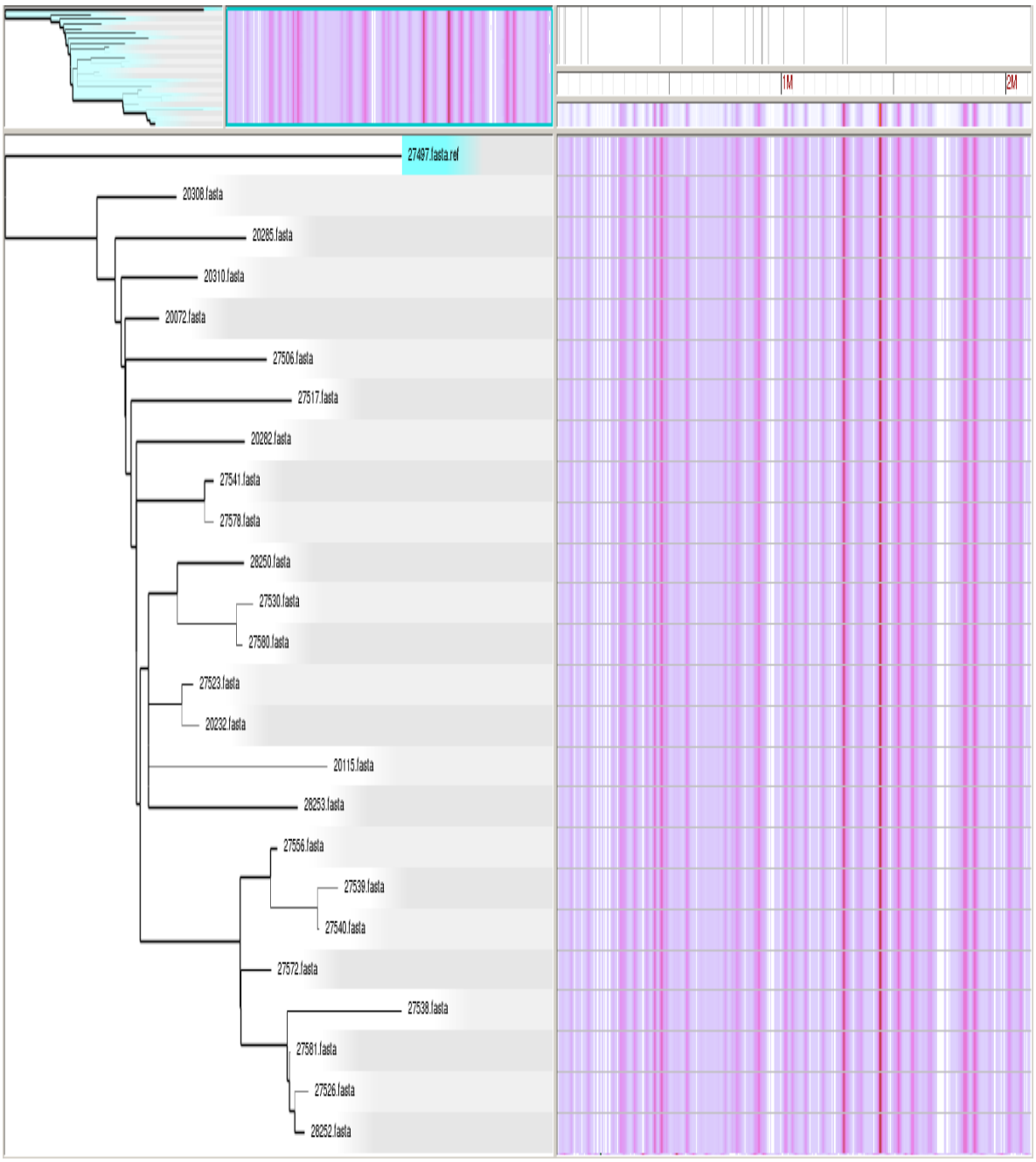
Isolate pair	Total genes in genome	Actual Variable Genes	Genic variation (mutabe genes/genome)	Mut Rate (genic variation /Month)
16976 (N51.1)- 17002 (N424.1)	1867	4	0.002142475	0.0003571
16975 (N241.1)- 17001 (N349.1)	1839	1	0.000543774	0.0002719
16977 (N52.1)- 17003 (N342.1)	1847	2	0.001082837	0.0003609
16979 (N58.1)- 17005 (N429.1)	1855	0	0	0
16980 (N59.1)- 17006 (N438.1)	1880	4	0.00212766	0.0003546
16981 (N88.1)- 17007 (N449.1)	1868	3	0.0016059957	0.000267666
16982 (N138.1)- 17008 (N331.1)	1858	4	0.002152853	0.0007176
16978 (N54.1)- 17004 (N343.1)	1847	19	0.010286952	0.003429
			0.0024928181	0.0007198
16983 (N64.1)- 17009 (N348.1)	1869	2	0.001070091	0.0003567
16984 (N117.1)- 17010 (N417.1)	1886	5	0.002651113	0.0004419
16985 (N124.1)- 17011 (N336.1)	1870	4	0.002139037	0.000713
16986 (N128.1)- 17012 (N420.1)	1899	35	0.018430753	0.0030717922
			0.0060727487	0.001145
16987 (N258.1)- 17013 (N431.1)	1853	1	0.000539665	0.0001079
16988 (N264.1)- 17014 (N359.1)	1890	4	0.0015873016	0.00079365
16989 (N259.1)- 17015 (N445.1)	1800	5	0.002777778	0.0005556
16990 (N222.1)- 17016 (N459.1)	1649	5	0.003032141	0.0005054
16997 (N188.1)- 17023 (N462.1)	1871		0	
			0.0016931972	0.0005567616
16991 (N114.1)- 17017 (N330.1)	1849	8	0.004326663	0.0014422
16992 (N134.1)- 17018 (N333.1)	1832	2	0.001091703	0.0003639
16993 (N185.1)- 16994 (N185.2)	1853	6	0.003237992	0.0005397
17019 (N456.1)- 17020 (N456.2)				
16995 (N262.1)- 17021 (N446.1)	1822	11	0.006037322	0.0010062
			0.0037099019	0.0008440822
16996 (N73.1)- 17022 (N450.1)	1779	3	0.001686341	0.0002811
16998 (N199.1)- 17024 (N378.1)	1889	1	0.000529381	0.0001765
16999 (N176.1)- 17025 (N408.1)	1985	1	0.000503778	8.40E-05
17000 (N86.1)- 17026 (N447.1)	1787	2	0.001119194	0.0001865
			0.0009497239	0.0001803448

34 The IGRs varied in more than one pair of isolates of 25 paired isolates belonging into different CCs.

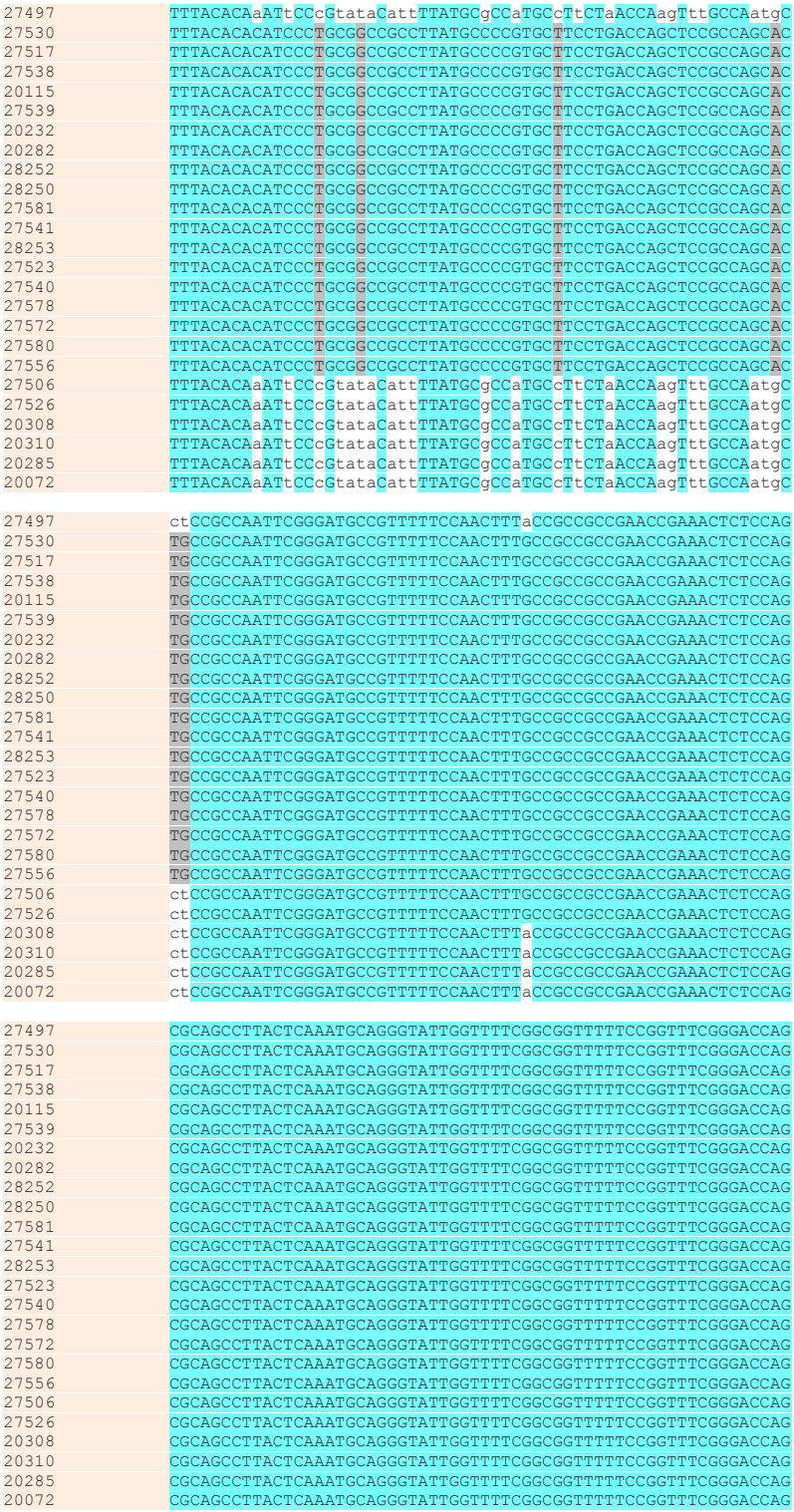
Isolate Pair and number of copies	CC-174 complex	Variable loci	NMB fraction	Type of variation
16975_17001				
3		*NEIS0651*	*NMB0700*	A to T
5		*NEIS0978*	*NMB0992*	C to T
16976_17002	CC-174 complex			
2		*NEIS0824*	*NMB0883*	A to T
4		*NEIS1615*	*NMB1699*	T to C
16977_17003	CC-174 complex			
2		*NEIS1468*		T To C and C to G
16978_17004	CC-174 complex			
8		*NEIS0405*	*NMB1815*	A to C
2		*NEIS0652*	*NMB0702*	T to C
2		*NEIS0825*		T to A
4		*NEIS1617*	*NMB1703*	GC to AT
16979_17005	CC-174 complex			
16980_17006	CC-174 complex			
4	Non-tri	*NEIS1428*	*NMB1497*	highly variable
16981_17007	CC-174 complex			
3	promoter	*NEIS0186*	*NMB0196*	highly variable
16982_17008	CC-174 complex			
2		*NEIS1587*	*NMB1669*	C to T
16983_17009				
3		*NEIS1994*	*NMB2015*	A to G
16984_17010	CC-167 complex			
4		*NEIS1900*	*NMB1926*	C to A,G to A,A to C, A to G
16985_17011	CC-167 complex			
16986_17012	CC-167 complex			
16987_17013	CC-23 complex/Cluster A3			
16988_17014	CC-23 complex/Cluster A3			
2		*BACT000032*	*NMB0143*	T To C
2		*NEIS0407*	*NMB1813*	highly variable
2		*NEIS2020*	*NMB2039*	T To C
16989_17015	CC-23 complex/Cluster A3			

16990_17016	CC-23 complex/Cluster A3			
4		*NEIS2153*	*NMB0015*	C to T
16992_17018	CC-60 complex			
2		*NEIS1492*		deletion T
2		*NEIS0202*	*NMB0210*	C toG
16998_17024	CC-269 complex			

35 Phylogenetic trees constructed using ParSnp software. Tree was constructed using maximum likelihood algorithm from entire assembled genomes of 25 CC-174 isolates. The panel on the right showed the recombinant SNPs depicted as line with pink color.



