Meningococcal genome dynamics: an allele based, population approach to define lineage structure

Holly Bea Bratcher
Hertford College

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2015

Meningococcal genome dynamics: an allele based, population approach to define lineage structure

Holly Bea Bratcher

Hertford College

A thesis submitted for the degree of Doctor of Philosophy, Trinity 2015

ABSTRACT:

Advances in genome sequencing technologies have rapidly expanded the number of bacterial genome sequences available for study, permitting the emergence of the discipline of population genomics. Bioinformatics platforms were used to exploit this resource by the provision of data in an easily accessible and uniform format.

The *de novo* assembly, combined with gene-by-gene annotation, generated high quality draft genomes in which the majority of protein-encoding genes were present with high accuracy. The approach catalogued diversity efficiently and was a practical approach to interpreting whole genome sequence data for a large bacterial population.

The hyperinvasive meningococcal Lineage 3 core and accessory genome was described and mobile genetic elements and undescribed proteins were found to be influencing the shape of the lineages evolution and population structure.

Commensal carriage of the meningococcus was examined using temporally paired isolates. Long-term carriage was found and the comparison of the genomes pairs found a highly conserved set of core genes.

The methods used generated novel insights into the biology of the meningococcus and improved our understanding of the whole population structure, not just disease causing lineages. This work contributes to knowledge of genomic evolution of bacteria and population structure within a species.

Declaration

The work herein is my own except where explicitly stated in Appendix B and in relevant chapters. No part of this work has been submitted for any other degree or professional qualification.

Holly B. Bratcher, Trinity 2015

CONTENTS

CHAPTER 4:

CHAPTER 5:

Chapter 1 General introduction

1. Genetics and genomics

In 1905, the term genetics was introduced by British biologist William Bateson after discovering and translating Gregor Mendel's paper, "Experiments with Plant Hybrids". Danish biologist Wilhelm Johannsen expanded these early concepts by defining the gene (the basic unit of heredity), the genotype (the genetic makeup of an organism), and the phenotype (that which reflects an organism's observable characteristics) [1]. As the field of classical genetics evolved, the terminology was accepted and assimilated. Thomas Morgan, in his chromosomal theory of heredity, proposed that genes are linked in a series and are responsible for hereditary traits [2]; however, it was the study of bacteria that provided the first evidence that DNA is the bearer of genetic information and what, in fact, a gene is composed of [3]. From this point on, most of the methods used in molecular biology and indeed much of what we know about molecular genetics, including the language of the genetic code, originated with the study of bacteria.

The first bacterial gene was isolated in 1969, and the first bacterial gene sequenced was completed in 1972. Within 20 years, the Sanger dideoxy chain termination sequencing method and PCR were being used to characterize a multitude of bacterial isolates using their genetic material [4-7]. The sequencing of DNA became one of the most powerful molecular tools of its time and still is today; revolutionizing biological research, medical diagnostics, forensic science, and biotechnology. Each organism, however, maintains hundreds and thousands of genes more than what was being targeted during these early years. The sequencing of complete genomes was not practicable at that point, primarily because of technical limitations and cost. The first large-scale genome sequencing capability became possible using cloning procedures, and the development was in parallel to a new analysis method where sequencing data could be collected directly by a computer [8]. Almost 20 years after the introduction of DNA sequencing the first completely sequenced bacterial genomes were published in 1995 (*Haemophilus influenzae* and *Mycoplasma genitalium*) [9, 10]. The automation of sequencing and the use of whole genome

shotgun sequencing procedures were key to the development and went on to produce several complete bacterial genomes [8]. It took nearly 150 years of progress, beginning with the study of heredity, for the field of genetics to emerge and to continue to expand and grow into what is today, a multifaceted field of genomic research.

2. Insights from microbial genomic studies

Classical genetic studies are concerned with the study of the mechanisms of heritable information, and over time the emphasis shifted to include studies of transmission. In bacteria, this includes the study of chromosomal DNA as well as plasmids or phages, with the ultimate aim to determine the transmission of genotypes from one generation to the next. Molecular genetics, as it is now known, goes one step further to study the structure, function and interaction of individual genes at the molecular level. With the focus on understanding how the genome encodes information in the size, structure, synteny and the distinction of or between genes. In many ways, each of these genetic disciplines has not changed with the introduction of whole genome-based studies but has instead been refined and expanded.

Genomics applies DNA sequencing methods and bioinformatics approaches to sequence, assemble and analyse genomes incorporating the disciplines of functional, structural and comparative genetics. For example, genomic analysis using pulsed-field gel electrophoresis (PFGE) provided data on genome size but could not distinguish the plasmid components from the chromosomal. Whole genome sequencing has the capability to provide a comprehensive analysis of the structure and content of microbial genomes. The capacity to sequence whole genomes changed our perception of bacterial evolution and its relationship to the population structure and biological variation; linking phenotypically important traits to the underlying genotypic structure. From this point, hypothesis-driven questions can be formed about speciation and the diversifying force that structures lineages, as well as ecology and niche adaptation, based on the genomic analysis [11, 12].

Two discoveries have changed our understanding of some assumptions regarding bacteria. The first, that genetic diversity within a bacterial species is far greater than expected, and second, that strictly clonal descent is a rare state found in only a few species, such as *Bacillus anthracis* and *Yersinia pestis* [13, 14]. The change in understanding includes the long-held belief that *Escherichia coli* is a 'typical' bacterium [15, 16]. The majority of bacterial genomes are dynamic, variable and can include phage inserts and extra-chromosomal DNA in the form of plasmids [17]. Both the chromosomal and extra-chromosomal material is exposed to various genetic events that mutate, duplicate, invert, transpose and recombine genomic sequence data. More importantly while any one of these exposures may produce neutral sequence changes in terms of evolution, events like horizontal gene transfer, recombination, insertions or deletions make reconstructing evolutionary phylogenies extremely difficult [18, 19].

Several genome sequencing studies of diverse pathogen collections have revealed insights into the evolution and transmission of pathogenic organisms. Studies can stretch globally across decades or focus within a single patient over many years. One such comparative study established that a 50-year history of pandemic *Vibrio cholerea* was, in fact, three overlapping waves with geographical associations [20]. Another study analysed the mutations of *Pseudomonas aeruginosa* isolates collected from a single cystic fibrosis patient over 20 years [21]. Whole genome sequencing (WGS) studies have also helped to elucidate the transmission routes of *Escherichia coli* in a foodborne outbreak in Germany and the increase in methicillin-resistant *Staphylococcus aureus* (MRSA) in hospital [22, 23].

The turn of the 21st century witnessed an enormous amount of change in the typing methods used, in particular for the meningococcus. The genetic taxonomy of the family *Neisseriaceae* pioneered in 1961 by Catlin and Cunningham demonstrated the close relationships between most *Neisseria* species using DNA transformation studies and base ratio analysis [24]. Throughout the late 1980s and 1990s, several molecular and genetic techniques were developed and used to explore targeted sequence diversity of *Neisseria meningitidis* for the precise characterization of isolates recovered from cases of invasive disease [25]. They included the first

two genetically based classification systems developed to type isolates, multilocus enzyme electrophoresis (MLEE) in 1986 and multilocus sequence typing (MLST) in 1998 [26, 27]. These genetic typing techniques complimented traditional microbial methods for phenotypic and serological typing and played a defining role in establishing the population structure and biology of many bacteria. The progression to the level of whole genome sequence analysis continues to yield more and more information, and to generate novel insights into the biology, population structure, and evolution of *N. meningitidis*.

3. Aims

The research chapters (Chapters Three to Five) are part of a much larger research program, of which one focus is to investigate why the *Neisseria* population is structured at the level of species and subspecies (also referred to as lineages), to determine if the link between genotype and phenotype is selective for the population structure observed, and to develop a theoretical framework that describes the evolution of this structure, see Appendix B, page 130). To accomplish this, a method that would facilitate the use of whole genome sequence data to explore the population biology of *N. meningitidis* was required.

The first aim (Chapter Three) was to establish the genomic methods needed to sequence, assemble, annotate, and analyse whole genome data that allowed a functional validation of the reliability, repeatability and robustness of the data generated. Building on from there, the next aim (Chapter Four) was to use the developed method with a gene-by-gene MLST approach to investigate the population structure of N. meningitidis using genomes from a large, diverse group of related strains to examine what defines a meningococcal lineage, its diversity, and its structure; using a population genomics approach to identify the characteristics of genetic structure. The analysis includes, but is not limited to, establishing a pan and core genome, the

ribosomal protein gene profile, the presence of phenotypic clustering, and categorizing variation as either genetic drift or shift within a representative collection of genomes, the ST-41/44 clonal complex. Over all this chapter also contributes knowledge to the larger program question, how the hyperivasive phenotypes emerge.

The final research chapter (Chapter Five) looks at the genetic diversity hidden in the predominant unit of analysis, the typing assignments such as clonal complex and sequence type (ST) which is based on sequence fragments of seven housekeeping genes. The aim was to assess how similar the genomes of the same ST are when recovered from the same individual over a year apart and whether long term carriage variation could be characterised as within host evolution or variation at the point of exposure/infection.

4. Evolutionary and genomic insights into meningococcal biology

The content was published as Bratcher HB, Bennett JS, Maiden MC. 'Evolutionary and genomic insights into meningococcal biology'. Future Microbiology 2012, 7:873-85. HB Bratcher drafted and wrote the manuscript, JS Bennett created Figures 1 and 2, and MC Maiden approved the outline, manuscript and contributed to the final editing.

Evolutionary and genomic insights into meningococcal biology

Holly B Bratcher[1], Julia S Bennett[1] & Martin CJ Maiden*[1]
[1]Department of Zoology, University of Oxford, Oxford OX1 3PS, UK
*Author for correspondence: Tel.: +44 1865 271 284 ■ Fax: +44 1865 281 275 ■ martin.maiden@zoo.ox.ac.uk

Epidemic disease caused by *Neisseria meningitidis*, the meningococcus, has been recognized for two centuries, but remains incompletely controlled and understood. There have been dramatic reductions in serogroup A and C meningococcal disease following the introduction of protein–polysaccharide conjugate vaccines, but there is currently no comprehensive vaccine against serogroup B meningococci. Genetic analyses of meningococcal populations have provided many insights into the biology, evolution and pathogenesis of this important pathogen. The meningococcus, and its close relative the gonococcus, are the only pathogenic members of the genus *Neisseria*, and the invasive propensity of meningococci varies widely, with approximately a dozen 'hyperinvasive lineages' responsible for most disease. Despite this, attempts to identify a 'pathogenome', a subset of genes associated with the invasive phenotypes, have failed; however, genome-wide studies of representative meningococcal isolates using high-throughput sequencing are beginning to provide details on the relationship of invasive phenotype and genotype in this fascinating organism and how this relationship has evolved.