36 This figure showing a part of alignment of the DNA sequence of 25 CC-174 in *NEIS1464* gene. The recombination was inferred between isolate ID: 27497 and other isolates IDs (27506, 27526, 20308, 20310, 20285 and 20072).



37 The number of recombination blocks, hot spot and large importation regions detected using the ClonalFrameML program.

These recombination events were detected within disease and carriage isolates of CC-174. The disease isolates were highlighted with yellow color. The largest import fragment occurred in the node N27 (node between ID: 27540 and ID: 27539) with length of (5594 bp) while the length was (5508 bp) in the 27506 isolate.

Isolates	Number of recombination block	Size Kbp (Lower-Upper limited)	Hot spot regions(three signals in the same location)	Large importation regions (> 3Kbp)
27497	40	21-3255	11	1
27517	23	29-3992	3	1
28253	18	9-1924		
20285	15	206-1790	1	1
20115	11	27-1651	2	
27506	11	108-5508	5	1
28250	11	27-1302	4	
20308	10	42-3388	3	1
20310	8	15-3580	3	1
27572	8	102-968	1	
20282	6	30-4858	3	2
27538	3	1223-3063		1
27541	3	50		
20232	1	130		
27523	1	140		
27530	1	10	1	
27539	1	2198		
27540	1	9	1	
27556	1	256		
27580	1	58		
27526	1	2	1	
27578	1	56		
28252	1	326		
Node_26	37	(17-2494)	9	
Node_27	11	(5-5594)	7	1
Node_28	10	(22-2145)	4	
Node_29	6	(99-1643)	4	

Node_30	3	(597-796)	3	
Node_31	5	(117-726)	4	
Node_32	2	(74-2571)	1	
Node_33	2	(31-92)	1	
Node_34	1	817		
Node_35	2	(2-58)	1	
Node_37	1	58		
Node_38	1	116		
Node_39	1	84		
Node_40	3	(65-2037)	2	
Node_41	9	(2-2460)	2	
Node_42	1	248	1	
Node_43	1	20	1	
Node_44	2	(9-1286)	1	
Node_45	3	(120-860)		
Node_46	4	(215-2518)		
Node_47	2	(153-1734)		

38 Listing the name of isolates, genes, size of repeat tracts, height of peak, repeat length and ON/OFF state. These are the results of GeneScan for four phase variable genes (*opc*, *hpuAB*, *nalP* and *hmbr*) and 15 isolates under the study.

Isolates	Gene	Size	Repeat length	ON/OFF state or Intermediate/high state
419.1	*opc*	240.6	8	Intermed.or less
419.2	*opc*	237.6	5	Intermed.or less
419.3	*opc*	237.5	5	Intermed.or less
419.4	*opc*	237.6	5	Intermed .or less
419.5	*opc*	237.5	5	Intermed.or less
419.6	*opc*	237.6	5	Intermed.or less
436.1	*opc*	245.3	13	high
436.2	*opc*	245.2	13	high
436.3	*opc*	242.2	10	Intermed.or less
436.4	*opc*	245.2	13	high
436.5	*opc*	246.1	14	high
436.6	*opc*	241.08	9	Intermed.or less
428.1	*opc*	242.2	10	Intermed.or less
428.2	*opc*	242.1	10	Intermed.or less
428.3	*opc*	246.2	14	high
419.1	*hpuAB*	362.1	12	off
419.2	*hpuAB*	362.1	12	off
419.3	*hpuAB*	361.2	11	off
419.4	*hpuAB*	361.1	11	off
419.5	*hpuAB*	361.1	11	off
419.6	*hpuAB*	361.2	11	off
436.1	*hpuAB*	365.9	15	off
436.2	*hpuAB*	357.8	7	On
436.3	*hpuAB*	357.5	7	On
436.4	*hpuAB*	357.5	7	On
436.5	*hpuAB*	357.5	7	On
436.6	*hpuAB*	353.8	5	off
428.1	*hpuAB*	353.8	5	off
428.2	*hpuAB*	360.2	10	On
428.3	*hpuAB*	354.07	6	off
419.1	*nalp*	315.1	9	off
419.2	*nalp*	315.4	9	off
419.3	*nalp*	315.1	9	off

419.4	*nalp*	314.1	8	off
419.5	*nalp*	314.5	8	off
419.6	*nalp*	314.1	8	off
436.1	*nalp*	316	10	On
436.2	*nalp*	316.4	10	On
436.3	*nalp*	314.1	8	off
436.4	*nalp*	316	10	On
436.5	*nalp*	316	10	On
436.6	*nalp*	316	10	On
428.1	*nalp*	315.4	9	off
428.2	*nalp*	315.4	9	off
428.3	*nalp*	315.2	9	off
419.1	*hmbr*	427.08	9	On
419.2	*hmbr*	427.1	9	On
419.3	*hmbr*	426.8	8	off
419.4	*hmbr*	426.9	8	off
419.5	*hmbr*	427.7	9	On
419.6	*hmbr*	430.1	12	On
436.1	*hmbr*	430	12	On
436.2	*hmbr*	427	9	On
436.3	*hmbr*	427	9	On
436.4	*hmbr*	427	9	On
436.5	*hmbr*	427	9	On
436.6	*hmbr*	430	12	On
428.1	*hmbr*	430	12	On
428.2	*hmbr*	427	9	On
428.3	*hmbr*	430	12	On

References

Aaron, R., Q., Ira, M., H., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 841–842.

Achaz, G., Coissac, E., Netter, P., Rocha, E.P., 2003. Associations between inverted repeats and the structural evolution of bacterial genomes. Genetics. 164, 1279-1289.

Adamiak, P., Calmettes, C., Moraes, T.F., Schryvers, A.B., 2015. Patterns of structural and sequence variation within isotype lineages of the *Neisseria meningitidis* transferrin receptor system. Microbiologyopen. 4, 491-504.

Ahmed-Abakur, E.H., 2014. Meningococcal Meningitis: Etiology, Diagnosis, Epidemiology and Treatment. American Journal of Medicine and Medical Sciences. 4, 266-271.

Aho, E.L., Botten, J.W., Hall, R.J., Larson, M.K., Ness, J.K., 1997. Characterization of a class II pilin expression locus from *Neisseria meningitidis*: evidence for increased diversity among pilin genes in pathogenic Neisseria species. Infection and Immunity. 65, 2613-2620.

Aho, E.L. and Cannon, J.G., 1988. Characterization of a silent pilin gene locus from *Neisseria meningitidis* strain FAM18. Microbial Pathogenesis. 5, 391-398.

Alamro, M., Bidmos, F.A., Chan, H., Oldfield, N.J., Newton, E., Bai, X., Aidley, J., Care, R., Mattick, C., Turner, D.P., 2014. Phase variation mediates reductions in expression of surface proteins during persistent meningococcal carriage. Infection and Immunity. 82, 2472-2484.

Ali, O., Aseffa, A., Bedru, A., Lema, T., Moti, T., Tekletsion, Y., Worku, A., Xabher, H.G., Yamuah, L., Boukary, R.M., 2015. The diversity of meningococcal carriage across the African meningitis belt and the impact of vaccination with a group A meningococcal conjugate vaccine. Journal of Infectious Diseases. 212, 1298-1307.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology. 215, 403-410.

Ambur, O.H., Frye, S.A., Tnjum, T., 2007. New functional identity for the DNA uptake sequence in transformation and its presence in transcriptional terminators. Journal of Bacteriology. 189, 2077-2085.

Ambur, O.H., Frye, S.A., Nilsen, M., Hovland, E., Tnjum, T., 2012. Restriction and sequence alterations affect DNA uptake sequence-dependent transformation in *Neisseria meningitidis*. PloS One. 7, e39742.

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. 175-176.

Andrews, T.D. and Gojobori, T., 2004. Strong positive selection and recombination drive the antigenic variation of the PilE protein of the human pathogen *Neisseria meningitidis*. Genetics. 166, 25-32.

Anisimova, M., Nielsen, R., Yang, Z., 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics. 164, 1229-1236.

Arber, W., 2014. Horizontal gene transfer among bacteria and its role in biological evolution. Life. 4, 217-224.

Assalkhou, R., Balasingham, S., Collins, R.F., Frye, S.A., Davidsen, T., Benam, A.V., Bjrs, M., Derrick, J.P., Tnjum, T., 2007. The outer membrane secretin *PilQ* from *Neisseria meningitidis* binds DNA. Microbiology. 153, 1593-1603.

Assefa, S., Keane, T.M., Otto, T.D., Newbold, C., Berriman, M., 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics. 25, 1968-1969.

Azad, R.K. and Lawrence, J.G., 2007. Detecting laterally transferred genes: use of entropic clustering methods and genome position. Nucleic Acids Research. 35, 4629-4639.

Bairoch, A.M., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro Rojas, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., 2005. The universal protein resource (UniProt). Nucleic Acids Research. 33, 154.

Bambini, S., De Chiara, M., Muzzi, A., Mora, M., Lucidarme, J., Brehony, C., Borrow, R., Masignani, V., Comanducci, M., Maiden, M.C., 2014. Neisseria adhesin A variation and revised nomenclature scheme. Clinical and Vaccine Immunology. 21, 966-971.

Bandyopadhyay, S., Chandramouli, K., Johnson, M.K., 2008. Iron–sulfur Cluster Biosynthesis. 1112-1119.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology. 19, 455-477.

Bansal, A.K., 2005. Bioinformatics in microbial biotechnology–a mini review. Microbial Cell Factories. 4, 19.

Bårnes, G.K., Brynildsrud, O.B., Børud, B., Workalemahu, B., Kristiansen, P.A., Beyene, D., Aseffa, A. and Caugant, D.A., 2017. Whole genome sequencing reveals within-host genetic changes in paired meningococcal carriage isolates from Ethiopia. BMC genomics, 18, 407.

Bart, A., Barnab, C., Achtman, M., Dankert, J., van der Ende, A., Tibayrenc, M., 2001. The population structure of *Neisseria meningitidis* serogroup A fits the predictions for clonality. Infection, Genetics and Evolution. 1, 117-122.

Bayliss, C.D., 2009. Determinants of phase variation rate and the fitness implications of differing rates for bacterial pathogens and commensals. FEMS Microbiology Reviews. 33, 504-520.

Bayliss, C.D., Field, D., Moxon, E.R., 2001. The simple sequence contingency loci of *Haemophilus influenzae* and *Neisseria meningitidis*. The Journal of Clinical Investigation. 107, 657-666.

Bayliss, C.D., Hoe, J.C., Makepeace, K., Martin, P., Hood, D.W., Moxon, E.R., 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, 5038-5048.

Beddek, A.J., Li, M., Kroll, J.S., Jordan, T.W., Martin, D.R., 2009. Evidence for capsule switching between carried and disease-causing *Neisseria meningitidis* strains. Infection and Immunity. 77, 2989-2994.

Behrouzi, A., Bouzari, S., Siadat, S.D., Irani, S., 2014. In silico studies of outer membrane of *Neisseria meningitidis* por a: its expression and immunogenic properties. International Journal of Molecular and Cellular Medicine. 3, 166.

Benam, A.V., Lång, E., Alfsnes, K., Fleckenstein, B., Rowe, A.D., Hovland, E., Ambur, O.H., Frye, S.A., Tønjum, T., 2011. Structure–function relationships of the competence lipoprotein ComL and SSB in meningococcal transformation. Microbiology. 157, 1329-1342.

Benghezal, M., Gauntlett, J.C., Debowski, A.W., Fulurija, A., Nilsson, H.O. and Marshall, B.J., 2014. Persistence of *Helicobacter pylori* infection: Genetic and epigenetic diversity. In Trends in Helicobacter pylori Infection. InTech.

Bennett, D.E. and Cafferkey, M.T., 2003. Multilocus restriction typing: a tool for *Neisseria meningitidis* strain discrimination. Journal of Medical Microbiology. 52, 781-787.

Bennett, J.S., Bentley, S.D., Vernikos, G.S., Quail, M.A., Cherevach, I., White, B., Parkhill, J. and Maiden, M.C., 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, 652.

Bentley, S.D., Vernikos, G.S., Snyder, L.A., Churcher, C., Arrowsmith, C., Chillingworth, T., Cronin, A., Davis, P.H., Holroyd, N.E., Jagels, K., 2007. Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. PLoS Genet. 3, e23.

Benz, I. and Schmidt, M.A., 2011. Structures and functions of autotransporter proteins in microbial pathogens. International Journal of Medical Microbiology. 301, 461-468.

Berry, J., Cehovin, A., McDowell, M.A., Lea, S.M., Pelicic, V., 2013. Functional analysis of the interdependence between DNA uptake sequence and its cognate *ComP* receptor during natural transformation in Neisseria species. PLoS Genet. 9, e1004014.

Berry, J., Xu, Y., Ward, P.N., Lea, S.M., Matthews, S.J., Pelicic, V., 2016. A comparative structure/function analysis of two type IV pilin DNA receptors defines a novel mode of DNA binding. Structure. 24, 926-934.

Beyene, G.T., Balasingham, S.V., Frye, S.A., Namouchi, A., Homberset, H., Kalayou, S., Riaz, T., Tnjum, T., 2016. Characterization of the *Neisseria meningitidis* Helicase RecG. PloS One. 11, e0164588.

Bidmos, F.A., Chan, H., Praekelt, U., Tauseef, I., Ali, Y.M., Kaczmarski, E.B., Feavers, I., Bayliss, C.D., 2015. Investigation into the Antigenic Properties and Contributions to Growth in Blood of the Meningococcal Haemoglobin Receptors, *HpuAB* and *HmbR*. PloS One. 10, e0133855.

Bidmos, F.A., Neal, K.R., Oldfield, N.J., Turner, D.P., Ala'Aldeen, D.A., Bayliss, C.D., 2011. Persistence, replacement, and rapid clonal expansion of meningococcal carriage isolates in a 2008 university student cohort. Journal of Clinical Microbiology. 49, 506-512.

Bilek, N., Ison, C.A., Spratt, B.G., 2009. Relative contributions of recombination and mutation to the diversification of the *opa* gene repertoire of *Neisseria gonorrhoeae*. Journal of Bacteriology. 191, 1878-1890.

Bille, E., Zahar, J., Perrin, A., Morelle, S., Kriz, P., Jolley, K.A., Maiden, M.C., Dervin, C., Nassif, X., Tinsley, C.R., 2005. A chromosomally integrated bacteriophage in invasive meningococci. Journal of Experimental Medicine. 201, 1905-1913.

Boan, P., Metasan, N., Tempone, S., Harnett, G., Speers, D.J., Keil, A.D., 2014. *Neisseria meningitidis porA*, *fetA* and *fHbp* gene distribution in Western Australia 2000 to 2011. BMC Infectious Diseases. 14, 686.

Boisier, P., Nicolas, P., Djibo, S., Taha, M., Jeanne, I., Maínassara, H.B., Tenebray, B., Kairo, K.K., Giorgini, D., Chanteau, S., 2007. Meningococcal meningitis: unprecedented

incidence of serogroup X—related cases in 2006 in Niger. Clinical Infectious Diseases. 44, 657-663.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. btu170.

Bosis, S., Mayer, A., Esposito, S., 2015. Meningococcal disease in childhood: epidemiology, clinical features and prevention. Journal of Preventive Medicine and Hygiene. 56, E121.

Bowler, L.D., Zhang, Q., Riou, J., Spratt, B.G., 1994. Interspecies recombination between the *penA* genes of *Neisseria meningitidis* and commensal Neisseria species during the emergence of penicillin resistance in *N. meningitidis*: natural events and laboratory simulation. Journal of Bacteriology. 176, 333-337.

Brandtzaeg, P., Ovsteb, R., Kierulf, P., 1995. Bacteremia and compartmentalization of LPS in meningococcal disease. Progress in Clinical and Biological Research. 392, 219.

Brandtzaeg, P., 2006. Pathogenesis and pathophysiology of invasive meningococcal disease. Handbook of Meningococcal Disease: Infection Biology, Vaccination, Clinical Management. 427-480.

Braun, J.M., Blackwell, C.C., Poxton, I.R., El Ahmer, O., Gordon, A.E., Al Madani, O.M., Weir, D.M., Giersen, S., Beuth, J., 2002. Proinflammatory responses to lipo-oligosaccharide of *Neisseria meningitidis* immunotype strains in relation to virulence and disease. Journal of Infectious Diseases. 185, 1431-1438.

Brehony, C., Jolley, K.A. and Maiden, M.C., 2006. Multilocus sequence typing for global surveillance of meningococcal disease. FEMS microbiology reviews. 31, 15-26.

Brooks, R., Woods, C.W., Benjamin, D.K., Rosenstein, N.E., 2006. Increased case-fatality rate associated with outbreaks of *Neisseria meningitidis* infection, compared with sporadic meningococcal disease, in the United States, 1994–2002. Clinical Infectious Diseases. 43, 49-54.

Bryant, J., Chewapreecha, C., Bentley, S.D., 2012. Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. Future Microbiology. 7, 1283-1296.

Buckee, C.O., Jolley, K.A., Recker, M., Penman, B., Kriz, P., Gupta, S., Maiden, M.C., 2008. Role of selection in the emergence of lineages and the evolution of virulence in *Neisseria meningitidis*. Proceedings of the National Academy of Sciences. 105, 15082-15087.

Budroni, S., Siena, E., Hotopp, J.C.D., Seib, K.L., Serruto, D., Nofroni, C., Comanducci, M., Riley, D.R., Daugherty, S.C., Angiuoli, S.V., 2011. *Neisseria meningitidis*is structured in clades associated with restriction modification systems that modulate homologous recombination. Proceedings of the National Academy of Sciences. 108, 4494-4499.

Bui, M. and Liu, Z., 2009. Simple allele-discriminating PCR for cost-effective and rapid genotyping and mapping. Plant Methods. 5, 1.

Buisine, N., Tang, C.M., Chalmers, R., 2002. Transposon-like Correia elements: structure, distribution and genetic exchange between pathogenic Neisseria sp. FEBS Letters. 522, 52-58.

Cahoon, L.A. and Seifert, H.S., 2011. Focusing homologous recombination: pilin antigenic variation in the pathogenic Neisseria. Molecular Microbiology. 81, 1136-1143.

Callaghan, M.J., Jolley, K.A., Maiden, M.C., 2006. Opacity-associated adhesin repertoire in hyperinvasive *Neisseria meningitidis*. Infection and Immunity. 74, 5085-5094.

Capecchi, B., Adu-Bobie, J., Di Marcello, F., Ciucchi, L., Masignani, V., Taddei, A., Rappuoli, R., Pizza, M., Aricò, B., 2005. *Neisseria meningitidis NadA* is a new invasin which promotes bacterial adhesion to and penetration into human epithelial cells. Molecular Microbiology. 55, 687-698.

Carbonnelle, E., Hlaine, S., Prouvensier, L., Nassif, X., Pelicic, V., 2005. Type IV pilus biogenesis in *Neisseria meningitidis*: *PilW* is involved in a step occurring after pilus assembly, essential for fibre stability and function. Molecular Microbiology. 55, 54-64.

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M., Barrell, B.G., Parkhill, J., 2005. ACT: the Artemis comparison tool. Bioinformatics. 21, 3422-3423.

Castillo-Ramirez, S., Harris, S.R., Holden, M.T., He, M., Parkhill, J., Bentley, S.D., Feil, E.J., 2011. The impact of recombination on dN/dS within recently emerged bacterial clones. PLoS Pathog. 7, e1002129.

Caugant, D.A., 2008. Genetics and evolution of Neisseria meningitidis: importance for the epidemiology of meningococcal disease. Infection, Genetics and Evolution. 8, 558-565.

Caugant, D.A., Hiby, E.A., Magnus, P., Scheel, O., Hoel, T., Bjune, G., Wedege, E., Eng, J., Frholm, L.O., 1994. Asymptomatic carriage of *Neisseria meningitidis* in a randomly sampled population. Journal of Clinical Microbiology. 32, 323-330.

Caugant, D.A. and Maiden, M.C., 2009. Meningococcal carriage and disease—population biology and evolution. Vaccine. 27, B70.

Cehovin, A., Simpson, P.J., McDowell, M.A., Brown, D.R., Noschese, R., Pallett, M., Brady, J., Baldwin, G.S., Lea, S.M., Matthews, S.J., 2013. Specific DNA recognition mediated by a type IV pilin. Proceedings of the National Academy of Sciences. 110, 3065-3070.

Che, D., Hasan, M.S., Chen, B., 2014. Identifying pathogenicity islands in bacterial pathogenomics using computational approaches. Pathogens. 3, 36-56.

Chen, I. & Gotschlich, E.C., 2001. ComE, a competence protein from *Neisseria gonorrhoeae* with DNA-binding activity. Journal of Bacteriology. 183, 3160-3168.

Cheng, K.C., Cahill, D.S., Kasai, H., Nishimura, S., Loeb, L.A., 1992. 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G----T and A----C substitutions. Journal of Biological Chemistry. 267, 166-172.

Chopra, I., O'Neill, A.J., Miller, K., 2003. The role of mutators in the emergence of antibiotic-resistant bacteria. Drug Resistance Updates. 6, 137-145.

Chow, J., Uadiale, K., Bestman, A., Kamau, C., Caugant, D.A., Shehu, A., Greig, J., 2016. Invasive Meningococcal Meningitis Serogroup C Outbreak in Northwest Nigeria, 2015-Third Consecutive Outbreak of a New Strain. PLoS Currents, 8 .

Claus, H., Elias, J., Meinhardt, C., Frosch, M., Vogel, U., 2007. Deletion of the meningococcal *fetA* gene used for antigen sequence typing of invasive and commensal isolates from Germany: frequencies and mechanisms. Journal of Clinical Microbiology. 45, 2960-2964.

Claus, H., Maiden, M.C., Wilson, D.J., McCarthy, N.D., Jolley, K.A., Urwin, R., Hessler, F., Frosch, M., Vogel, U., 2005. Genetic analysis of meningococci carried by children and young adults. Journal of Infectious Diseases. 191, 1263-1271.

Cloward, J.M. and Shafer, W.M., 2013. MtrR control of a transcriptional regulatory pathway in *Neisseria meningitidis* that influences expression of a gene (*nadA*) encoding a vaccine candidate. PloS One. 8, e56097.

Cohn, A.C. and Harrison, L.H., 2013. Meningococcal vaccines: current issues and future strategies. Drugs. 73, 1147-1155.

Comanducci, M., Bambini, S., Brunelli, B., Adu-Bobie, J., Aric, B., Capecchi, B., Giuliani, M.M., Masignani, V., Santini, L., Savino, S., 2002. *NadA*, a novel vaccine candidate of *Neisseria meningitidis*. Journal of Experimental Medicine. 195, 1445-1454.

Cook, G.C., 1992. Royal Society of Tropical Medicine and Hygiene Meeting at Manson House, London, 10 December 1992.George Carmichael Low FRCP: Twelfth President of the Society and Underrated Pioneer of Tropical Medicine. 355-360.

Corander, J., Connor, T.R., O'Dwyer, C.A., Kroll, J.S., Hanage, W.P., 2011. Population structure in the Neisseria, and the biological significance of fuzzy species. Journal of the Royal Society Interface. rsif20110601.