Review
Future Microbiology

A short history of the meningococcus & meningococcal disease

Historical evidence suggests that epidemic meningococcal disease emerged at the beginning of the 19th century, with the first likely outbreaks reported only 2 years apart but in different continents: in Geneva, Switzerland in 1805 [1] and in Medfield, MA, USA a year later [2]. Throughout the succeeding decades, the disease became widely recognized throughout the USA and Europe with a number of large-scale outbreaks. Epidemic meningococcal disease was not described in Africa until 1905, but major epidemics occurred in the meningitis belt for much of the 20th century [3]. The causative organism, now known as *Neisseria meningitidis* or the meningococcus, was isolated from human cerebrospinal fluid in 1887 [4], and by 1890, it was acknowledged that the organism could be found in an asymptomatic carrier state, in addition to being isolated from cases of invasive disease [5]. Fulminant meningococcal septicemia, or Waterhouse–Friderichsen syndrome, the most clinically serious manifestation of the disease, was recognized in the early 20th century [6] and, as the century progressed, serological analyses were increasingly used to identify distinct types of meningococcal isolates, eventually recognized as capsular serogroups, of which serogroups A, B, C, W and Y are most commonly associated with disease [7]. The incidence of meningococcal disease is influenced by the virulence potential of circulating meningococci, which depends at least in part on host and environmental factors, and on host susceptibility to disease [8]. At the time of writing, the disease had an annual incidence rate that ranged from one to 1000 cases per 100,000 individuals in different parts of the world.

The advent of penicillin and a variety of other antimicrobial agents in the mid-20th century made a major impact on mortality as invasive meningococci are very susceptible to treatment with systemic antimicrobials. Indeed, as long as an appropriate antimicrobial is administered promptly, prognosis is extremely good [9]; however, the principal problem with the treatment of meningococcal disease is that it can develop extraordinarily quickly, with an apparently healthy individual becoming a medical emergency within a matter of hours [8]. This problem is exacerbated by the nonspecific nature of early symptoms. Therefore, prophylaxis, especially the development and implementation of vaccines, is the best public health intervention against this disease.

The first effective vaccines against this disease, the 'plain' polysaccharide vaccines, were developed in the 1960s [10] but were poorly immunogenic and ineffective in interrupting person-to-person transmission over more than a few weeks [11,12]. At the end of the 20th and beginning of the 21st century, these were replaced with protein conjugate polysaccharide vaccines [13], which were much more immunogenic and

Keywords
- molecular epidemiology
- molecular evolution
- *Neisseria meningitidis*
- pathogenomics
- population genomics

Future Medicine part of fsg

capable of interrupting transmission [14], thereby generating herd immunity [15]. While these vaccines proved to be extremely effective against meningococci of serogroups A, C, W and Y, the similarity of the B polysaccharide to the human antigen NCAM has precluded the development of a serogroup B vaccine [16,17]. As serogroup B meningococci are the most prevalent causes of disease in many countries [18], the development of a 'group B' substitute vaccine remains a major research priority [14]; however, the high antigenic and genetic variability of the meningococcus have proved major obstacles in the development of such a vaccine.

Initial studies of meningococcal variation employed serological analyses (i.e., the comparison of the reactivity of sera raised against a particular meningococcus with other meningococci). This led to the identification of capsular serogroups, which are the consequence of the expression of one of 11 immunochemically distinct capsular polysaccharides [19]. Later studies identified additional antigenic variation in the outer membrane proteins (OMPs) and short-chain lipopolysaccharides (often referred to as 'lipooligosaccharides') of this Gram-negative organism, leading to the development of meningococcal serotypes and serosubtypes (OMPs PorB and PorA, respectively) and immunotypes (lipopolysaccharides) [20]. Since the widespread availability of nucleotide sequencing, sequence-based typing has become the method of choice for characterizing the highly variable surface proteins of this organism [21].

The meningococcus has been in the forefront of work on the population biology of bacteria in general and pathogens in particular, being among the first organisms to be investigated by multilocus enzyme electrophoresis [22] and the first to be examined by multilocus sequence typing (MLST) [23]. Despite intensive research effort over several decades, which has led to a detailed appreciation of many aspects of the genetics of this organism, much remains to be learned concerning its biology and, particularly, why some meningococci are appreciably more invasive than others. Much has been learned, however, by placing this organism in the context of its close relatives, which in general do not cause a pathology similar to the meningococcus, and comparative studies have provided valuable insights into meningococcal biology and continue to represent a promising approach for improving our understanding of the pathogenicity of this bacterium in the genomic era [24].

The genus *Neisseria*

The genus *Neisseria* comprises Gram-negative, oxidase-positive, aerobic β-proteobacteria that inhabit human and animal mucosal and dental surfaces [25]. Apart from the meningococcus, only one other species, *Neisseria gonorrhoeae,* the gonococcus, is regarded as pathogenic, the other members of the genus being apparently harmless commensal inhabitants of the microbiota that remain relatively stable colonizers over time, despite regular minor perturbations to the oral environment [26].

Unlike its relatives, which inhabit the human oropharynx, the gonococcus colonizes the mucosal surface of the urogenital tract and is considered to be an obligate mucosal pathogen that occasionally causes disseminated infection [27], although it can on rare occasions also inhabit other mucosal sites. Gonococci are most commonly found to infect young adults and do not always cause symptoms, especially in women. The gonococcus is the least diverse of the *Neisseria* and the present day gonococcal population is likely to be descended, relatively recently, from a single clone, which adapted to the change in niche [28]. Unusually for a 'single clone pathogen' (i.e., a pathogen that has arisen by the invasion of a single clone into a pathogenic niche accompanied by genetic isolation) [29], the gonococcus has continued to recombine within its own population while being reproductively isolated from other *Neisseria*, and it has exhibited a remarkable ability to acquire novel genetic elements encoding phenotypes such as antimicrobial-resistance determinants [28,30]. A distinguishing feature of the gonococcus relative to the meningococcus is that, despite the absence of a polysaccharide capsule, it is nonetheless virulent and causes symptomatic disease quite unlike invasive meningococcal strains.

In addition to the two pathogens, the other member of the genus that has been more extensively investigated is *Neisseria lactamica*, principally because of its lack of association with invasive disease, despite being found at high rates of carriage in infants and young children [31]. Acquisition of *N. lactamica* is very rapid in the first few years of life and this organism establishes a long-term relationship with its host, with carriage rates declining gradually as the age of the host population rises [32]. In European populations, carriage of *N. lactamica* is low in adolescents and young adults [33], while the carriage of the meningococcus increases to very high levels at this age [34], especially in closed

or semiclosed environments with populations in close association such as military recruits and university dormitories [35–37]. This has led to the suggestion on more than one occasion, that carriage of *N. lactamica* is responsible for the increase in immunity against meningococci that is observed during childhood, and that deliberate colonization or immunization with *N. lactamica* is a potential approach to prophylaxis against the meningococcus [31,38].

Members of the genus *Neisseria* are closely related genetically and poorly resolved by techniques for bacterial speciation such as DNA–DNA hybridization [39] and 16S rRNA sequencing [40], leading to the suggestion that they are not distinct species; however, this suggestion has never found favor among medical microbiologists due to the fact that the defined species nevertheless correspond to particular pathogenic phenotypes. An analysis of three *Neisseria* genomes using the Artemis Comparison Tool [41] indicates large syntenic regions of similarity; however, further examination of these homologous regions show variability that is distinct to each species (Figure 1). Recent genomic analyses show that the microbiologically defined species are, by and large, genealogically coherent and distinct groups, which can be reconstructed in a variety of ways, including ribosomal MLST, as seen in Figure 2 [42,43]. The genus provides an excellent model system for studying the genetic traits of closely related organisms that exhibit very different phenotypes, which includes two globally significant diseases [24,44]. For example, why does the meningococcus, normally a generally harmless commensal organism, occasionally cause disease, and what are the differences between it and *N. lactamica*, which occurs in the same niche? What were the genetic events that led to the gonococcus establishing itself in the urogenital tract and why has this only occurred once [28,45]? These questions are amenable to population genomic studies that compare the genomes of representative collections of phenotypically defined isolates.

Population structure, epidemiology & evolution

The widespread acceptance of seven-locus MLST [46] for the characterization of *Neisseria*

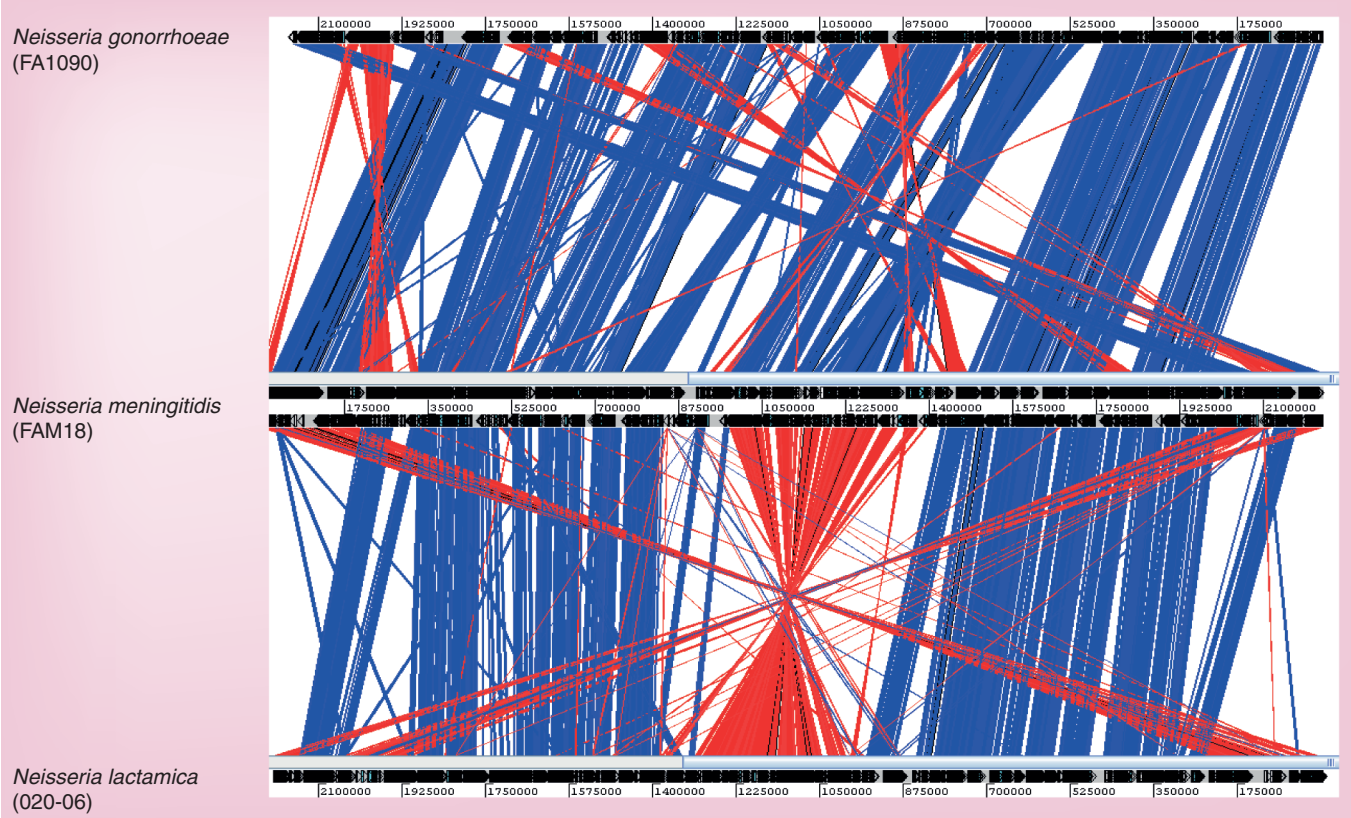


Figure 1. A comparison of the genome organization of *Neisseria meningitis* FAM18 (GenBank accession no. AM421808), *Neisseria gonorrhoeae* FA1090 (GenBank accession no. NC_002946) and *Neisseria lactamica* 020-06 (GenBank accession no. FN995097). The red and blue bars indicate regions of similarity, with blue bars indicating corresponding regions that are oriented similarly and red bars indicating regions oriented in opposite directions. The coordinates represent the relative nucleotide positions on the genomes.