Criss, A.K., Bonney, K.M., Chang, R.A., Duffin, P.M., LeCuyer, B.E., Seifert, H.S., 2010. Mismatch correction modulates mutation frequency and pilus phase and antigenic variation in *Neisseria gonorrhoeae*. Journal of Bacteriology. 192, 316-325.

Croucher, N.J., Vernikos, G.S., Parkhill, J., Bentley, S.D., 2011. Identification, variation and transcription of pneumococcal repeat sequences. BMC Genomics. 12, 120.

Crum-Cianflone, N. and Sullivan, E., 2016. Meningococcal vaccinations. Infectious Diseases and Therapy. 5, 89-112.

Davidsen, T., Tuven, H.K., Bjrs, M., Rdland, E.A., Tnjum, T., 2007. Genetic interactions of DNA repair pathways in the pathogen *Neisseria meningitidis*. Journal of Bacteriology. 189, 5728-5737.

Davies, J.K., Harrison, P.F., Lin, Y., Bartley, S., Khoo, C.A., Seemann, T., Ryan, C.S., Kahler, C.M., Hill, S.A., 2014. The use of high-throughput DNA sequencing in the investigation of antigenic variation: application to Neisseria species. PloS One. 9, e86704.

Del Tordello, E., Bottini, S., Muzzi, A., Serruto, D., 2012. Analysis of the regulated transcriptome of *Neisseria meningitidis* in human blood using a tiling array. Journal of Bacteriology. 194, 6217-6232.

Del Tordello, E., Vacca, I., Ram, S., Rappuoli, R., Serruto, D., 2014. *Neisseria meningitides NalP* cleaves human complement C3, facilitating degradation of C3b and survival in human serum. Proceedings of the National Academy of Sciences. 111, 427-432.

DeLano, W.L., 2002. Pymol: An open-source molecular graphics tool. CCP4 Newsletter On Protein Crystallography, 40, pp.82-92.

Delany, I., Rappuoli, R., Scarlato, V., 2004. Fur functions as an activator and as a repressor of putative virulence genes in *Neisseria meningitidis*. Molecular Microbiology. 52, 1081-1090.

Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O., Salzberg, S.L., 1999. Alignment of whole genomes. Nucleic Acids Research. 27, 2369-2376.

Denamur, E., Lecointre, G., Darlu, P., Tenaillon, O., Acquaviva, C., Sayada, C., Sunjevaric, I., Rothstein, R., Elion, J., Taddei, F., 2000. Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. Cell. 103, 711-721.

Denamur, E. and Matic, I., 2006. Evolution of mutation rates in bacteria. Molecular Microbiology. 60, 820-827.

Didelot, X., Urwin, R., Maiden, M.C., Falush, D., 2009. Genealogical typing of *Neisseria meningitidis*. Microbiology. 155, 3176-3186.

Didelot, X. and Wilson, D.J., 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 11, e1004041.

Doran, K.S., Fulde, M., Gratz, N., Kim, B.J., Nau, R., Prasadarao, N., Schubert-Unkmeir, A., Tuomanen, E.I., Valentin-Weigand, P., 2016. Host–pathogen interactions in bacterial meningitis. Acta Neuropathologica. 131, 185-209.

Draskovic, I. & Dubnau, D., 2005. Biogenesis of a putative channel protein, ComEC, required for DNA uptake: membrane topology, oligomerization and formation of disulphide bonds. Molecular Microbiology. 55, 881-896.

Duffin, P.M. and Seifert, H.S., 2010. DNA uptake sequence-mediated enhancement of transformation in *Neisseria gonorrhoeae* is strain dependent. Journal of Bacteriology. 192, 4436-4444.

Dyet, K.H. & Martin, D.R., 2005. Sequence variation in the *porB* gene from B: P1. 4 meningococci causing New Zealand's epidemic. Journal of Clinical Microbiology. 43, 838-842.

Echols, H., 1990. Nucleoprotein structures initiating DNA replication, transcription, and site-specific recombination. J.Biol.Chem. 265, 14697-14700.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 32, 1792-1797.

Enrquez, R., Abad, R., Chanto, G., Corso, A., Cruces, R., Gabastou, J.M., Gorla, M.C., Maldonado, A., Moreno, J., Muros-Le Rouzic, E., 2010. Deletion of the Correia element in the mtr gene complex of *Neisseria meningitidis*. Journal of Medical Microbiology. 59, 1055-1060.

Enríquez, R., Abad, R., Salcedo, C., Pérez, S. and Vázquez, J.A., 2007. Fluoroquinolone resistance in *Neisseria meningitidis* in Spain. Journal of antimicrobial chemotherapy. 61, 286-290.

Etherington, G.J., Dicks, J., Roberts, I.N., 2005. Recombination Analysis Tool (RAT): a program for the high-throughput detection of recombination. Bioinformatics. 21, 278-281.

Fagnocchi, L., Biolchi, A., Ferlicca, F., Boccadifuoco, G., Brunelli, B., Brier, S., Norais, N., Chiarot, E., Bensi, G., Kroll, J.S., 2013. Transcriptional regulation of the nadA gene in *Neisseria meningitides* impacts the prediction of coverage of a multicomponent meningococcal serogroup B vaccine. Infection and Immunity. 81, 560-569.

Fan, X., Li, Y., He, R., Li, Q., He, W., 2016. Comparative analysis of prophage-like elements in Helicobacter sp. genomes. PeerJ. 4, e2012.

Fàbrega, A., Madurga, S., Giralt, E. and Vila, J., 2009. Mechanism of action of and resistance to quinolones. Microbial biotechnology. 2, 40-61.

Feavers, I.M., Heath, A.B., Bygraves, J.A., Maiden, M., 1992. Role of horizontal genetic exchange in the antigenic variation of the class 1 outer membrane protein of *Neisseria meningitidis*. Molecular Microbiology. 6, 489-495.

Feil, E., Carpenter, G., Spratt, B.G., 1995. Electrophoretic variation in adenylate kinase of *Neisseria meningitidisis* due to inter-and intraspecies recombination. Proceedings of the National Academy of Sciences. 92, 10535-10539.

Feil, E.J., Li, B.C., Aanensen, D.M., Hanage, W.P. and Spratt, B.G., 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. Journal of bacteriology. 186, 1518-1530.

Feil, E.J., Holmes, E.C., Bessen, D.E., Chan, M., Day, N.P., Enright, M.C., Goldstein, R., Hood, D.W., Kalia, A., Moore, C.E., 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proceedings of the National Academy of Sciences. 98, 182-187.

Feil, E.J., Maiden, M.C., Achtman, M., Spratt, B.G., 1999. The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. Molecular Biology and Evolution. 16, 1496-1502.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 783-791.

Fijalkowska, I.J., Schaaper, R.M., Jonczyk, P., 2012. DNA replication fidelity in *Escherichia coli*: a multi-DNA polymerase affair. FEMS Microbiology Reviews. 36, 1105-1121.

Fournier, P., Dubourg, G., Raoult, D., 2014. Clinical detection and characterization of bacterial pathogens in the genomics era. Genome Medicine. 6, 114.

Fransen, F., Heckenberg, S.G., Hamstra, H.J., Feller, M., Boog, C.J., van Putten, J.P., van de Beek, D., van der Ende, A., van der Ley, P., 2009. Naturally occurring lipid A mutants in *Neisseria meningitides* from patients with invasive meningococcal disease are associated with reduced coagulopathy. PLoS Pathog. 5, e1000396.

Freudenthal, B.D., Beard, W.A., Cuneo, M.J., Dyrkheeva, N.S., Wilson, S.H., 2015. Capturing snapshots of APE1 processing DNA damage. Nature Structural and Molecular Biology.

Frye, S.A., Nilsen, M., Tønjum, T. and Ambur, O.H., 2013. Dialects of the DNA uptake sequence in Neisseriaceae. PLoS Genet, 9, e1003458

Gabutti, G., Stefanati, A., Kuhdari, P., 2015. Epidemiology of *Neisseria meningitides* infections: case distribution by age and relevance of carriage. Journal of Preventive Medicine and Hygiene. 56, E116.

Gally, D. L., Rucker, T. J. & Blomfield, I. C. 1994. The leucine-responsive regulatory protein binds to the fim switch to control phase variation of type 1 fimbrial expression in *E. coli* K-12. J Bacteriol, 176, 5665-72.

Gasparini, R., Panatto, D., Bragazzi, N.L., Lai, P.L., Bechini, A., Levi, M., Durando, P., Amicizia, D., 2015. How the knowledge of interactions between meningococcus and the

human immune system has been used to prepare effective *Neisseria meningitides* vaccines. Journal of Immunology Research. 2015, .

Gasparini, R. and Panatto, D., 2011. Meningococcal glycoconjugate vaccines. Human Vaccines. 7, 170-182.

Gault, J., Ferber, M., Machata, S., Imhaus, A., Malosse, C., Charles-Orszag, A., Millien, C., Bouvier, G., Bardiaux, B., Phau-Arnaudet, G., 2015. *Neisseria meningitides* type IV pili composed of sequence invariable pilins are masked by multisite glycosylation. PLoS Pathog. 11, e1005162.

Geoffroy, M., Floquet, S., Mtais, A., Nassif, X., Pelicic, V., 2003. Large-scale analysis of the meningococcus genome by gene disruption: resistance to complement-mediated lysis. Genome Research. 13, 391-398.

Giltner, C.L., Nguyen, Y., Burrows, L.L., 2012. Type IV pilin proteins: versatile molecular modules. Microbiology and Molecular Biology Reviews. 76, 740-772.

Gioia, C.A.C., de Lemos, Ana P Silva, Gorla, M.C.O., Mendoza-Sassi, R.A., Ballester, T., Von Groll, A., Wedig, B., de Vargas Ethur, N., Bragana, L., Milagres, L.G., 2015. Detection of *Neisseria meningitides* in asymptomatic carriers in a university hospital from Brazil. Revista Argentina De Microbiologa. 47, 322-327.

Givan, K.F., Thomas, B.W., Johnston, A.G., 1977. Isolation of *Neisseria meningitides* from the urethra, cervix, and anal canal: further observations. The British Journal of Venereal Diseases. 53, 109-112.

Gordon, L., Chervonenkis, A.Y., Gammerman, A.J., Shahmuradov, I.A., Solovyev, V.V., 2003. Sequence alignment kernel for recognition of promoter regions. Bioinformatics. 19, 1964-1971.

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution. 27, 221-224.

Grindley, N.D., 1997. Site-specific recombination: synapsis and strand exchange revealed. Current Biology. 7, R612.

Grindley, N.D., Whiteson, K.L., Rice, P.A., 2006. Mechanisms of site-specific recombination. Annu.Rev.Biochem. 75, 567-605.

Guindon, S., Dufayard, J., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology. 59, 307-321.

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 29, 1072-1075.

Haas, R. and Meyer, T.F., 1986. The repertoire of silent pilus genes in *Neisseria gonorrhoeae*: evidence for gene conversion. Cell. 44, 107-115.

Hallet, B. and Sherratt, D.J., 1997. Transposition and site-specific recombination: adapting DNA cut-and-paste mechanisms to a variety of genetic rearrangements. FEMS Microbiology Reviews. 21, 157-178.

Hamilton, H.L. and Dillard, J.P., 2006. Natural transformation of *Neisseria gonorrhoeae*: from DNA donation to homologous recombination. Molecular Microbiology. 59, 376-385.

Hamoen, L.W., Venema, G., Kuipers, O.P., 2003. Controlling competence in *Bacillus subtilis*: shared use of regulators. Microbiology. 149, 9-17.

Hansen, A., Gu, Y., Li, M., Andrykovitch, M., Waugh, D.S., Jin, D.J., Ji, X., 2005. Structural basis for the function of stringent starvation protein a as a transcription factor. Journal of Biological Chemistry. 280, 17380-17391.

Hao, W., 2013. Extensive genomic variation within clonal bacterial groups resulted from homologous recombination. Mobile Genetic Elements. 3, e33971.

Harcourt, B.H., et al, 2015, Population-based surveillance of *Neisseria meningitides* antimicrobial resistance in the United States, Open forum infectious diseases, 2015, Oxford University Press ppofv117.

Hardy, S.J., Christodoulides, M., Weller, R.O., Heckels, J.E., 2000. Interactions of *Neisseria meningitides* with cells of the human meninges. Molecular Microbiology. 36, 817-829.

Harrison, L.H., 2010. Epidemiological profile of meningococcal disease in the United States. Clinical Infectious Diseases. 50, S44.

Harrison, L.H., Trotter, C.L., Ramsay, M.E., 2009. Global epidemiology of meningococcal disease. Vaccine. 27, B63.

Harrison, O.B., Bennett, J.S., Derrick, J.P., Maiden, M.C., Bayliss, C.D., 2013. Distribution and diversity of the haemoglobin–haptoglobin iron-acquisition systems in pathogenic and non-pathogenic Neisseria. Microbiology. 159, 1920-1930.

Hedberg, S.T., Fredlund, H., Nicolas, P., Caugant, D.A., Olcn, P., Unemo, M., 2009. Antibiotic susceptibility and characteristics of *Neisseria meningitides* isolates from the African meningitis belt, 2000 to 2006: phenotypic and genotypic perspectives. Antimicrobial Agents and Chemotherapy. 53, 1561-1566.

Hill, D.J., Griffiths, N.J., Borodina, E., Virji, M., 2010. Cellular and molecular biology of *Neisseria meningitidis* colonization and invasive disease. Clinical Science. 118, 547-564.

Hill, S.A. and Davies, J.K., 2009. Pilin gene variation in *Neisseria gonorrhoeae*: reassessing the old paradigms. FEMS Microbiology Reviews. 33, 521-530.

Hobbs, M.M., Malorny, B., Prasad, P., Morelli, G., Kusecek, B., Heckels, J.E., Cannon, J.G., Achtman, M., 1998. Recombinational reassortment among *opa* genes from ET-37 complex *Neisseria meningitidis* isolates of diverse geographical origins. Microbiology. 144, 157-166.

Holmes, E.C., Urwin, R., Maiden, M.C., 1999. The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. Molecular Biology and Evolution. 16, 741-749.

Hong, E., Hedberg, S.T., Abad, R., Fazio, C., Enríquez, R., Deghmane, A.E., Jolley, K.A., Stefanelli, P., Unemo, M., Vazquez, J.A. and Veyrier, F.J., 2013. Target gene sequencing to

define the susceptibility of *Neisseria meningitidis* to ciprofloxacin. Antimicrobial agents and chemotherapy. 57, 1961-1964.

Hotopp, J.C.D., Grifantini, R., Kumar, N., Tzeng, Y.L., Fouts, D., Frigimelica, E., Draghi, M., Giuliani, M.M., Rappuoli, R., Stephens, D.S., 2006. Comparative genomics of *Neisseria meningitidis*: core genome, islands of horizontal transfer and pathogen-specific genes. Microbiology. 152, 3733-3749.

Huber, C., 2011. Genetic Diversity and Immune Evasion of Bacterial Pathogens. (Doctoral dissertation, University_of_Basel).

Hubert, K., Pawlik, M., Claus, H., Jarva, H., Meri, S., Vogel, U., 2012. *Opc* expression, LPS immunotype switch and pilin conversion contribute to serum resistance of unencapsulated meningococci. PLos One. 7, e45132.

Hughes, A.L. and Friedman, R., 2004. Patterns of sequence divergence in 5′ intergenic spacers and linked coding regions in 10 species of pathogenic bacteria reveal distinct recombinational histories. Genetics. 168, 1795-1803.

Huson, D.H. and Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution. 23, 254-267.

Jayaraman, R., 2009. Mutators and hypermutability in bacteria: the *Escherichia coli* paradigm. Journal of Genetics. 88, 379-391.

John, C.M., Phillips, N.J., Din, R., Liu, M., Rosenqvist, E., Hiby, E.A., Stein, D.C., Jarvis, G.A., 2016. Lipooligosaccharide structures of invasive and carrier isolates of *Neisseria meningitidis* are correlated with pathogenicity and carriage. Journal of Biological Chemistry. 291, 3224-3238.

Johnston, C., Martin, B., Fichant, G., Polard, P., Claverys, J., 2014. Bacterial transformation: distribution, shared mechanisms and divergent control. Nature Reviews Microbiology. 12, 181-196.

Jolivet-Gougeon, A., Kovacs, B., Le Gall-David, S., Le Bars, H., Bousarghin, L., Bonnaure-Mallet, M., Lobel, B., Guill, F., Soussy, C., Tenke, P., 2011. Bacterial hypermutation: clinical implications. Journal of Medical Microbiology. 60, 563-573.

Jolley, K.A. and Maiden, M.C., 2010. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 11, 595.

Jones, C.H., Mohamed, N., Rojas, E., Andrew, L., Hoyos, J., Hawkins, J.C., McNeil, L.K., Jiang, Q., Mayer, L.W., Wang, X., 2016. Comparison of phenotypic and genotypic approaches to capsule typing of *Neisseria meningitidis* by use of invasive and carriage isolate collections. Journal of Clinical Microbiology. 54, 25-34.

Jones, D.M., Borrow, R., Fox, A.J., Gray, S., Cartwright, K.A., Poolman, J.T., 1992. The lipooligosaccharide immunotype as a virulence determinant in *Neisseria meningitidis*. Microbial Pathogenesis. 13, 219-224.

Jordan, P.W. and Saunders, N.J., 2009. Host iron binding proteins acting as niche indicators for *Neisseria meningitidis*. PLoS One. 4, e5198.

Joseph, B., Schwarz, R.F., Linke, B., Blom, J., Becker, A., Claus, H., Goesmann, A., Frosch, M., Mller, T., Vogel, U., 2011. Virulence evolution of the human pathogen *Neisseria meningitidis* by recombination in the core and accessory genome. PLoS One. 6, e18441.

Kahler, C.M., Blum, E., Miller, Y.K., Ryan, D., Popovic, T., Stephens, D.S., 2001. exl, an exchangeable genetic island in *Neisseria meningitidis*. Infection and Immunity. 69, 1687-1696.

Kanehisa, M., Sato, Y. and Morishima, K., 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. Journal of molecular biology, 428, 726-731.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2015. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research. gkv1070.

Katoh, K. and Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution. 30, 772-780.

Kattner, C., Toussi, D.N., Zaucha, J., Wetzler, L.M., Rppel, N., Zachariae, U., Massari, P., Tanabe, M., 2014. Crystallographic analysis of *Neisseria meningitidis* PorB extracellular loops potentially implicated in TLR2 recognition. Journal of Structural Biology. 185, 440-447.

Khairalla, A.S., Omer, S.A., Mahdavi, J., Aslam, A., Dufailu, O.A., Self, T., Jonsson, A., Geörg, M., Sjölinder, H., Royer, P., 2015. Nuclear trafficking, histone cleavage and induction of apoptosis by the meningococcal *App* and *MspA* autotransporters. Cellular Microbiology. 17, 1008-1020.

Klauser, T., Pohlner, J., Meyer, T.F., 1993. The secretion pathway of IgA protease-type proteins in gram-negative bacteria. Bioessays. 15, 799-805.

Klee, S.R., Nassif, X., Kusecek, B., Merker, P., Beretti, J., Achtman, M., Tinsley, C.R., 2000. Molecular and biological analysis of eight genetic islands that distinguish *Neisseria meningitidis* from the closely related pathogen *Neisseria gonorrhoeae*. Infection and Immunity. 68, 2082-2095.

Klimenko, A.I., Matushkin, Y.G., Kolchanov, N.A., Lashin, S.A., 2016. Bacteriophages affect evolution of bacterial communities in spatially distributed habitats: a simulation study. BMC Microbiology. 16, S10.

Klughammer, J., Dittrich, M., Blom, J., Mitesser, V., Vogel, U., Frosch, M., Goesmann, A., Mller, T., Schoen, C., 2017. Comparative Genome Sequencing Reveals Within-Host Genetic Changes in *Neisseria meningitidis* during Invasive Disease. Plos One. 12, e0169892.

Kong, Y., Ma, J.H., Warren, K., Tsang, R.S., Low, D.E., Jamieson, F.B., Alexander, D.C., Hao, W., 2013. Homologous recombination drives both sequence diversity and gene content variation in *Neisseria meningitidis*. Genome Biology and Evolution. 5, 1611-1627.

Koonin, E.V., 2015. The turbulent network dynamics of microbial evolution and the statistical tree of life. Journal of Molecular Evolution. 80, 244-250.

Korber B. (2000). HIV Signature and Sequence Variation Analysis. Computational Analysis of HIV Molecular Sequences, Chapter 4, pages 55-72. Allen G. Rodrigo and Gerald H. Learn, eds. Dordrecht, Netherlands: Kluwer Academic Puublishers.

Kosakovsky Pond, S.L., Poon, A.F.Y.S. and Frost, D.W., 2009. The Phylogenetic Handbook. A Practical Approach to Phylogenetic Analysis and Hypothesis Testing, chapter Estimating selection pressures on alignments of coding sequences. Number, 14, pp.2375-85.

Krauland, M.G., Hotopp, J.C.D., Riley, D.R., Daugherty, S.C., Marsh, J.W., Messonnier, N.E., Mayer, L.W., Tettelin, H., Harrison, L.H., 2012. Whole genome sequencing to investigate the emergence of clonal complex 23 *Neisseria meningitidis* serogroup Y disease in the United States. PloS One. 7, e35699.

Ladhani, S.N., Flood, J.S., Ramsay, M.E., Campbell, H., Gray, S.J., Kaczmarski, E.B., Mallard, R.H., Guiver, M., Newbold, L.S., Borrow, R., 2012. Invasive meningococcal disease in England and Wales: implications for the introduction of new vaccines. Vaccine. 30, 3710-3716.

Ladhani, S.N., Giuliani, M.M., Biolchi, A., Pizza, M., Beebeejaun, K., Lucidarme, J., Findlow, J., Ramsay, M.E., Borrow, R., 2016. Effectiveness of meningococcal B vaccine against endemic hypervirulent *Neisseria meningitidis* W strain, England. Emerging Infectious Diseases. 22, 309.

Lamelas, A., Harris, S.R., Rltgen, K., Dangy, J., Hauser, J., Kingsley, R.A., Connor, T.R., Sie, A., Hodgson, A., Dougan, G., 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, 1974.

Langille, M.G. and Brinkman, F.S., 2009. Bioinformatic detection of horizontally transferred DNA in bacterial genomes. F1000 Biol.Rep. 1, 25.

Lavezzo, E., Toppo, S., Franchin, E., Di Camillo, B., Finotello, F., Falda, M., Manganelli, R., Pal, G., Barzon, L., 2013. Genomic comparative analysis and gene function prediction in infectious diseases: application to the investigation of a meningitis outbreak. BMC Infectious Diseases. 13, 554.

Lee, J. and Helmann, J.D., 2007. Functional specialization within the Fur family of metalloregulators. Biometals. 20, 485-499.

Lewis-Rogers, N., Crandall, K.A., Posada, D., 2004. Evolutionary analyses of genetic recombination. Dynamical Genetics. 408, .

Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 25, 1754-1760.

Librado, P. and Rozas, J., 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 25, 1451-1452.

Lin, Y., Ryan, C.S., 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, 5728-5736.

Linz, B., Schenker, M., Zhu, P., Achtman, M., 2000. Frequent interspecific genetic exchange between commensal Neisseriae and *Neisseria meningitidis*. Molecular Microbiology. 36, 1049-1058.

Litschko, C., Romano, M.R., Pinto, V., Claus, H., Vogel, U., Berti, F., Gerardy-Schahn, R., Fiebig, T., 2015. The capsule polymerase CslB of *Neisseria meningitidis* serogroup L catalyzes the synthesis of a complex trimeric repeating unit comprising glycosidic and phosphodiester linkages. Journal of Biological Chemistry. 290, 24355-24366.