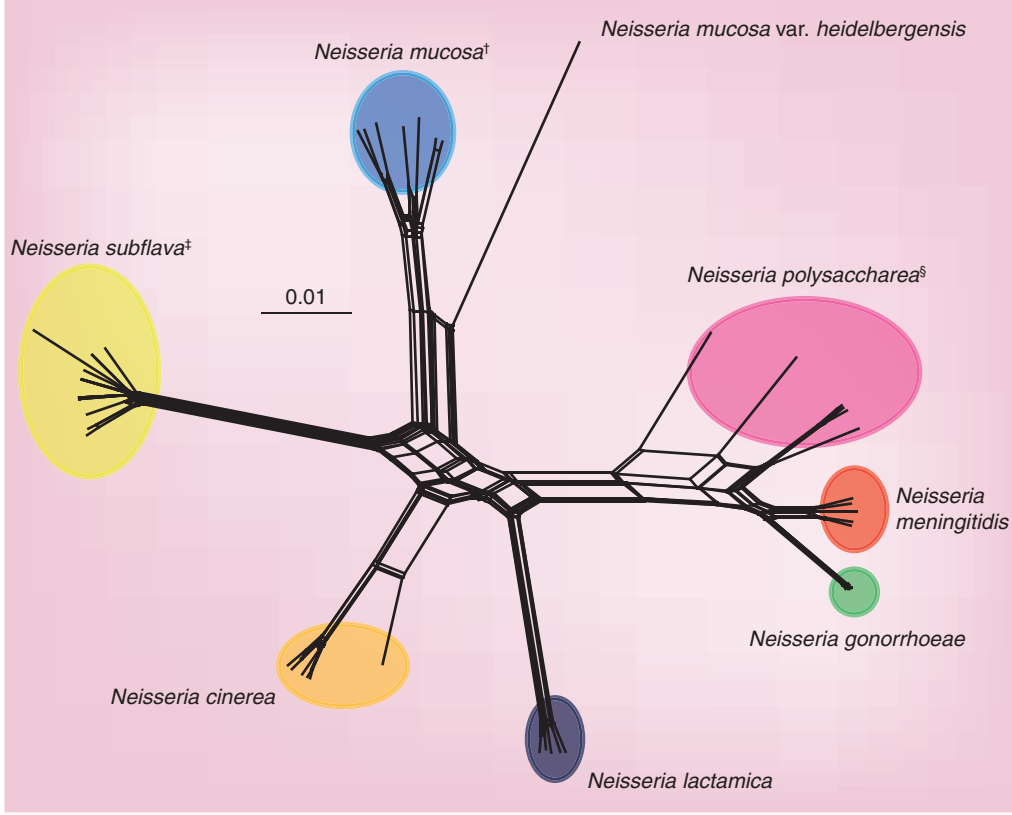


Figure 2. Evolutionary relationships among *Neisseria* isolates based on concatenated sequences of 53 ribosomal protein genes (ribosomal multilocus sequence typing).
The evolutionary history was inferred using the neighbor-net method. The analysis involved 49 nucleotide sequences consisting of 20,927 nucleotides.
Scale bar indicates the number of nucleotide substitutions per site.
†This group contains the species *Neisseria sicca* and *Neisseria macacae*, which are closely related to *Neisseria mucosa* and are likely to be variants of this species.
‡This group also contains the species *Neisseria flavescens*, which is closely related to *Neisseria subflava* and is likely to be a variant of this species.
§This is a polyphyletic group, which suggests that more than one species is present among strains microbiologically defined as *Neisseria polysaccharea*.

isolates provided a reproducible and unambiguous means of typing the meningococcus [23], gonococcus [28] and *N. lactamica* [32], and indeed other species [42]. MLST data confirmed many of the observations made with its predecessor technique, multilocus enzyme electrophoresis [22,47], and have enabled studies of the molecular epidemiology [48], evolution [49] and population genetics of these organisms [34]. Other sequence-based identification and typing schemes have been developed, including antigen gene sequence typing, or 'fine typing', for classification by indexing variation at OMP genes [21] and genes responsible for reduced susceptibility to antimicrobials [50]. These highly reproducible and portable schemes have enabled the assembly of a growing public repository, pubMLST.org/neisseria [51,101], comprising data from over 17,000 meningococcal isolates from both disease and carriage. At the time of writing, this database also contained more than 300 gonococcal, 500 *N. lactamica* and over 80 isolates of other *Neisseria* species.

MLST and antigen gene sequence typing data have demonstrated that meningococcal populations, while highly diverse and dominated by frequent recombination, are nevertheless highly structured into groups of related organisms referred to as clonal complexes [52]. These clonal complexes correspond to clades reconstructed with phylogenetic and genealogical analyses, although with seven- and even 20-locus data, relationships among these complexes are not apparent [53]. An important insight gained from the comparative study of isolates from invasive disease and asymptomatic carriage [48] is that some of these clonal complexes are more invasive than others: these are referred to as the 'hyperinvasive lineages', and only a handful of these have caused the majority

of reported disease globally in the last half of the 20th century [52]. These lineages exhibit stability in their antigenic and disease phenotype, including their epidemiology, although these properties are different among different hyperinvasive lineages (Table 1). This has established a robust framework for the comparison of *Neisseria* genomes, both among and within recognized species groups, with the aim of elucidating the relationship between genotype and phenotype, with the ultimate aim of designing improved interventions to reduce or prevent disease.

Horizontal genetic exchange was once thought to be uncommon in most prokaryotic species, but is now known to be a major factor in the evolution of most bacterial species [54]; indeed, the lack of genetic exchange is now regarded as a special case (e.g., the genetically monomorphic single clone pathogens such as *Mycobacterium tuberculosis*, *Bacillus anthracis*, *Yersinia pestis* and *Salmonella enterica* var. Typhi) [29]. Work on the *Neisseria*, and specifically the meningococcus, was influential in establishing this paradigm [55,56], as they are naturally competent for DNA uptake, and horizontal genetic exchange is most

Table 1. Stability of the *Neisseria* clonal complexes over the past decade[†].

Clonal complex	*Neisseria lactamica*			*Neisseria meningitidis*			*Neisseria* species[‡]		
	Disease	*Carriage*	*Undefined*	*Disease*	*Carriage*	*Undefined*	*Disease*	*Carriage*	*Undefined*
ST-41/44 complex				1474	783	288			
ST-11 complex				1208	158	126			
ST-5 complex				623	36	20			
ST-8 complex				370	12	26			
ST-1 complex				177	46	11			
ST-4 complex				18		27			
ST-32 complex				969	254	182			
ST-269 complex				421	137	82			
ST-23 complex				192	247	63			1
ST-22 complex				155	272	42			
ST-35 complex				155	243	55			
ST-18 complex				239	52	36			
ST-213 complex				114	141	58			
ST-53 complex				6	290	15			
ST-60 complex				113	155	29			1
ST-167 complex				78	141	20			
ST-198 complex				8	196	34		1	
ST-254 complex				49	120	24			
ST-162 complex				74	78	32			
ST-103 complex				89	58	31			
ST-624 complex		50	23					4	
ST-640 complex		42						5	
ST-613 complex		44						2	
ST-1494 complex		38					1	8	
ST-595 complex		32						1	
ST-1540 complex		21						1	
Unassigned STs		244	35	1298	1452	355		1	34

[†]*The table was generated in March 2012 from the pubMLST.org/neisseria database* [101] *and includes the top 20 of 51* Neisseria meningitidis *clonal complexes, six defined* Neisseria lactamica *clonal complexes and unspeciated* Neisseria *strains considered to be either* N. meningitidis *or* N. lactamica *on the basis of their multilocus sequence typing profiles. The list shows 90% of the profiles in the database and represents examples of isolate profiles from 1917 to 2012.*
[‡]*Definitive speciation for 60* Neisseria *isolates is currently inconclusive and based on multilocus sequence typing and/or fine typing only.*
ST: Sequence type.

frequently mediated by means of transformation followed by integration into the genome, usually by homologous recombination. Although mostly this exchange occurs among meningococci, occasionally this genetic transfer can occur over large phylogenetic distances, an example being the superoxide dismutase gene, *sodC*, which appears to have been acquired by *N. meningitidis* via horizontal transfer from *Haemophilus influenzae* [57].

This type of 'localized sex' in bacteria is most often thought of as a means of generating genetic diversity, but more recent evidence suggests that this process is normally conservative, rather than diversifying [58]. The *Neisseria* have a number of mechanisms that make the transformation process very specific. The first of these is provided by the DNA uptake sequences distributed in the *Neisseria* chromosome, which play a major role in both the DNA uptake and its integration into the chromosome [58]. In addition, *Neisseria* have a large number of restriction modification systems, which in the meningococci are characteristic of particular clonal complexes [59], suggesting that a given meningococcus is most likely to incorporate DNA from its very closest relatives [60]. This view is supported by the distribution of the DNA uptake sequences, which are concentrated in conserved rather than variable regions of the chromosome, suggesting that under normal circumstances, horizontal genetic exchange is used as a repair mechanism, rather than as a mechanism for diversification in the meningococci [58].

Genetic differences among *Neisseria* isolates with different phenotypes

The existence of genealogically related organisms with distinct properties presents the possibility of conducting 'association studies', where the relationships of particular characteristics of organisms with their phenotypes are investigated. As the amount of genetic information increases, so does the opportunity of investigating the relationships of particular genetic variants with specific phenotypes. In the case of the meningococcus, the phenotype of most interest is the propensity to cause disease, and the genetic and phenotypic diversity within the genus presents many opportunities to investigate this. Once an association is established, it is then necessary to establish the causality of this association. It is worth reflecting that this general approach is not new, and historically was the way in which the predominant virulence factor of the meningococcus, the polysaccharide capsule, was characterized [19] – virtually all invasive meningococci are capsulate, a property which confers serum resistance. This serum resistance provides a mechanism for the observed association [61]. While a necessary property for invasion, the expression of one of the disease-associated capsules is not, however, sufficient in itself, as many capsulate meningococci are nevertheless unlikely to cause disease [48].

Despite the increases in the data and technology available to undertake association studies, there has been rather limited success in identifying additional determinants that are consistently associated with invasion. Various studies have employed hybridization approaches with microarrays or filters to compare the genes present among panels of isolates against various reference sequences [62–68]. This approach has the advantage that many isolates can be compared, but has the limitation that the method can only efficiently identify the presence or absence of genes present in the reference isolate or isolates – additional genes present in the tested isolates remain unknown, as does the influence of sequence variants within genes that are universally present. Thus, in such studies, inference is limited to identifying the presence and absence of those members of the accessory genome present in the reference genome(s) in the tested strains, and sequence variation in the core genome and shared members of the accessory genome are not detected.

Most comparative genomic hybridization studies have concluded that, with the exception of the *cps* region, which encodes the capsule, the invasive phenotype is polygenic in meningococci, with extensive lateral gene transfer reassorting putative virulence determinates [66]. The meningococcal disease-associated island, which was identified by comparisons of a coherent epidemiological isolate collection [69], contributes to the virulence of certain clonal complexes in teenagers, but interestingly not in infants [70], and appears to be an integrated bacteriophage. Other studies have shown particular iron metabolism genes to be positively or negatively associated with the likelihood of causing invasion, and these are associated with particular clonal complexes, but the associations are not absolute [71,72]. Both human and bacterial factors are likely to be important in determining the outcome of infection, and although human genetic association studies are difficult to perform for meningococcal disease because of the low number of well-characterized patients, variants in human factor H have been implicated in the severity of meningococcal disease [73].

Meningococcal genomics & vaccine design

The meningococcus was among the first bacteria to have its complete sequence determined [74], an endeavor that was, to a large extent, motivated by the attempts to develop a 'serogroup B' meningococcal vaccine [75]. The first isolate to be genome sequenced was a serogroup B meningococcal variant of the ST-32 complex, the laboratory strain MC58, which was derived from an isolate from a protracted hyperendemic outbreak in Stroud, Gloucestershire, UK [76]. The Stroud outbreak was one of many hyperendemics caused by members of this clonal complex (originally termed the ET-5 complex) that occurred globally from the late 1960s onwards and has continued until the time of writing (2012) [77–83]. The MC58 sequence was followed by a serogroup A (Z2491) [84] and then a serogroup C isolate (FAM18) [85].

The MC58 genome was used as the basis of the 'reverse vaccinology' approach, which was developed to identify novel vaccine candidates from genome sequences [86]. This approach started with the genome sequence, exploiting the sequences to express the genes identified in the annotation process. These were then used in a mouse antibody screen to identify novel antigens that would be potentially protective and hopefully conserved. While this approach has led to the identification of four proteins that are included in a vaccine undergoing trials, it was necessary to include them with an existing outer membrane vesicle vaccine, and most of them are not conserved among meningococci but are highly diverse [87]. One of these components, the factor H-binding protein, was also discovered by more conventional biochemical methods [88] and displays appreciable antigenic diversity [89] such that multiple variants are required for vaccine coverage.

In an era of whole-genome sequences, alternative approaches to developing vaccines can exploit genome data from populations, rather than single genomes, as the starting point of the analysis. Such approaches can take advantage of models of strain structure imposed by immune selection in order to identify proteins that are under immune selection and to rationally design vaccine cocktails that take advantage of the biology and antigenic variation, rather than trying to circumvent it [90]. The long lifespans of most hyperinvasive meningococcal strain types, which are measured in several decades at least, combined with their antigenic stability, suggest that this is a feasible approach.

Opportunities & challenges presented by advances in genome sequencing

The scale and efficiency of high-throughput 'next-generation' parallel sequencing methods, combined with their reduced costs, provides the unprecedented prospect of large numbers of whole-genome sequences becoming available for the meningococcus in the near future. At least in the short term, these will not be fully annotated, closed genome sequences, but rather a number of contiguous sequence assemblies generated from short-read sequencing technologies, such as the Illumina HiSeq™ 2000 (Illumina, Inc., CA, USA) instrument. With sequence data availability increasing more rapidly than conventional analyses can be performed, novel strategies for data analysis, storage and output are required if full advantage is to be taken of them.

The initial studies using this type of data concentrated on bacteria with relatively limited genetic variation: either the study of relatively 'young' single clone pathogens or closely related members of more diverse species, often initially chosen on the basis of MLST data [91,92]. For such analyses, it was possible to borrow analysis techniques from studies of human populations, which are not very diverse, for the identification of single nucleotide polymorphisms (SNPs) by mapping against a complete closed reference genome. This approach, however, is not appropriate with more diverse bacteria, such as the meningococcus where there is high diversity (i.e., extensive sequence polymorphism rather than SNPs), and especially when this diversity is reassorted by recombination. Another limitation of these studies was the requirement of a reference genome, a high-quality, finished genome that is closely related to the genomes being analyzed and against which SNPs can be called. These approaches are not scalable to a situation where many hundreds or thousands of partial genome sequences are being generated for diverse members of a bacterial species.

We have recently developed an alternative paradigm for the analysis of whole-genome sequence data, which places allelic diversity among bacterial isolates at the center of multiple-genome analysis (Figure 3). This approach is effectively an expansion of the analysis concept that has been central to the success of MLST [23] and has the advantage that multiple genomes that are highly diverse can be simultaneously analyzed. While most MLST schemes index variation at one to ten 'housekeeping loci', usually fragments of genes under stabilizing selection [46], there is, in principle, no limit to the number or type of loci that can be analyzed in this way, up to all

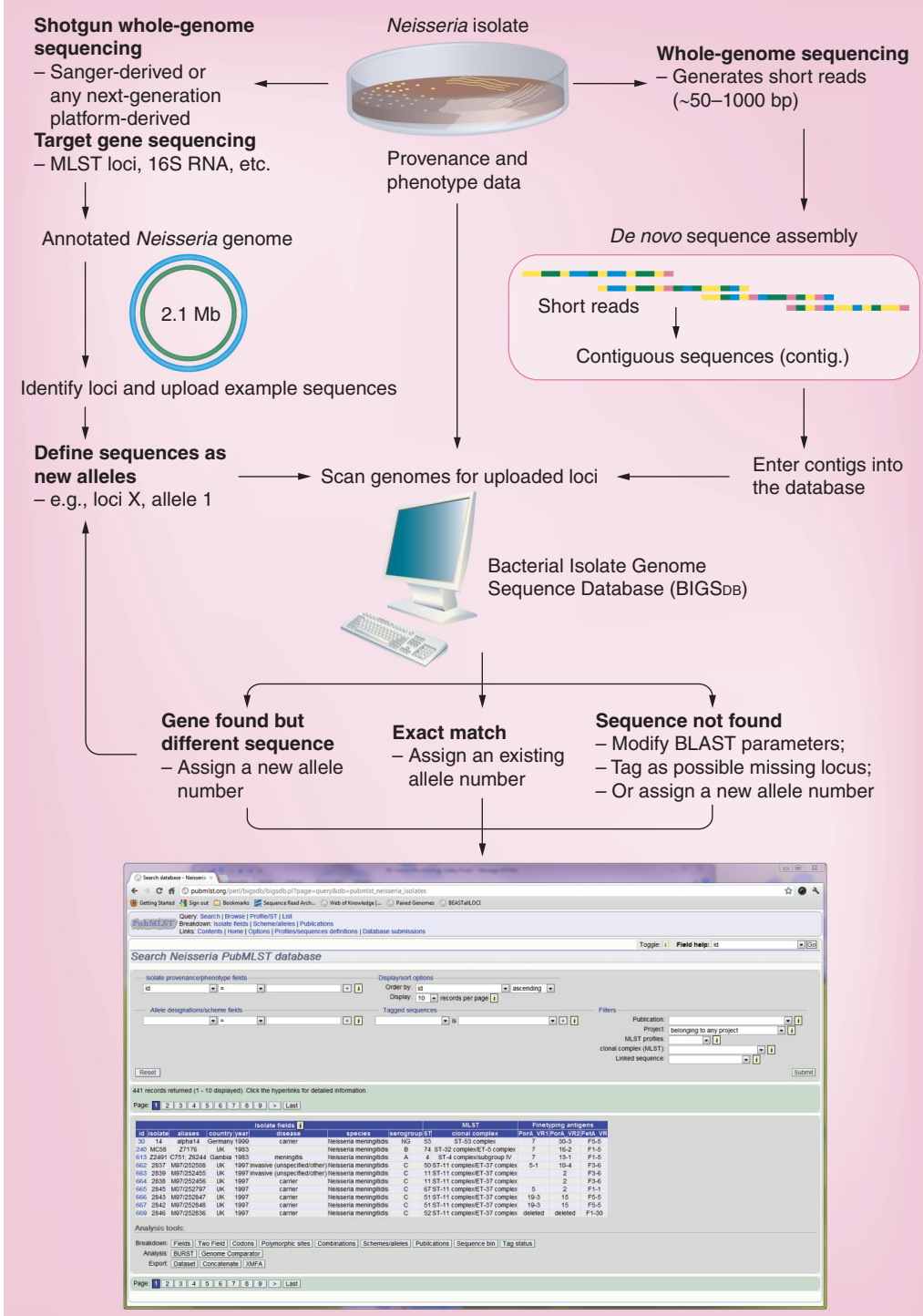


Figure 3. The Bacterial Isolate Genome Sequence Database (BIGSDB)†, an integrated online database that links provenance, phenotype, bibliographic and sequence data for use in applications for epidemiological, evolutionary and functional studies.

†See [101].

BLAST: Basic local alignment search tool; MLST: Multilocus sequence typing.

Adapted with permission from [99].

the loci present in a genome. For the purposes of characterizing and cataloging genetic variation, a locus can be defined as any sequence string, nucleotide or peptide, which can be identified by its sequence, genomic context or a combination of the two. As in MLST, each novel variant of each locus is assigned an allele number, and the sequence and the number are stored in a curated

data table, providing a comprehensive catalog of the variation observed to date for that locus. These tables are easily expanded as new variants are detected. The advantage of this approach is that the sequence variation at a given locus can be summarized as a single number, and a genome can be rapidly identified as having a known variant at a particular locus, or a novel variant that can then be added to the curated data set for that locus.

For organisms in which recombination is common [49], this analysis approach has the advantage that analyses based on the alleles present, rather than their sequences or individual nucleotides, inherently correct for bias introduced by nucleotide-based phylogenetic or genealogical analyses by horizontal genetic exchange; indeed, this was why the approach was used for MLST of such organisms [46]. In organisms where recombination is absent or rare, the sequences can be used directly, as the sequence variation conforms to tree-like evolutionary models [93]. A further level of efficiency in cataloging variation is achieved by grouping alleles, as is done in MLST, to generate sequence types that effectively summarize variation data at seven loci with a single number. There is no limit to the number of such schemes that can be defined, each of which will group genetic variation in different ways (e.g., membership of the same macromolecular structure, such as the ribosome [43], or contribution to a particular phenotype, such as antibiotic resistance). Once the genetic variation of a given isolate is cataloged in this way, it can readily be associated with provenance and phenotype data for the isolate, often referred to as metadata.

This approach has been implemented with the BIGSdb [51] software on the pubMLST.org/neisseria website [101]. There are three fundamental levels of inter-related information within the website: first, a table of isolate provenance and phenotype data; second, a 'sequence bin' associated with each isolate; and third, tables of reference sequences. The isolate information table contains data on the provenance and phenotype of the isolate, together with any alternative names that have been used and links to relevant records in external databases, such as PubMed. The sequence bin can contain any type of compiled sequence data, including: single locus sequence data, such as MLST data, 16S rRNA genes and/or antimicrobia-resistance determinants; assembled draft genomes, such as those generated from 'next-generation' short-read sequences; or complete closed genome sequences. Each of these types of data can reside in a single sequence bin, associated with an experiment type, which indicates the likely quality of the data and can be used to analyze the data preferentially, such that complete closed genome data, where available, will be used in preference to draft genome data.