Liu, S.V., Saunders, N.J., Jeffries, A., Rest, R.F., 2002. Genome analysis and strain comparison of correia repeats and correia repeat-enclosed elements in pathogenic Neisseria. Journal of Bacteriology. 184, 6163-6173.

Liu, Y., Zhang, D., Engstrm, k., Mernyi, G., Hagner, M., Yang, H., Kuwae, A., Wan, Y., Sjlinder, M., Sjlinder, H., 2016. Dynamic niche-specific adaptations in *Neisseria meningitidis* during infection. Microbes and Infection. 18, 109-117.

Loeb, L.A. and Preston, B.D., 1986. Mutagenesis by apurinic/apyrimidinic sites. Annual Review of Genetics. 20, 201-230.

Lovett, S.T., 2004. Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. Molecular Microbiology. 52, 1243-1253.

Lovett, S.T., 2011. The DNA exonucleases of escherichia coli. EcoSal Plus, 4.

Lucidarme, J., Findlow, J., Chan, H., Feavers, I.M., Gray, S.J., Kaczmarski, E.B., Parkhill, J., Bai, X., Borrow, R., Bayliss, C.D., 2013. The distribution and 'in vivo'phase variation status of haemoglobin receptors in invasive meningococcal serogroup B disease: genotypic and phenotypic analysis. PloS One. 8, e76932.

Macfadyen, L.P., Chen, D., Vo, H.C., Liao, D., Sinotte, R., Redfield, R.J., 2001. Competence development by *Haemophilus influenzae* is regulated by the availability of nucleic acid precursors. Molecular Microbiology. 40, 700-707.

MacLennan, J., Kafatos, G., Neal, K., Andrews, N., Cameron, J.C., Roberts, R., Evans, M.R., Cann, K., Baxter, D.N., Maiden, M.C., 2006. Social behavior and meningococcal carriage in British teenagers. Emerging Infectious Diseases. 12, 950-957.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., 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, 3140-3145.

Malito, E., Biancucci, M., Faleri, A., Ferlenghi, I., Scarselli, M., Maruggi, G., Surdo, P.L., Veggi, D., Liguori, A., Santini, L., 2014. Structure of the meningococcal vaccine antigen *NadA* and epitope mapping of a bactericidal antibody. Proceedings of the National Academy of Sciences. 111, 17128-17133.

Manchanda, V., Gupta, S., Bhalla, P., 2006. Meningococcal disease: history, epidemiology, pathogenesis, clinical manifestations, diagnosis, antimicrobial susceptibility and prevention. Indian Journal of Medical Microbiology. 24, 7.

Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., 2014. CDD: NCBI's conserved domain database. Nucleic Acids Research. gku1221.

Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Research. 39, D229.

Marri, P.R., Paniscus, M., Weyand, N.J., Rendn, M.A., Calton, C.M., Hernndez, D.R., Higashi, D.L., Sodergren, E., Weinstock, G.M., Rounsley, S.D., 2010. Genome sequencing reveals widespread virulence gene exchange among human Neisseria species. PloS One. 5, e11835.

Marsay, L., Dold, C., Green, C.A., Rollier, C.S., Norheim, G., Sadarangani, M., Shanyinde, M., Brehony, C., Thompson, A.J., Sanders, H., 2015. A novel meningococcal outer membrane vesicle vaccine with constitutive expression of *FetA*: A phase I clinical trial. Journal of Infection. 71, 326-337.

Marsh, J.W., O'Leary, M.M., Shutt, K.A., Harrison, L.H., 2007. Deletion of *fetA* gene sequences in serogroup B and C *Neisseria meningitidis* isolates. Journal of Clinical Microbiology. 45, 1333-1335.

Martin, D.P., Murrell, B., Golden, M., Khoosal, A. and Muhire, B., 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evolution. 1.

Martin, P., Makepeace, K., Hill, S.A., Hood, D.W., Moxon, E.R., 2005. Microsatellite instability regulates transcription factor binding and gene expression. Proceedings of the National Academy of Sciences of the United States of America. 102, 3800-3804.

Matschiner, M., 2016. Fitchi: haplotype genealogy graphs based on the Fitch algorithm. Bioinformatics. 32, 1250-1252.

Matus-Garcia, M., Nijveen, H., van Passel, M.W., 2012. Promoter propagation in prokaryotes. Nucleic Acids Research. gks787.

Maydt, J. and Lengauer, T., 2006. Recco: recombination analysis using cost optimization. Bioinformatics. 22, 1064-1071.

McGill, F., Heyderman, R.S., Michael, B.D., Defres, S., Beeching, N.J., Borrow, R., Glennie, L., Gaillemin, O., Wyncoll, D., Kaczmarski, E., 2016. The UK joint specialist societies guideline on the diagnosis and management of acute meningitis and meningococcal sepsis in immunocompetent adults. Journal of Infection. 72, 405-438.

Mechergui, A., Achour, W., Hassen, A.B., 2015. Principles and applications of typing methods for commensal Neisseria. Reviews in Medical Microbiology. 26, 47-52.

Mendelman, P.M., Campos, J., Chaffin, D.O., Serfass, D.A., Smith, A.L., Saez-Nieto, J.A., 1988. Relative penicillin G resistance in *Neisseria meningitidis* and reduced affinity of penicillin-binding protein 3. Antimicrobial Agents and Chemotherapy. 32, 706-709.

Merz, A.J., Enns, C.A., So, M., 1999. Type IV pili of pathogenic Neisseriae elicit cortical plaque formation in epithelial cells. Molecular Microbiology. 32, 1316-1332.

Merz, A.J., Rifenbery, D.B., Arvidson, C.G., So, M., 1996. Traversal of a polarized epithelium by pathogenic Neisseriae: facilitation by type IV pili and maintenance of epithelial barrier function. Molecular Medicine. 2, 745.

Metruccio, M.M., Pigozzi, E., Roncarati, D., Scorza, F.B., Norais, N., Hill, S.A., Scarlato, V., Delany, I., 2009. A novel phase variation mechanism in the meningococcus driven by a ligand-responsive repressor and differential spacing of distal promoter elements. PLoS Pathog. 5, e1000710.

Meyers, L.A., Levin, B.R., Richardson, A.R., Stojiljkovic, I., 2003. Epidemiology, hypermutation, within–host evolution and the virulence of *Neisseria meningitidis*. Proceedings of the Royal Society of London B: Biological Sciences. 270, 1667-1677.

Milenović, Ž, 2011. Application of Mann-Whitney U test in research of professional training of primary school teachers. Metodički Obzori. 6, 73-79.

Mller, M.G., Ing, J.Y., Cheng, M.K., Flitter, B.A., Moe, G.R., 2013. Identification of a phage-encoded Ig-binding protein from invasive *Neisseria meningitidis*. The Journal of Immunology. 191, 3287-3296.

Molesworth, A.M., Thomson, M.C., Connor, S.J., Cresswell, M.P., Morse, A.P., Shears, P., Hart, C.A., Cuevas, L.E., 2002. Where is the meningitis belt? Defining an area at risk of epidemic meningitis in Africa. Transactions of the Royal Society of Tropical Medicine and Hygiene. 96, 242-249.

Morelli, G., Malorny, B., Müller, K., Seiler, A., Wang, J., Del Valle, J., Achtman, M., 1997. Clonal descent and microevolution of *Neisseria meningitidis* during 30 years of epidemic spread. Molecular Microbiology. 25, 1047-1064.

Morgenthau, A., Beddek, A., Schryvers, A.B., 2014. The negatively charged regions of lactoferrin binding protein B, an adaptation against anti-microbial peptides. PLOS One. 9, e86243.

Moxon, R., Bayliss, C., Hood, D., 2006. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. Annu.Rev.Genet. 40, 307-333.

Muzzi, A., Mora, M., Pizza, M., Rappuoli, R., Donati, C., 2013. Conservation of meningococcal antigens in the genus Neisseria. MBio. 4, 163.

Nadel, S., 2016. Treatment of meningococcal disease. Journal of Adolescent Health. 59, S28.

Nagorska, K., Silhan, J., Li, Y., Pelicic, V., Freemont, P.S., Baldwin, G.S., Tang, C.M., 2012. A network of enzymes involved in repair of oxidative DNA damage in *Neisseria meningitidis*. Molecular Microbiology. 83, 1064-1079.

Najafi, M.B.H. and Pezeshki, P., 2014. Bacterial mutation; Types, Mechanisms and mutant detection methods: a review. European Scientific Journal, ESJ. 9, .

Nash, H.A., 1996. Site-specific recombination: integration, excision, resolution, and inversion of defined DNA segments. *Escherichia Coli* and *Salmonella*: Cellular and Molecular Biology. 2, 2363-2376

Nassif, X., 2010. Interactions between encapsulated *Neisseria meningitidis* and host cells. International Microbiology. 2, 133-136.

Nei, M. and Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution. 3, 418-426.

Nei, M. and Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Niehus, R., Mitri, S., Fletcher, A.G., Foster, K.R., 2015. Migration and horizontal gene transfer divide microbial genomes into multiple niches. Nature Communications, 6.

Nocek, B., Mulligan, R., Bargassa, M., Collart, F., Joachimiak, A., 2008. Crystal structure of aminopeptidase N from human pathogen *Neisseria meningitidis*. Proteins: Structure, Function, and Bioinformatics. 70, 273-279.

Noinaj, N., Cornelissen, C.N., Buchanan, S.K., 2013. Structural insight into the lactoferrin receptors from pathogenic Neisseria. Journal of Structural Biology. 184, 83-92.

Norbury, C.J. and Hickson, I.D., 2001. Cellular responses to DNA damage. Annual Review of Pharmacology and Toxicology. 41, 367-401.

Norheim, G., Sanders, H., Mellesdal, J.W., Sundfr, I., Chan, H., Brehony, C., Vipond, C., Dold, C., Care, R., Saleem, M., 2015. An OMV Vaccine Derived from a Capsular Group B Meningococcus with Constitutive FetA Expression: Preclinical Evaluation of Immunogenicity and Toxicity. PloS One. 10, e0134353.

O'Ryan, M., Stoddard, J., Toneatto, D., Wassil, J., Dull, P.M., 2014. A multi-component meningococcal serogroup B vaccine (4CMenB): the clinical development program. Drugs. 74, 15-30.

Obergfell, K.P. and Seifert, H.S., 2015. Mobile DNA in the pathogenic Neisseria. Microbiology Spectrum. 3, .

Oldfield, N.J., Harrison, O.B., Bayliss, C.D., Maiden, M.C., Ala'Aldeen, D.A., Turner, D.P., 2016. Genomic Analysis of Serogroup Y *Neisseria meningitidis* Isolates Reveals Extensive Similarities Between Carriage-Associated and Disease-Associated Organisms. Journal of Infectious Diseases. 213, 1777-1785.

Oren, Y., Smith, M.B., Johns, N.I., Zeevi, M.K., Biran, D., Ron, E.Z., Corander, J., Wang, H.H., Alm, E.J., Pupko, T., 2014. Transfer of noncoding DNA drives regulatory rewiring in bacteria. Proceedings of the National Academy of Sciences. 111, 16112-16117.

Orihuela, C.J., Mahdavi, J., Thornton, J., Mann, B., Wooldridge, K.G., Abouseada, N., Oldfield, N.J., Self, T., Ala'Aldeen, D.A., Tuomanen, E.I., 2009. Laminin receptor initiates bacterial contact with the blood brain barrier in experimental meningitis models. The Journal of Clinical Investigation. 119, 1638-1646.

Ota, T. and Nei, M., 1994. Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. Molecular Biology and Evolution. 11, 469-482.

Pace, D. and Pollard, A.J., 2012. Meningococcal disease: clinical presentation and sequelae. Vaccine. 30, B9. Page, P.S., MOLECULAR EVENTS IN MICROBIAL PATHOGENESIS. J.Cell Sci. 105, 699-710.

Parkhill, J., Achtman, M., James, K.D., Bentley, S.D., Churcher, C., Klee, S.R., Morelli, G., Basham, D., Brown, D., Chillingworth, T., 2000. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature. 404, 502-506.