The tables of reference sequences store any number of reference sequences from any number of loci, grouped into any number of schemes, enabling rapid and flexible annotation of the sequences within the sequence bin using well-established search algorithms such as the basic local alignment search tool (BLAST). Once identified, the loci are tagged within the sequence bin for easy future identification, and the allelic designations are reported back to the isolate information table. This scanning process is automated so that as new genomes are added, they are automatically annotated against the loci defined in the reference tables. The reference tables themselves not only define variants and associate them with allele numbers, but can also contain links to other data, including publications, which indicate the function of the loci. These tables also give alternative locus names. To facilitate unambiguous labeling, each locus is assigned a unique identifier of the form NEIS0001. Loci are curated, but multiple curators are possible, such that individual experts in a particular locus can have responsibility for that locus. All features of the database are accessible through a web interface, so that the system provides a flexible and backwards-compatible means of cataloging and analyzing genome wide information. The PubMLST database employing the BIGSdb software provides a platform upon which population genomic analyses of *Neisseria* can be efficiently performed. It contains sequence data generated by single locus, multilocus and whole-genome approaches that can be analyzed together. A number of data analysis, summary and export tools are built into the system, and it is possible to conduct phylogenetic and genealogical analysis on the same data set that is being used for functional studies, enabling the effective fusion of these two powerful but not always well-integrated functions. This is especially important in bacteria such as the meningococcus, where population structure has to be taken into account when performing association studies to determine which genes or genetic variants are associated with particular phenotypes. The pathogenic phenotype in the meningococcus is polygenic and complex, but the data accumulated to date, which have demonstrated the existence of defined genotypes and their association with particular phenotypes, suggest that unraveling

the genetic elements associated with these traits will be achievable in the foreseeable future.

Conclusion

Large-scale studies of sequence variation in meningococcal populations have established the central role of horizontal genetic exchange in bacterial evolution. The analyses of these data have demonstrated that, notwithstanding high rates of recombination, meningococcal populations are highly structured [94]. This structure, which is evident from only a handful of housekeeping genes located at various positions on the chromosome, is associated with particular phenotypic properties, including the expression of particular antigens and the likelihood of causing invasive disease [52]. The current models of bacterial population structures and how they evolve were developed in the 1990s from sequence data [95], and the recent increase in sequencing capacity provides access to the whole genome, enabling these models to be refined, developed and exploited. It is perhaps worth noting, however, that the genomic revolution has not, at least as yet, led to a fundamental change in these models [96]. A particularly intriguing prospect made possible with whole-genome data is to definitively establish the components of the accessory genome associated with particular clonal complexes and the relationship of this repertoire with sequence variation in the core genome, which has been used to identify clonal complexes. As the data and methods to analyze them expand, we are poised to make exciting novel insights into the biology of this intriguing and dangerous, if accidental [97], pathogen. These insights will lead to the development of improved public health interventions.

Future perspective

There is little doubt that the number of meningococcal genomes available is going to increase dramatically in the next 5–10 years. At the time of writing, it was already more cost effective to obtain whole-genome sequences on the Illumina platform than to perform MLST using seven PCR reactions, at least at major genome centers. It is very likely that within 10 years, whole-genome sequencing will become the method of choice for the characterization of clinical isolates of the meningococcus, and perhaps even the sequencing of clinical specimens such as blood and cerebrospinal fluid, from which meningococci cannot be cultured due to prior antibiotic treatment [98]. In addition to resolving clinical questions, this will generate an invaluable resource for understanding the genetics of the meningococcus, perhaps in conjunction with the genetics of the host. However, exploitation of this resource will depend on the data being available in a structured form and comparison with equivalent data from meningococcal isolates obtained from asymptomatic carriage and other *Neisseria* species, as exemplified by the gene-by-gene analysis and annotation approach implemented on pubMLST.org/neisseria [101].

Financial & competing interests disclosure

The authors are grateful for the support of the Wellcome Trust. MCJ Maiden is a Wellcome Trust Senior Research Fellow. This publication made use of the Neisseria *Multilocus Sequence Typing website developed by K Jolley and sited at the University of Oxford, UK. The development of this site has been funded by the Wellcome Trust and European Union. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.*

No writing assistance was utilized in the production of this manuscript.

Executive summary

- Epidemic meningococcal disease, which takes the form of meningitis and septicemia, was first recognized in the early 19th century and continues to have a global impact on human health.
- With the exception of the meningococcus, *Neisseria meningitidis* and the gonococcus, *Neisseria gonorrhoeae*, the causative agent of gonorrhea, the members of the genus *Neisseria* are usually harmless commensal members of the microbiota of humans and other animals.
- Towards the end of the 20th century, the application of genetic methods established that meningococcal populations are very diverse but highly structured, with some meningococcal genotypes being much more likely to cause disease than others.
- Comparisons of a number of meningococcal and other genomes have failed to reveal a 'pathogenome' (i.e., a set of genes that are limited to the more invasive meningococci). Rather, these studies suggest that there are many ways in which a meningococcus can become more invasive.
- Population genomics, the application of 'next-generation' sequencing methods to large numbers of meningococcal isolates, is generating novel insights into the biology of the meningococcus, with implications for improved understanding of the biology of this disease and the development of novel methods for its control.

References

Papers of special note have been highlighted as:
- ▪ of interest
- ▪▪ of considerable interest

1. Vieusseux G. Memoire sur le maladie qui a régné a Genêve au printemps de 1805. *Journal de Médecine, Chirurgie et Pharmacie* II, 163–165 (1806).
2. Danielson L, Mann E. The history of a singular and very mortal disease which lately made its appearance in Medfield. *Med. Agric. Reg.* 1, 65–69 (1806).
3. Greenwood B. Manson Lecture. Meningococcal meningitis in Africa. *Trans. R. Soc. Trop. Med. Hyg.* 93(4), 341–353 (1999).
4. Weichselbaum A. Uber die Aetiologie der akuten Meningitis cerebrospinalis. *Fortschr.* 5, 573–575 (1887).
5. Kiefer F. Zur differential Diagnose des Erregers der epidemischen Cerebrospinalmeningitis und der Gonorrhoea. *Berl. Klin. Wochenschr.* 33, 628–630 (1896).
6. Waterhouse R. A case of suprarenal apoplexy. *Lancet* 1, 577–578 (1911).
7. Flexner S. The results of the serum treatment in thirteen hundred cases of epidemic meningitis. *J. Exp. Med.* 17, 553–576 (1913).
8. Stephens DS, Greenwood B, Brandtzaeg P. Epidemic meningitis, meningococcaemia, and *Neisseria meningitidis*. *Lancet* 369(9580), 2196–2210 (2007).
9. Cartwright K, Reilly S, White D, Stuart J. Early treatment with parenteral penicillin in meningococcal disease. *BMJ* 305, 143–147 (1992).
10. Gotschlich EC, Goldschneider I, Artenstein MS. Human immunity to the meningococcus IV. Immunogenicity of group A and group C meningococcal polysaccharides. *J. Exp. Med.* 129, 1367–1384 (1969).
11. Artenstein MS, Gold R, Zimmerly JG, Wyle FA, Schneider H, Harkins C. Prevention of meningococcal disease by group C polysaccharide vaccine. *N. Engl. J. Med.* 282, 417–420 (1970).
12. Sivonen A. Effect of *Neisseria meningitidis* group A polysaccharide vaccine on nasopharyngeal carrier rates. *J. Infect.* 3(3), 266–272 (1981).
13. Robbins JB, Schneerson R, Szu SC *et al.* Prevention of invasive bacterial diseases by immunization with polysaccharide-protein conjugates. *Curr. Top. Microbiol. Immunol.* 146, 169–180 (1989).
14. Tan LKK, Carlone GM, Borrow R. Advances in the development of vaccines against *Neisseria meningitidis*. *N. Engl. J. Med.* 362(16), 1511–1520 (2010).
15. Maiden MC, Ibarz-Pavon AB, Urwin R *et al.* Impact of meningococcal serogroup C conjugate vaccines on carriage and herd immunity. *J. Infect. Dis.* 197(5), 737–743 (2008).
16. Finne J, Leinonen M, Makela PH. Antigenic similarities between brain components and bacteria causing meningitis. Implications for vaccine development and pathogenesis. *Lancet* 2(8346), 355–357 (1983).
17. Robbins JB, Schneerson R, Xie G, Hanson LA, Miller MA. Capsular polysaccharide vaccine for group B *Neisseria meningitidis*, *Escherichia coli* K1, and *Pasteurella haemolytica* A2. *Proc. Natl Acad. Sci. USA* 108(44), 17871–17875 (2011).
18. Harrison LH, Trotter CL, Ramsay ME. Global epidemiology of meningococcal disease. *Vaccine* 27(Suppl. 2), B51–B63 (2009).
19. Vedros NA. Development of meningococcal serogroups. In: *Evolution of Meningococcal Disease*, Vedros NA (Ed.). CRC Press Inc., FL, USA, 33–37 (1987).
20. Frasch CE, Zollinger WD, Poolman JT. Serotype antigens of *Neisseria meningitidis* and a proposed scheme for designation of serotypes. *Rev. Infect. Dis.* 7, 504–510 (1985).
21. Jolley KA, Brehony C, Maiden MC. Molecular typing of meningococci: recommendations for target choice and nomenclature. *FEMS Microbiol. Rev.* 31(1), 89–96 (2007).
22. Caugant DA, Bovre K, Gaustad P *et al.* Multilocus genotypes determined by enzyme electrophoresis of *Neisseria meningitidis* isolated from patients with systemic disease and from healthy carriers. *J. Gen. Microbiol.* 132, 641–652 (1986).
23. Maiden MCJ, Bygraves JA, Feil E *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl Acad. Sci. USA* 95(6), 3140–3145 (1998).
24. Maiden MC. Population genomics: diversity and virulence in the *Neisseria*. *Curr. Opin. Microbiol.* 11(5), 1–5 (2008).
25. Tonjum T. Genus I. *Neisseria*. In: *Bergey's Manual of Systematic Bacteriology*, Garrity GM, Brenner DJ, Krieg NR, Staley JR (Eds). Springer-Verlag, NY, USA, 777–798 (2005).
26. Marsh PD, Percival RS. The oral microflora – friend or foe? Can we decide? *Int. Dent. J.* 56(4 Suppl. 1), 233–239 (2006).
27. Morse SA, Knapp JS. The genus *Neisseria*. In: *The Prokaryotes*. Balows A, Trüper HG, Dworkin M, Harder W, Schleifer K-H (Eds). Springer-Verlag, NY, USA, 2495–2559 (1992).
28. Bennett JS, Jolley KA, Sparling PF *et al.* Species status of *Neisseria gonorrhoeae*: evolutionary and epidemiological inferences from multilocus sequence typing. *BMC Biol.* 5(1), 35 (2007).
- ▪ **Examination of the relationships of the gonococcus, meningococcus and *Neisseria lactamica* using multilocus sequence typing data, showing that these are three phylogenetically coherent groups with minimal interspecies genetic exchange.**
29. Achtman M. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu. Rev. Microbiol.* 62, 53–70 (2008).
30. Lewis DA. The gonococcus fights back: is this time a knock out? *Sex. Transm. Infect.* 86(6), 415–421 (2010).
31. Gold R, Goldschneider I, Lepow ML, Draper TF, Randolph M. Carriage of *Neisseria meningitidis* and *Neisseria lactamica* in infants and children. *J. Infect. Dis.* 137(2), 112–121 (1978).
32. Bennett JS, Griffiths DT, McCarthy ND *et al.* Genetic diversity and carriage dynamics of *Neisseria lactamica* in infants. *Infect. Immun.* 73(4), 2424–2432 (2005).
33. Cartwright KA, Stuart JM, Jones DM, Noah ND. The Stonehouse survey: nasopharyngeal carriage of meningococci and *Neisseria lactamica*. *Epidemiol. Infect.* 99(3), 591–601 (1987).
34. Claus H, Maiden MC, Wilson DJ *et al.* Genetic analysis of meningococci carried by children and young adults. *J. Infect. Dis.* 191(8), 1263–1271 (2005).
35. Caugant DA, Fogg C, Bajunirwe F *et al.* Pharyngeal carriage of *Neisseria meningitidis* in 2–19-year-old individuals in Uganda. *Trans. R. Soc. Trop. Med Hyg.* 100(12), 1159–1163 (2006).
36. Glitza IC, Ehrhard I, Muller-Pebody B *et al.* Longitudinal study of meningococcal carrier rates in teenagers. *Int. J. Hyg. Environ. Health* 211(3–4), 263–272 (2008).
37. Caugant DA, Tzanakaki G, Kriz P. Lessons from meningococcal carriage studies. *FEMS Microbiol. Rev.* 31(1), 52–63 (2007).
38. Oliver KJ, Reddin KM, Bracegirdle P *et al.* *Neisseria lactamica* protects against experimental meningococcal infection. *Infect. Immun.* 70(7), 3621–3626 (2002).
39. Guibourdenche M, Popoff MY, Riou JY. Deoxyribonucleic acid relatedness among *Neisseria gonorrhoeae*, *N. meningitidis, N. lactamica, N. cinerea* and "*Neisseria polysaccharea*". *Ann. Inst. Pasteur Microbiol.* 1373(2), 177–185 (1986).

40. Harmsen D, Singer C, Rothganger J *et al.* Diagnostics of *Neisseriaceae* and *Moraxellaceae* by ribosomal DNA sequencing: ribosomal differentiation of medical microorganisms. *J. Clin. Microbiol.* 39(3), 936–942 (2001).
41. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. ACT: the Artemis Comparison Tool. *Bioinformatics* 21(16), 3422–3423 (2005).
42. Bennett JS, Jolley KA, Earle SG *et al.* A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus *Neisseria*. *Microbiology* doi:10.1099/mic.0.056077-0 (2012) (Epub ahead of print).
43. Jolley KA, Bliss CM, Bennett JS *et al.* Ribosomal multi-locus sequence typing: universal characterisation of bacteria from domain to strain. *Microbiology* 158, 1005–1015 (2012).

▪▪ **First description of a bacterial isolate characterization technique that can provide information at all levels of resolution, from domain to strain, by indexing variation in the 53 ribosomal protein subunit genes.**

44. Falush D. Toward the use of genomics to study microevolutionary change in bacteria. *PLoS Genet.* 5(10), e1000627 (2009).
45. Bennett JS, Bentley SD, Vernikos GS *et al.* 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 (2010).
46. Maiden MC. Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.* 60, 561–588 (2006).
47. Caugant DA. Population genetics and molecular epidemiology of *Neisseria meningitidis*. *APMIS* 106(5), 505–525 (1998).
48. Yazdankhah SP, Kriz P, Tzanakaki G *et al.* Distribution of serogroups and genotypes among disease-associated and carried isolates of *Neisseria meningitidis* from the Czech Republic, Greece, and Norway. *J. Clin. Microbiol.* 42(11), 5146–5153 (2004).
49. Holmes EC, Urwin R, Maiden MCJ. The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. *Mol. Biol. Evol.* 16(6), 741–749 (1999).
50. Taha MK, Vazquez JA, Hong E *et al.* Target gene sequencing to characterize the penicillin G susceptibility of *Neisseria meningitidis*. *Antimicrob. Agents Chemother.* 51(8), 2784–2792 (2007).
51. Jolley KA, Maiden MC. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 11(1), 595 (2010).
52. Caugant DA, Maiden MC. Meningococcal carriage and disease – population biology and evolution. *Vaccine* 27(Suppl. 2), B64–B70 (2009).
53. Didelot X, Urwin R, Maiden MC, Falush D. Genealogical typing of *Neisseria meningitidis*. *Microbiology* 155(10), 3176–3186 (2009).
54. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* 323(5915), 741–746 (2009).
55. Maynard Smith J, Dowson CG, Spratt BG. Localized sex in bacteria. *Nature* 349, 29–31 (1991).
56. Maynard Smith J, Smith NH, O'Rourke M, Spratt BG. How clonal are bacteria? *Proc. Natl Acad. Sci. USA* 90(10), 4384–4388 (1993).
57. Kroll JS, Wilks KE, Farrant JL, Langford PR. Natural genetic exchange between *Haemophilus* and *Neisseria*: intergeneric transfer of chromosomal genes between major human pathogens. *Proc. Natl Acad. Sci. USA* 95(21), 12381–12385 (1998).
58. Treangen TJ, Ambur OH, Tonjum T, Rocha EP. The impact of the neisserial DNA uptake sequences on genome evolution and stability. *Genome Biol.* 9(3), R60 (2008).

▪▪ **Examination of the distribution of DNA-uptake sequences in the *Neisseria* genome, suggesting a conservative, rather than diversifying, role for genetic exchange in these organisms.**

59. Budroni S, Siena E, Hotopp JCD *et al.* *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc. Natl Acad. Sci. USA* 108(11), 4494–4499 (2011).

▪▪ **Shows that despite high rates of genetic recombination, whole-genome analysis reveals major clades within the meningococcal population structure, providing an evolutionary framework for investigations of virulence of hyperinvasive meningococci.**

60. Claus H, Stoevesandt J, Frosch M, Vogel U. Genetic isolation of meningococci of the electrophoretic type 37 complex. *J. Bacteriol.* 183(8), 2570–2575 (2001).
61. Ram S, Mackinnon FG, Gulati S *et al.* The contrasting mechanisms of serum resistance of *Neisseria gonorrhoeae* and group B *Neisseria meningitidis*. *Mol. Immunol.* 36(13–14), 915–928 (1999).
62. Tinsley CR, Nassif X. Analysis of the genetic differences between *Neisseria meningitidis* and *Neisseria gonorrhoeae*: two closely related bacteria expressing two different pathogenicities. *Proc. Natl Acad. Sci. USA* 93(20), 11109–11114 (1996).
63. Perrin A, Nassif X, Tinsley C. Identification of regions of the chromosome of *Neisseria meningitidis* and *Neisseria gonorrhoeae* which are specific to the pathogenic *Neisseria* species. *Infect. Immun.* 67(11), 6119–6129 (1999).
64. Stabler RA, Marsden GL, Witney AA *et al.* Identification of pathogen-specific genes through microarray analysis of pathogenic and commensal *Neisseria* species. *Microbiology* 151(Pt 9), 2907–2922 (2005).
65. Hotopp JC, Grifantini R, Kumar N *et al.* Comparative genomics of *Neisseria meningitidis*: core genome, islands of horizontal transfer and pathogen-specific genes. *Microbiology* 152(Pt 12), 3733–3749 (2006).
66. Joseph B, Schwarz RF, Linke B *et al.* Virulence evolution of the human pathogen *Neisseria meningitidis* by recombination in the core and accessory genome. *PLoS One* 6(4), e18441 (2011).
67. Snyder LA, Saunders NJ. The majority of genes in the pathogenic *Neisseria* species are present in non-pathogenic *Neisseria lactamica*, including those designated as 'virulence genes'. *BMC Genomics* 7, 128 (2006).

▪ **Highlights the problems and difficulties of using microarray data to identify a meningococcal pathogenome.**

68. Joseph B, Schneiker-Bekel S, Schramm-Gluck A *et al.* Comparative genome biology of a serogroup B carriage and disease strain supports a polygenic nature of meningococcal virulence. *J. Bacteriol.* 192(20), 5363–5377 (2010).
69. Bille E, Zahar JR, Perrin A *et al.* A chromosomally integrated bacteriophage in invasive meningococci. *J. Exp. Med.* 201(12), 1905–1913 (2005).
70. Bille E, Ure R, Gray SJ *et al.* Association of a bacteriophage with meningococcal disease in young adults. *PLoS one* 3(12), e3885 (2008).
71. Tauseef I, Harrison OB, Wooldridge KG *et al.* Influence of the combination and phase variation status of the haemoglobin receptors HmbR and HpuAB on meningococcal virulence. *Microbiology* 157, 1446–1456 (2011).
72. Harrison OB, Evans NJ, Blair JM *et al.* Epidemiological evidence for the role of the hemoglobin receptor, HmbR, in meningococcal virulence. *J. Infect. Dis.* 200(1), 94–98 (2009).
73. Davila S, Wright VJ, Khor CC *et al.* Genome-wide association study identifies variants in the CFH region associated with

host susceptibility to meningococcal disease. *Nat. Genet.* 42(9), 772–763 (2010).

74. Tettelin H, Saunders NJ, Heidelberg J *et al.* Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* 287(5459), 1809–1815 (2000).
75. Jodar L, Feavers IM, Salisbury D, Granoff DM. Development of vaccines against meningococcal disease. *Lancet* 359(9316), 1499–1508 (2002).
76. Cartwright KA, Stuart JM, Noah ND. An outbreak of meningococcal disease in Gloucestershire. *Lancet* 2, 558–561 (1986).
77. Wylie PA, Stevens D, Drake W 3rd, Stuart J, Cartwright K. Epidemiology and clinical management of meningococcal disease in west Gloucestershire: retrospective, population based study. *BMJ* 315(7111), 774–779 (1997).
78. Takahashi H, Kuroki T, Watanabe Y *et al.* Characterization of *Neisseria meningitidis* isolates collected from 1974 to 2003 in Japan by multilocus sequence typing. *J. Med. Microbiol.* 53(Pt 7), 657–662 (2004).
79. De Filippis I, Vicente AC. Multilocus sequence typing and repetitive element-based polymerase chain reaction analysis of *Neisseria meningitidis* isolates in Brazil reveal the emergence of 11 new sequence types genetically related to the ST-32 and ST-41/44 complexes and high prevalence of strains related to hypervirulent lineages. *Diagn. Microbiol. Infect. Dis.* 53(3), 161–167 (2005).
80. Smith I, Caugant DA, Hoiby EA, Wentzel-Larsen T, Halstensen A. High case–fatality rates of meningococcal disease in western Norway caused by serogroup C strains belonging to both sequence type (ST)-32 and ST-11 complexes, 1985–2002. *Epidemiol. Infect.* 134(6), 1195–1202 (2006).
81. Gottfredsson M, Diggle MA, Lawrie DI *et al.* *Neisseria meningitidis* sequence type and risk for death, Iceland. *Emerg. Infect. Dis.* 12(7), 1066–1073 (2006).
82. Climent Y, Urwin R, Yero D *et al.* The genetic structure of *Neisseria meningitidis* populations in Cuba before and after the introduction of a serogroup BC vaccine. *Infect. Genet. Evol.* 10(4), 546–554 (2010).
83. De Filippis I, De Lemos APS, Hostetler JB *et al.* Molecular epidemiology of *Neisseria meningitidis* serogroup B in Brazil. *PLoS One* 7(3), e33016 (2012).
84. Parkhill J, Achtman M, James KD *et al.* Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* 404(6777), 502–506 (2000).
85. Bentley SD, Vernikos GS, Snyder LA *et al.* Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. *PLoS Genet.* 3(2), e23 (2007).
86. Pizza M, Scarlato V, Masignani V *et al.* Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 287(5459), 1816–1820 (2000).