Parrow, N.L., Fleming, R.E., Minnick, M.F., 2013. Sequestration and scavenging of iron in infection. Infection and Immunity. 81, 3503-3514.

Peak, I.R., Jennings, C.D., Jen, F.E., Jennings, M.P., 2014. Role of *Neisseria meningitidis* *PorA* and *PorB* expression in antimicrobial susceptibility. Antimicrobial Agents and Chemotherapy. 58, 614-616.

Peng, J., Yang, L., Yang, F., Yang, J., Yan, Y., Nie, H., Zhang, X., Xiong, Z., Jiang, Y., Cheng, F., 2008. Characterization of ST-4821 complex, a unique *Neisseria meningitidis* clone. Genomics. 91, 78-87.

Perkins-Balding, D., Ratliff-Griffin, M., Stojiljkovic, I., 2004. Iron transport systems in *Neisseria meningitidis*. Microbiology and Molecular Biology Reviews. 68, 154-171.

Perrin, A., Bonacorsi, S., Carbonnelle, E., Talibi, D., Dessen, P., Nassif, X., Tinsley, C., 2002. Comparative genomics identifies the genetic islands that distinguish *Neisseria meningitidis*, the agent of cerebrospinal meningitis, from other Neisseria species. Infection and Immunity. 70, 7063-7072.

Persky, N.S. and Lovett, S.T., 2008. Mechanisms of recombination: lessons from E. coli. Critical Reviews in Biochemistry and Molecular Biology. 43, 347-370.

Picard, C., Casanova, J., Puel, A., 2011. Infectious diseases in patients with IRAK-4, MyD88, NEMO, or IκBα deficiency. Clinical Microbiology Reviews. 24, 490-497.

Putonti, C., Nowicki, B., Shaffer, M., Fofanov, Y., Nowicki, S., 2013. Where does Neisseria acquire foreign DNA from: an examination of the source of genomic and pathogenic islands and the evolution of the Neisseria genus. BMC Evolutionary Biology. 13, 184.

Ram, S. and Vogel, U., 2006. Role of complement in defense against meningococcal infection. Handbook of Meningococcal Disease: Infection Biology, Vaccination, Clinical Management. 273-293.

Sciences. 99, 6103-6107.

Richardson, A.R. and Stojiljkovic, I., 2001. Mismatch repair and the regulation of phase variation in *Neisseria meningitidis*. Molecular microbiology, 40(3), pp.645-655.

Richardson, A.R., Yu, Z., Popovic, T., Stojiljkovic, I., 2002. Mutator clones of *Neisseria meningitidis* in epidemic serogroup A disease. Proceedings of the National Academy of

Rishishwar, L., Katz, L.S., Sharma, N.V., Rowe, L., Frace, M., Thomas, J.D., Harcourt, B.H., Mayer, L.W., Jordan, I.K., 2012. Genomic basis of a polyagglutinating isolate of *Neisseria meningitidis*. Journal of Bacteriology. 194, 5649-5656.

Roberts, S.B., Spencer-Smith, R., Shah, M., Nebel, J., Cook, R.T., Snyder, L.A., 2016. Correia repeat enclosed elements and non-coding RNAs in the Neisseria species. Microorganisms. 4, 31.

Rogozin, I.B., Makarova, K.S., Natale, D.A., Spiridonov, A.N., Tatusov, R.L., Wolf, Y.I., Yin, J., Koonin, E.V., 2002. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. Nucleic Acids Research. 30, 4264-4271.

Rohde, K.H., Gillaspy, A.F., Hatfield, M.D., Lewis, L.A., Dyer, D.W., 2002. Interactions of haemoglobin with the *Neisseria meningitidis* receptor *HpuAB*: the role of *TonB* and an intact proton motive force. Molecular Microbiology. 43, 335-354.

Romero, J.D. and Outschoorn, I.M., 1994. Current status of meningococcal group B vaccine candidates: capsular or noncapsular? Clinical Microbiology Reviews. 7, 559-575.

Ross, T.D., 2003. Accurate confidence intervals for binomial proportion and Poisson rate estimation. Computers in Biology and Medicine. 33, 509-531.

Rouphael, N.G. and Stephens, D.S., 2012. *Neisseria meningitidis*: biology, microbiology, and epidemiology. *Neisseria Meningitidis*: Advanced Methods and Protocols. 1-20.

Rouquette-Loughlin, C.E., Balthazar, J.T., Hill, S.A., Shafer, W.M., 2004. Modulation of the mtrCDE-encoded efflux pump gene complex of *Neisseria meningitidis* due to a Correia element insertion sequence. Molecular Microbiology. 54, 731-741.

Ruiz-Perez, F. and Nataro, J.P., 2014. Bacterial serine proteases secreted by the autotransporter pathway: classification, specificity, and role in virulence. Cellular and Molecular Life Sciences. 71, 745-770.

Rusniok, C., Vallenet, D., Floquet, S., Ewles, H., Mouz-Soulama, C., Brown, D., Lajus, A., Buchrieser, C., Mdigue, C., Glaser, P., 2009. NeMeSys: a biological resource for narrowing

the gap between sequence and function in the human pathogen *Neisseria meningitidis*. Genome Biology. 10, R110.

Russell, J.E., 2004. *PorA* Variable Regions of *Neisseria meningitidis*-Volume 10, Number 4—April 2004-Emerging Infectious Disease journal-CDC.

Sabat, A.J., Budimir, A., Nashev, D., S-Leo, R., Van Dijl, J.M., Laurent, F., Grundmann, H., Friedrich, A.W., ESCMID Study Group of Epidemiological Markers (ESGEM), 2013. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro Surveill. 18, 20380.

Sadarangani, M., Pollard, A.J., Gray-Owen, S.D., 2011. *Opa* proteins and CEACAMs: pathways of immune engagement for pathogenic Neisseria. FEMS Microbiology Reviews. 35, 498-514.

Sadowski, P., 1986. Site-specific recombinases: changing partners and doing the twist. Journal of Bacteriology. 165, 341.

Salipante, S.J., SenGupta, D.J., Cummings, L.A., Land, T.A., Hoogestraat, D.R., Cookson, B.T., 2015. Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology. Journal of Clinical Microbiology. 53, 1072-1079.

Sanders, H., Norheim, G., Chan, H., Dold, C., Vipond, C., Derrick, J.P., Pollard, A.J., Maiden, M.C., Feavers, I.M., 2015. *FetA* antibodies induced by an outer membrane vesicle vaccine derived from a serogroup B meningococcal isolate with constitutive feta expression. PloS One. 10, e0140345.

Sarkari, J., Pandit, N., Moxon, E.R., Achtman, M., 1994. Variable expression of the *Opc* outer membrane protein in *Neisseria meningitidis* is caused by size variation of a promoter containing poly-cytidine. Molecular Microbiology. 13, 207-217.

Schneider, M. C., Exley, R. M., Ram, S., Sim, R. B. & Tang, C. M. 2007. Interactions between *N. meningitidis* and the complement system. Trends Microbiol, 15, 233-40.

Schneider, M.C., Prosser, B.E., Caesar, J.J., Kugelberg, E., Li, S., Zhang, Q., Quoraishi, S., Lovett, J.E., Deane, J.E., Sim, R.B., 2009. *Neisseria meningitidis* recruits factor H using protein mimicry of host carbohydrates. Nature. 458, 890-893.

Schoen, C., Blom, J., Claus, H., Schramm-Glck, A., Brandt, P., Mller, T., Goesmann, A., Joseph, B., Konietzny, S., Kurzai, O., 2008. Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in *Neisseria meningitidis*. Proceedings of the National Academy of Sciences. 105, 3473-3478.

Schoen, C., Kischkies, L., Elias, J., Ampattu, B.J., 2014. Metabolism and virulence in *Neisseria meningitidis*. Front Cell Infect Mi, 4.

Schoen, C., Tettelin, H., Parkhill, J., Frosch, M., 2009. Genome flexibility in *Neisseria meningitidis*. Vaccine. 27, B111.

Scholten, R., Kuipers, B., Valkenburg, H.A., Dankert, J., Zollinger, W.D., Poolman, J.T., 1994. Lipo-oligosaccharide immunotyping of *Neisseria meningitidis* by a whole-cell ELISA with monoclonal antibodies. Journal of Medical Microbiology. 41, 236-243.

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics. btu153.

Segal, E., Billyard, E., So, M., Storzbach, S., Meyer, T.F., 1985. Role of chromosomal rearrangement in *N. gonorrhoeae* pilus phase variation. Cell. 40, 293-300.

Serruto, D., Adu-Bobie, J., Scarselli, M., Veggi, D., Pizza, M., Rappuoli, R., Aricò, B., 2003. *Neisseria meningitidis App*, a new adhesin with autocatalytic serine protease activity. Molecular Microbiology. 48, 323-334.

Shea, M.W., 2013. The long road to an effective vaccine for meningococcus group B (MenB). Annals of Medicine and Surgery. 2, 53-56.

Shultz, T.R., White, P.A., Tapsall, J.W., 2005. In vitro assessment of the further potential for development of fluoroquinolone resistance in *Neisseria meningitidis*. Antimicrobial Agents and Chemotherapy. 49, 1753-1760.

Siddique, A., Buisine, N., 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 Genet. 7, e1001277.

Sim, R.J., Harrison, M.M., Moxon, E.R., Tang, C.M., 2000. Underestimation of meningococci in tonsillar tissue by nasopharyngeal swabbing. The Lancet. 356, 1653-1654.

Sinclair, D., Preziosi, M., Jacob John, T., Greenwood, B., 2010. The epidemiology of meningococcal disease in India. Tropical Medicine and International Health. 15, 1421-1435.

Skaar, E.P., Lazio, M.P., Seifert, H.S., 2002. Roles of the recJ and recN genes in homologous recombination and DNA repair pathways of *Neisseria gonorrhoeae*. Journal of Bacteriology. 184, 919-927.

Smith, J.M., 1992. Analyzing the mosaic structure of genes. Journal of Molecular Evolution. 34, 126-129.

Smith, N.H., Smith, J.M., Spratt, B.G., 1995. Sequence evolution of the *porB* gene of *Neisseria gonorrhoeae* and *Neisseria meningitidis*: evidence of positive Darwinian selection. Molecular Biology and Evolution. 12, 363-370.

Snape, M.D., Dawson, T., Oster, P., Evans, A., John, T.M., Ohene-Kena, B., Findlow, J., Yu, L., Borrow, R., Ypma, E., 2010. Immunogenicity of two investigational serogroup B meningococcal vaccines in the first year of life: a randomized comparative trial. The Pediatric Infectious Disease Journal. 29, e79.

Snyder, L.A., Cole, J.A., Pallen, M.J., 2009. Comparative analysis of two *Neisseria gonorrhoeae* genome sequences reveals evidence of mobilization of Correia Repeat Enclosed Elements and their role in regulation. BMC Genomics. 10, 70.

Snyder, L.A., Jarvis, S.A., Saunders, N.J., 2005. Complete and variant forms of the ‘gonococcal genetic island’in *Neisseria meningitidis*. Microbiology. 151, 4005-4013.

Snyder, L.A., McGowan, S., Rogers, M., Duro, E., O'farrell, E., Saunders, N.J., 2007. The repertoire of minimal mobile elements in the Neisseria species and evidence that these are

involved in horizontal gene transfer in other bacteria. Molecular Biology and Evolution. 24, 2802-2815.

Sohpal, V.K., Dey, A. and Singh, A., 2011. Substitution model analysis of human herpes simplex virus using molecular evolutionary genetic analysis.

Solovyev, V. and Salamov, A., 2011. Automatic annotation of microbial genomes and metagenomic sequences. Metagenomics and its applications in agriculture, biomedicine and environmental studies, pp.61-78.

Soucy, S.M., Huang, J., Gogarten, J.P., 2015. Horizontal gene transfer: building the web of life. Nature Reviews Genetics. 16, 472-482.

Spies, M. and Kowalczykowski, S.C., 2005. Homologous recombination by RecBCD and RecF pathways. The Bacterial Chromosome. 389-403.