▪ **Description of the process of reverse vaccinology, illustrated by its application to discover novel potential vaccine components against the meningococcus.**

87. Scarselli M, Arico B, Brunelli B *et al.* Rational design of a meningococcal antigen inducing broad protective immunity. *Sci. Transl. Med.* 3(91), 91ra62 (2011).
88. Fletcher LD, Bernfield L, Barniak V *et al.* Vaccine potential of the *Neisseria meningitidis* 2086 lipoprotein. *Infect. Immun.* 72(4), 2088–2100 (2004).
89. Brehony C, Wilson D, Maiden M. Variation of the factor H-binding protein of *Neisseria meningitidis*. *Microbiology* 155, 4155–4169 (2009).
90. Buckee CO, Gupta S, Kriz P, Maiden MCJ, Jolley KA. Long-term evolution of antigen repertoires among carried meningococci. *Proc. Biol. Sci.* 277(1688), 1635–1641 (2010).
91. Croucher NJ, Harris SR, Fraser C *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* 331(6016), 430–434 (2011).
92. Harris SR, Feil EJ, Holden MTG *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327(5964), 469–474 (2010).

▪ **First published large-scale use of whole-genome sequence data to analyze a large collection of disease isolates in order to resolve clinical questions.**

93. Didelot X, Maiden MC. Impact of recombination on bacterial evolution. *Trends Microbiol.* 18(7), 315–322 (2010).

▪ **Review of the impact of recombination on the population structure of different bacteria, highlighting the difficulty of distinguishing patterns of admixture due to descent from that caused by recent horizontal genetic exchange.**

94. Buckee CO, Jolley K, Recker M *et al.* Role of selection in the emergence of lineages and the evolution of virulence in *Neisseria meningitidis*. *Proc. Natl Acad. Sci. USA* 105(39), 15082–15087 (2008).
95. Spratt BG, Maiden MCJ. Bacterial population genetics, evolution and epidemiology. *Proc. R. Soc. Lond. B Biol. Sci.* 354(1384), 701–710 (1999).
96. Enright MC, Spratt BG. Genomics. The genomic view of bacterial diversification. *Science* 331(6016), 407–409 (2011).
97. Maiden MC. Dynamics of bacterial carriage and disease: lessons from the meningococcus. *Adv. Exp. Med. Biol.* 549, 23–29 (2004).
98. Ni H, Knight AI, Cartwright K, Palmer WH, Mcfadden J. Polymerase chain reaction for diagnosis of meningococcal meningitis. *Lancet* 340, 1432–1434 (1992).
99. Sheppard SK, Jolley KA, Maiden MCJ. A gene-by-gene approach to bacterial population genomics: whole genome MLST of *Campylobacter*. *Genes* 3(2), 261–277 (2012).

Website

101. *Neisseria* sequence typing home page. http://pubMLST.org/neisseria

Chapter 2 Materials and Methods

1. Isolate collections

1.1. Chapter Three isolates

Archived cultures of 108 previously characterized isolates representing multiple clonal complexes recovered globally in the 20th century were used for analysis [26]. The data set includes several hyper-invasive lineages and a smaller number of carriage strains. The isolates represent the serogroups A, B, C, E, W, X, Y, Z, and the capsule null locus (*cnl*). Isolate records were created in the *Neisseria* PubMLST.org isolate database for each, and each record contains known provenance and previous clinical typing data (Table 1).

Table 1. Chapter Three isolates. Provenance and clinical typing data for 108 global reference *N. meningitidis* isolates of the 20th century. The table also includes the ribosomal sequence type (rST). An empty bracket in strain designation indicates no clonal complex association.

id	isolate	country	year	disease	strain designation	rST
1	A4/M1027	USA	1937	invasive	A: P1.5-2,10: F1-5: ST-4 (cc4)	2504
2	120M	Pakistan	1967	invasive	A: P1.5-2,10: F5-1: ST-1 (cc1)	2532
7	7891	Finland	1975	invasive	A: P1.20,9: F3-1: ST-5 (cc5)	2434
10	6748[a]	Canada	1971	invasive	A: P1.18-1,3: F5-1: ST-1 (cc1)	2530
11	129E	Germany	1964	invasive	A: P1.5-2,10: F3-6: ST-1 (cc1)	2528
13	139M[b]	Philippines	1968		A: P1.5-2,10: F5-1: ST-1 (cc1)	2433
19	S3131	Ghana	1973	invasive	A: P1.7,13-1: F1-5: ST-4 (cc4)	2285
24	S4355	Denmark	1974	invasive	A: P1.5-1,9: F3-1: ST-5 (cc5)	2434
31	10	Burkina Faso	1963	invasive	A: P1.7,13-1: F1-5: ST-4 (cc4)	2286
34	20	Niger	1963	invasive	A: P1.5-2,10: F1-7: ST-1 (cc1)	2951
35	26	Niger	1963	invasive	A: P1.7,13: F1-5: ST-4 (cc4)	2285
52	243	Cameroon	1966		A: P1.7,13: F1-5: ST-4 (cc4)	2286
61	393	Greece	1968	carrier	A: P1.5-2,10: F5-1: ST-1 (cc1)	2433
64	254	Djibouti	1966	invasive	A: P1.5-2,10: F1-7: ST-1 (cc1)	2537
67	S5611	Australia	1977	invasive	A: P1.5-2,10: F5-1: ST-1 (cc1)	2433
82	11-004	China	1984	invasive	A: P1.20,9: F3-8: ST-5 (cc5)	2434
84	IAL2229	Brazil	1976		A: P1.20,9: F2-1: ST-5 (cc5)	2434
90	CN100	UK	1941	invasive	A: P1.5-2,10: F3-9: ST-21 ()	2529
120	F4698[a]	Saudi Arabia	1987	carrier	A: P1.20,9: F3-1: ST-5 (cc5)	2434
128	F6124	Chad	1988	invasive	A: P1.20,9: F3-1: ST-5 (cc5)	2434
160	1014	Sudan	1985	invasive	A: P1.7,13-1: F1-5: ST-4 (cc4)	2287
210	H1964	UK	1987	invasive	A: P1.20,9: F3-1: ST-5 (cc5)	2434
237	H44/76	Norway	1976	invasive	B: P1.7,16: F3-3: ST-32 (cc32)	2291

id	isolate	country	year	disease	strain designation	rST
238	153	China	1966	invasive	A: P1.20,9: F3-1: ST-5 (cc5)	2468
239	154	China	1966	invasive	A: P1.20,9: F3-1: ST-6 (cc5)	2503
299	80049	China	1963	carrier	A: P1.5-2,10: F1-5: ST-5 (cc5)	2511
314	D1	Mali	1989	carrier	C: P1.5,2-1: F5-4: ST-11 (cc11)	2336
316	D8	Mali	1990	invasive	A: P1.7,13-1: F1-5: ST-4 (cc4)	2285
340	196/87	Norway	1987		C: P1.7-2,16-12: F3-3: ST-32 (cc32)	2371
343	500	Italy	1984		C: P1.5,2: F1-1: ST-11 (cc11)	2338
349	38VI	USA	1964		B: P1.5,2: F1-1: ST-11 (cc11)	2328
369	M597[a]	Israel	1988	invasive	C: P1.5,2-1: F5-5: ST-11 (cc11)	2337
387	2059001	Mali	1990	invasive	A: P1.7,13: F1-5: ST-4 (cc4)	2285
391	90/18311	UK	1990		C: P1.5,2-1: F5-5: ST-11 (cc11)	2333
398	BZ 10	The Netherlands	1967	invasive	B: P1.5-1,2-2: F3-9: ST-8 (cc8)	2365
400	BZ 83	The Netherlands	1984	invasive	B: P1.5-2,10: F5-1: ST-34 (cc32)	2290
403	BZ 147	The Netherlands	1963	invasive	B: P1.18-2,1-1: F3-9: ST-48 (cc41/44)	2506
407	BZ 163	The Netherlands	1979	invasive	B: P1.21,16: F1-7: ST-9 (cc8)	2366
408	BZ 169[a]	The Netherlands	1985	invasive	B: P1.7-2,16: F3-3: ST-32 (cc32)	2290
409	BZ 198	The Netherlands	1986	invasive	B: P1.7-2,4: F1-5: ST-41 (cc41/44)	2498
410	BZ 232	The Netherlands	1964	invasive	B: P1.5-2,2-2: F5-2: ST-38 (cc37)	2541
411	DK 24[a]	Denmark	1940	invasive	B: *deleted*: F1-8: ST-16 (cc4240/6688)	2546
412	DK 353	Denmark	1962	invasive	B: P1.5-2,2-2: F5-7: ST-37 (cc37)	2538
414	EG 327	Germany	1985	invasive	B: P1.7,14-1,14: F5-2: ST-19 (cc18)	2520
415	EG 329	Germany	1985	invasive	B: P1.7-1,16: F1-2: ST-32 (cc32)	2309
416	EG 011	Germany	1986	invasive	B: P1.18-1,3: F4-1: ST-36 ()	2539
417	NG 3/88	Norway	1988	invasive	B: P1.7-1,1: F5-8: ST-12 ()	2526
418	NG 4/88	Norway	1988	invasive	B: P1.18,25: F1-5: ST-30 ()	2411
419	NG 6/88	Norway	1988	invasive	B: P1.7-1,1: F5-6: ST-13 (cc269)	2524
420	NG F26	Norway	1988	carrier	B: P1.31,16: F3-9: ST-13 (cc269)	2533
421	NG H15	Norway	1988	carrier	B: P1.19,15-2: F1-5: ST-43 (cc41/44)	2465
423	NG H38	Norway	1988	carrier	B: P1.18-1,3: F4-1: ST-36 ()	2544
424	NG E31	Norway	1988	carrier	B: P1.12-1,13-1: F1-6: ST-15 (cc364)	2534
425	NG G40[a]	Norway	1988	carrier	B: P1.19,15-1: F5-2: ST-25 ()	2543;3259[c]
427	NG E30	Norway	1988	carrier	B: P1.21,16: F1-7: ST-44 (cc41/44)	2496
428	NG H36	Norway	1988	carrier	B: P1.5-1,2-2: F1-7: ST-47 (cc41/44)	2471
430	NG 080	Norway	1981	invasive	B: P1.7,16: F3-3: ST-32 (cc32)	2290
431	NG144/82	Norway	1982	invasive	B: P1.7,16: F3-3: ST-32 (cc32)	2290
434	NG PB24	Norway	1985	invasive	B: P1.7-2,16-7: F3-3: ST-32 (cc32)	2290
441	8680	Chile	1987	invasive	B: P1.7-2,3: F3-1: ST-32 (cc32)	2293
442	297-0[a]	Chile	1987	carrier	B: P1.19,15-1: F3-6: ST-8523 (cc254)	2505
443	3906	China	1977	invasive	B: P1.12-1,16-8: F5-9: ST-17 ()	2509
445	528[a]	Russia	1989	invasive	B: P1.22,14: F1-5: ST-18 (cc18)	2518
446	1000	Russia	1988	invasive	B: P1.5-1,10-4: F1-9: ST-20 (cc18)	2356
451	14/1455[a]	Russia	1970	invasive	A: P1.20,9: F3-1: ST-5 (cc5)	2434
466	371	India	1980	invasive	A: P1.5-2,10: F5-1: ST-1 (cc1)	2525
467	690	India	1980	invasive	A: P1.7,13-1: F1-5: ST-4 (cc4)	2288