Spratt, B.G., Hanage, W.P., Feil, E.J., 2001. The relative contributions of recombination and point mutation to the diversification of bacterial clones. Current Opinion in Microbiology. 4, 602-606.

Stabler, R.A., Marsden, G.L., Witney, A.A., Li, Y., Bentley, S.D., Tang, C.M., Hinds, J., 2005. Identification of pathogen-specific genes through microarray analysis of pathogenic and commensal Neisseria species. Microbiology. 151, 2907-2922.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30, 1312-1313.

Stefanelli, P., Miglietta, A., Pezzotti, P., Fazio, C., Neri, A., Vacca, P., Voller, F., D'Ancona, F.P., Guerra, R., Iannazzo, S., 2016. Increased incidence of invasive meningococcal disease of serogroup C/clonal complex 11, Tuscany, Italy, 2015 to 2016. Eurosurveillance, 21 .

Stephens, D.S., 2009. Biology and pathogenesis of the evolutionarily successful, obligate human bacterium *Neisseria meningitidis*. Vaccine. 27, B77.

Stokes, R.H., Oakhill, J.S., Joannou, C.L., Gorringe, A.R., Evans, R.W., 2005. Meningococcal transferrin-binding proteins A and B show cooperation in their binding kinetics for human transferrin. Infection and Immunity. 73, 944-952.

Strelow, V.L. and Vidal, J.E., 2013. Invasive meningococcal disease. Arquivos De Neuro-Psiquiatria. 71, 653-658.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution. 28, 2731-2739.

Tan, A., Hill, D.M., Harrison, O.B., Srikhanta, Y.N., Jennings, M.P., Maiden, M.C., Seib, K.L., 2016. Distribution of the type III DNA methyltransferases *modA*, *modB* and *modD* among *Neisseria meningitidis* genotypes: implications for gene regulation and virulence. Scientific Reports, 6.

Tauseef, I., Harrison, O.B., Wooldridge, K.G., Feavers, I.M., Neal, K.R., Gray, S.J., Kriz, P., Turner, D.P., Ala'Aldeen, D.A., Maiden, M.C., 2011. Influence of the combination and phase variation status of the haemoglobin receptors *HmbR* and *HpuAB* on meningococcal virulence. Microbiology. 157, 1446-1456.

Telisinghe, L., Waite, T., Gobin, M., Ronveaux, O., Fernandez, K., Stuart, J., Scholten, R., 7 th March 2014. Systematic review of the effect of antibiotics and/or vaccination in preventing subsequent disease among household contacts of cases of meningococcal disease WHO protocol to inform the revision of meningitis outbreak response guidelines.

Tettelin, H., Saunders, N.J., Heidelberg, J., Jeffries, A.C., Nelson, K.E., Eisen, J.A., Ketchum, K.A., Hood, D.W., Peden, J.F., Dodson, R.J., 2000. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science. 287, 1809-1815.

Thompson, E.A., Feavers, I.M., Maiden, M.C., 2003. Antigenic diversity of meningococcal enterobactin receptor *FetA*, a vaccine component. Microbiology. 149, 1849-1858.

Thompson, M.J., Ninis, N., Perera, R., Mayon-White, R., Phillips, C., Bailey, L., Harnden, A., Mant, D., Levin, M., 2006. Clinical recognition of meningococcal disease in children and adolescents. The Lancet. 367, 397-403.

Thorpe, H.A., Bayliss, S.C., Hurst, L.D., Feil, E.J., 2016. The large majority of intergenic sites in bacteria are selectively constrained, even when known regulatory elements are excluded. bioRxiv. 069708.

Thorvaldsdttir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics. 14, 178-192.

Traverse, C.C. and Ochman, H., 2016. Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. Proceedings of the National Academy of Sciences. 113, 3311-3316.

Treangen, T.J., Ambur, O.H., Tonjum, T., Rocha, E.P., 2008. The impact of the neisserial DNA uptake sequences on genome evolution and stability. Genome Biology. 9, R60.

Treangen, T.J., Ondov, B.D., Koren, S., Phillippy, A.M., 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biology. 15, 1-15.

Tryfinopoulou, K., Kesanopoulos, K., Xirogianni, A., Marmaras, N., Papandreou, A., Papaevangelou, V., Tsolia, M., Jasir, A., Tzanakaki, G., 2016. Meningococcal Carriage in Military Recruits and University Students during the Pre MenB Vaccination Era in Greece (2014-2015). PloS One. 11, e0167404.

Turner, P.C., Thomas, C.E., Stojiljkovic, I., Elkins, C., Kizel, G., Ala'Aldeen, D.A., Sparling, P.F., 2001. Neisserial TonB-dependent outer-membrane proteins: detection, regulation and distribution of three putative candidates identified from the genome sequences. Microbiology. 147, 1277-1290.

Tzeng, Y., Noble, C., Stephens, D.S., 2003. Genetic basis for biosynthesis of the (α1→ 4)-linked N-acetyl-d-glucosamine 1-phosphate capsule of *Neisseria meningitidis* serogroup X. Infection and Immunity. 71, 6712-6720.

Tzeng, Y. and Stephens, D.S., 2015. Antimicrobial peptide resistance in *Neisseria meningitidis*. Biochimica Et Biophysica Acta (BBA)-Biomembranes. 1848, 3026-3031.

Tzeng, Y. and Stephens, D.S., 2000. Epidemiology and pathogenesis of *Neisseria meningitidis*. Microbes and Infection. 2, 687-700.

Uberos J., Molina-Oya, M., Martinez-Serrano, S., Fernández-López, L., 2015. Surface adhesion and host response as pathogenicity factors of *Neisseria meningitidis*. World J Clin Infect Dis. 5(2), 37-43.

Urwin, R., Holmes, E.C., Fox, A.J., Derrick, J.P., Maiden, M.C., 2002. Phylogenetic evidence for frequent positive selection and recombination in the meningococcal surface antigen *PorB*. Molecular Biology and Evolution. 19, 1686-1694.

Van der Ende, A., Hopman, C., Dankert, J., 1999. Deletion of *porA* by recombination between clusters of repetitive extragenic palindromic sequences in *Neisseria meningitidis*. Infection and Immunity. 67, 2928-2934.

van Deuren, M., van der Ven-Jongekrijg, J., Bartelink, A.K., van Dalen, R., Sauerwein, R.W., van der Meer, Jos WM, 1995. Correlation between proinflammatory cytokines and anti-inflammatory mediators and the severity of disease in meningococcal infections. Journal of Infectious Diseases. 172, 433-439.

Vernikos, G. and Medini, D., 2014. Bexsero® chronicle. Pathogens and Global Health. 108, 305-316.

Vipond, C., Care, R., Feavers, I.M., 2012. History of meningococcal vaccines and their serological correlates of protection. Vaccine. 30, B17.

Virji, M., 2009. Pathogenic neisseriae: surface modulation, pathogenesis and infection control. Nature Reviews Microbiology. 7, 274-286.

Virji, M., Makepeace, K., Ferguson, D.J., Achtman, M., Sarkari, J., Moxon, E.R., 1992. Expression of the Opc protein correlates with invasion of epithelial and endothelial cells by *Neisseria meningitidis*. Molecular Microbiology. 6, 2785-2795.

Vogel, U., Taha, M., Vazquez, J.A., Findlow, J., Claus, H., Stefanelli, P., Caugant, D.A., Kriz, P., Abad, R., Bambini, S., 2013. Predicted strain coverage of a meningococcal multicomponent vaccine (4CMenB) in Europe: a qualitative and quantitative assessment. The Lancet Infectious Diseases. 13, 416-425.

Vos, M., 2009. Why do bacteria engage in homologous recombination? Trends in Microbiology. 17, 226-232.

West, D., Reddin, K., Matheson, M., Heath, R., Funnell, S., Hudson, M., Robinson, A., Gorringe, A., 2001. Recombinant *Neisseria meningitidis* transferrin binding protein A protects against experimental meningococcal infection. Infection and Immunity. 69, 1561-1567.

Wisniewski-Dy, F. and Vial, L., 2008. Phase and antigenic variation mediated by genome modifications. Antonie Van Leeuwenhoek. 94, 493-515.

Witz, G. and Stasiak, A., 2010. DNA supercoiling and its role in DNA decatenation and unknotting. *Nucleic Acids Research*. 83, 2119–2133.

Wolfgang, M., Park, H., Hayes, S.F., Van Putten, J.P., Koomey, M., 1998. Suppression of an absolute defect in type IV pilus biogenesis by loss-of-function mutations in *pilT*, a twitching motility gene in *Neisseria gonorrhoeae*. Proceedings of the National Academy of Sciences. 95, 14973-14978.

Wong, C.T., Xu, Y., Gupta, A., Garnett, J.A., Matthews, S.J., Hare, S.A., 2015. Structural analysis of haemoglobin binding by *HpuA* from the Neisseriaceae family. Nature Communications, 6.

Woods, J.P. and Cannon, J.G., 1990. Variation in expression of class 1 and class 5 outer membrane proteins during nasopharyngeal carriage of *Neisseria meningitidis*. Infection and Immunity. 58, 569-572.

Wright, B.E., 2000. A biochemical mechanism for nonrandom mutations and evolution. Journal of Bacteriology. 182, 2993-3001.

Yazdankhah, S.P. and Caugant, D.A., 2004. *Neisseria meningitidis*: an overview of the carriage state. Journal of Medical Microbiology. 53, 821-832.

Yero, D., Vipond, C., Climent, Y., Sardinas, G., Feavers, I.M., Pajon, R., 2010. Variation in the *Neisseria meningitidis* FadL-like protein: an evolutionary model for a relatively low-abundance surface antigen. Microbiology. 156, 3596-3608.

Yogev, R. and Tan, T., 2011. Meningococcal disease: The advances and challenges of meningococcal disease prevention. Human Vaccines. 7, 828-837.

Yu, D., Jin, Y., Yin, Z., Ren, H., Zhou, W., Liang, L., Yue, J., 2014. A genome-wide identification of genes undergoing recombination and positive selection in Neisseria. BioMed Research International, 2014.

Zerbino, D.R., 2010. Using the velvet de novo assembler for short-read sequencing technologies. Current Protocols in Bioinformatics. 11.5. 12.

Zheng, W., Mutha, N.V., Heydari, H., Dutta, A., Siow, C.C., Jakubovics, N.S., Wee, W.Y., Tan, S.Y., Ang, M.Y., Wong, G.J., 2016. NeisseriaBase: a specialised Neisseria genomic resource and analysis platform. PeerJ. 4, e1698.

Zhang, Y., Yang, J., Xu, L., Zhu, Y., Liu, B., Shao, Z., Zhang, X., Jin, Q., 2014. Complete genome sequence of *Neisseria meningitidis* serogroup A strain NMA510612, isolated from a patient with bacterial meningitis in China. Genome Announcements. 2, 360.

Zhou, J., Bowler, L.D., Spratt, B.G., 1997. Interspecies recombination, and phylogenetic distortions, within the glutamine synthetase and shikimate dehydrogenase genes of *Neisseria meningitidis* and commensal Neisseria species. Molecular Microbiology. 23, 799-812.

Zhou, J., Lefebvre, B., Deng, S., Gilca, R., Deceuninck, G., Law, D.K., De Wals, P., Tsang, R.S., 2012. Invasive serogroup B *Neisseria meningitidis* in Quebec, Canada, 2003 to 2010:

persistence of the ST-269 clone since it first emerged in 2003. Journal of Clinical Microbiology. 50, 1545-1551.

Zhu, P., Morelli, G., Achtman, M., 1999. The *opcA* and *ΨopcB* regions in Neisseria: genes, pseudogenes, deletions, insertion elements and DNA islands. Molecular Microbiology. 33, 635-650.

Zhu, W., Hunt, D.J., Richardson, A.R., Stojiljkovic, I., 2000. Use of heme compounds as iron sources by pathogenic neisseriae requires the product of the hemO gene. Journal of Bacteriology. 182, 439-447.