id	isolate	country	year	disease	strain designation	rST
468	BRAZ10	Brazil	1976		C: P1.5-1,10-1: F1-10: ST-11 (cc11)	2328
488	106	Morocco	1967	invasive	A: P1.5-2,10: F5-1: ST-1 (cc1)	2433
492	79128	China	1979	invasive	A: P1.7-1,10: F5-5: ST-3 (cc1)	2527
493	322/85	Germany	1985	invasive	A: P1.5-2,10: F5-2: ST-2 (cc1)	2507
494	79126	China	1979	invasive	A: P1.7-3,10-5: F5-5: ST-3 (cc1)	2519
507	MA-5756	Spain	1985		C: P1.5,2-1: F5-5: ST-11 (cc11)	2328
597	92001	China	1992	invasive	A: P1.20,9: F3-1: ST-7 (cc5)	2501
613	Z2491	The Gambia	1983	invasive	A: P1.7,13-1: F1-5: ST-4 (cc4)	2285
638	G2136	UK	1986	invasive	B: P1.5-2,10-1: F3-6: ST-8 (cc8)	2363
639	B6116/77	Iceland	1977	invasive	B: P1.5-1,2-2: F1-4: ST-10 (cc8)	2362
640	SB25	South Africa	1990	invasive	C: P1.18-1,3: F3-9: ST-8 (cc8)	2364
641	94/155	New Zealand	1994	invasive	C: P1.5,2: F3-9: ST-66 (cc8)	2367
642	312 901	UK	1996	invasive	C: P1.5,2: F1-7: ST-8 (cc8)	2359
643	AK22	Greece	1992	invasive	B: P1.5-2,10: F3-9: ST-153 (cc8)	2361
644	L93/4286	UK	1993	invasive	C: P1.5-1,10-4: F3-6: ST-11 (cc11)	2343
645	204/92	Cuba	1992	invasive	B: P1.19,15: F5-1: ST-33 (cc32)	2300
646	400	Austria	1991	invasive	B: P1.7-2,13-2: F1-5: ST-40 (cc41/44)	2442
647	AK50	Greece	1992	invasive	B: P1.7-2,4: F1-5: ST-41 (cc41/44)	2448
648	M-101/93	Iceland	1993	invasive	B: P1.7-2,4: F1-5: ST-41 (cc41/44)	2490
649	50/94	Norway	1994	invasive	B: P1.7-2,4: F1-5: ST-45 (cc41/44)	2448
650	M40/94	Chile	1994	invasive	B: P1.7-2,4: F1-5: ST-41 (cc41/44)	2448
651	931905	The Netherlands	1993	invasive	B: P1.7-2,4: F1-5: ST-41 (cc41/44)	2448
652	N45/96	Norway	1996	invasive	B: P1.7-2,4: F1-5: ST-41 (cc41/44)	2448
653	91/40	New Zealand	1991	invasive	B: P1.7-2,4: F1-5: ST-42 (cc41/44)	2442
654	88/03415[a]	UK	1988	invasive	B: P1.7-2,4: F1-5: ST-46 (cc41/44)	2448
655	E32	Norway	1988	carrier	Z: P1.5-2,10-9: F1-5: ST-31 (cc334)	2517
656	E26[a]	Norway	1988	carrier	X: P1.18,25-1: F5-5: ST-39 (cc198)	2531
657	860060	The Netherlands	1986	invasive	X: P1.12-1,13-5: F5-5: ST-24 (cc750)	2508
659	A22	Norway	1986	carrier	W: P1.18-1,3: F4-1: ST-22 (cc22)	2415
660	71/94	Norway	1994	invasive	Y: P1.5-1,2-2: F5-8: ST-23 (cc23)	2521
661	860800	The Netherlands	1986	invasive	Y: P1.5-1,10-4: F4-1: ST-29 (cc167)	2523
698	FAM18	USA	1983	invasive	C: P1.5,2: F1-30: ST-11 (cc11)	2380
35956	255	Burkina Faso	1966	invasive	A: P1.7-2,13-1: F1-5: ST-4 (cc4)	2285
35957	890326	The Netherlands	1989	invasive	Z: P1.5-4,2-2: F3-9: ST-28 (cc103)	2406
35958	BZ 133	The Netherlands	1977	invasive	B: P1.18-1,3: F5-1: ST-10300 (cc1)	3301
35959	EG 328	Germany	1985	invasive	B: P1.22,14: F5-5: ST-18 (cc18)	2512
35960	F1576	Ghana	1984		C: P1.5,2: F1-1: ST-11 (cc11)	2328
35961	NG E28	Norway	1988	carrier	B: P1.22-1,14: F5-2: ST-26 ()	2540
35962	NG H41	Norway	1988	carrier	B: P1.5-2,10: F3-6: ST-27 ()	2412
35963	NG P20	Norway	1969	invasive	B: P1.5,2: F1-1: ST-11 (cc11)	2328
35964	SWZ107	Switzerland	1986	invasive	B: P1.22-1,14: F4-1: ST-35 (cc35)	2548

a – isolate was sequenced twice; b – isolate was extracted twice and sequenced three times; c – this genome was sequenced twice, and there was a 1 nucleotide difference, C>A, in the *rpsB* locus.

1.2. Chapter Four isolates

Three databases, two private and one public, were queried for all isolates typed as clonal complex ST-41/44 (41/44cc). The public *Neisseria* PubMLST.org isolate database was searched using the advanced queries option for all isolate records submitted to the database where an isolate was cultured, typed as 41/44cc and contained whole genome sequence data. There were 882 meningococcal genomes deposited in the public database from instances of carriage and disease. Partial typing records and sequence type (ST) profiles unassigned to a clonal complex were assessed to determine if any could belong to the 41/44cc and 11 additional isolate records were identified. The isolate record information contains known provenance, clinical and phenotypic data, and genomic profiles. The majority of the isolates (410) were part of the Meningitis Research Foundation (MRF) Meningococcal Genome Library hosted on the PubMLST.org website. Permission was requested and granted for access and analysis of an additional 256 genomic profiles contained in private projects. The UK meningococcal carriage study database and the Maiden Group database, both private, were searched using the advanced queries option for all culture records of an isolate typed as 41/44cc. The database and culture archive contained 207 isolate records; of which 12 cultures were requested from and provided by the Norwegian Institute of Public Health, Oslo, Norway. Partial typing records and ST profiles unassigned to a clonal complex were assessed to determine if any could belong to the 41/44cc and ten additional isolate records were identified. The isolates are from 1986 to 2005 and consist of samples recovered from blood, cerebral spinal fluid (CSF), and throat swabs. Isolate records were created in the public *Neisseria* PubMLST.org isolate definition database for each of the 217 isolates and included available provenance and clinical typing data so that all analysis could be done using one database. Due to the size of the dataset a complete list of the 1109 genomes can be found in Appendix A, page 106.

1.3. Chapter Five isolates

The isolates are from a subset of cultures collected and archived during the UK meningococcal carriage study between 1999 and 2001 [29]. The dataset consists of 96 isolates recovered from 48 students, 15 to 19 years of age and attending an Oxfordshire school. The isolates had ST information and were previously characterised with seven-locus MLST and fine typing antigens. Records were created in the *Neisseria* PubMLST.org isolate definition database for each isolate and contained provenance and previous clinical typing data (Table 2).

Table 2. Chapter Five isolates. Provenance and clinical typing data for 96 *N. meningitidis* carriage isolates and their ribosomal sequence types (rST). An empty bracket in strain designation indicates no clonal complex association. Related pairs are indicated in Table 4 of Chapter Five.

id	isolate	country	year	Disease	Source	strain designation	rST
890	OX9930407	UK	1999	Carrier	throat swab	NG: P1.22-11,13-22: F3-9: ST-2121 ()	3250
915	OX9931754	UK	1999	Carrier	throat swab	W: P1.18-1,3: F4-1: ST-1673 (cc22)	*I*
2913	OX0050673	UK	2000	Carrier	throat swab	NG: P1.18,25: F5-5: ST-1956 (cc198)	3008
3101	OX0051091	UK	2000	Carrier	throat swab	B: P1.19-1,30-1: F5-1: ST-2091 (cc269)	3004
3103	OX0051143	UK	2000	Carrier	throat swab	NG: P1.17-4,9: F5-5: ST-823 (cc198)	2996
3106	OX0050971	UK	2000	Carrier	throat swab	NG: P1.22-11,13-22: F3-9: ST-2121 ()	3250
3413	OX9931682	UK	1999	Carrier	throat swab	E: P1.21-7,16: F5-36: ST-1419 (cc1157)	2933
3415	OX9931776	UK	1999	Carrier	throat swab	B: P1.22,14: F5-5: ST-1644 (cc213)	3010
3419	OX9931971	UK	1999	Carrier	throat swab	E: P1.7,30: F1-2: ST-53 (cc53)	2975
3420	OX9931975	UK	1999	Carrier	throat swab	B: P1.22,14: F5-9: ST-755 (cc162)	2429
3661	OX01061392	UK	2001	Carrier	throat swab	NG: P1.18-1,3: F4-1: ST-2638 (cc22)	2415
4228	OX9931983	UK	1999	Carrier	throat swab	B: P1.19-3,15: F5-5: ST-3073 (cc35)	3235
4231	OX01061987	UK	2001	Carrier	throat swab	B: P1.19-3,15: F5-5: ST-3073 (cc35)	*I*
14595	OX9931590	UK	1999	Carrier	throat swab	NG: P1.7,30-9: F1-2: ST-53 (cc53)	2975
28885	OX9931715	UK	1999	Carrier	throat swab	E: P1.7-2,4-13: F1-5: ST-41 (cc41/44)	2991
28901	OX0050209	UK	2000	Carrier	throat swab	NG: P1.7-2,4-13: F1-5: ST-41 (cc41/44)	2991
28904	OX0050570	UK	2000	carrier	throat swab	B: P1.7-2,*deleted*: F1-5: ST-41 (cc41/44)	3114

id	isolate	country	year	Disease	Source	strain designation	rST
28907	OX0050873	UK	2000	Carrier	throat swab	B: P1.7-2,4: F1-5: ST-41 (cc41/44)	2554
28908	OX0050898	UK	2000	Carrier	throat swab	NG: P1.19-1,26: F1-7: ST-44 (cc41/44)	2978
28954	OX01061840	UK	2001	Carrier	throat swab	NG: P1.19-1,26: F1-7: ST-44 (cc41/44)	2978
28959	OX01062252	UK	2001	Carrier	throat swab	B: P1.7-2,4: F1-5: ST-41 (cc41/44)	2554
29461	OX0053039	UK	2000	Carrier	throat swab	NG: P1.21-7,16: F5-36: ST-1419 (cc1157)	2933
29462	OX0053048	UK	2000	Carrier	throat swab	ND: P1.7,30-4: F1-2: ST-53 (cc53)	2975
29464	OX01062316	UK	2001	Carrier	throat swab	NG: P1.5-1,10-1: F1-3: ST-767 (cc167)	3037
35096	OX0052328	UK	2000	Carrier	throat swab	Y: P1.5-1,10-8: F1-3: ST-767 (cc167)	3037
35885	OX9932169	UK	1999	Carrier	throat swab	ND: P1.19-1,15-11: F5-1: ST-269 (cc269)	2423
35886	OX9931422	UK	1999	Carrier	throat swab	ND: P1.5-1,2-2: F3-6: ST-254 (cc254)	*i*
35887	OX9930244	UK	1999	Carrier	throat swab	ND: P1.7,30: F1-2: ST-53 (cc53)	2984
35888	OX9932212	UK	1999	Carrier	throat swab	ND: P1.18,25: F5-5: ST-1956 (cc198)	
35889	OX9931942	UK	1999	Carrier	throat swab	ND: P1.22,14: F5-9: ST-162 (cc162)	2429
35890	OX9930103	UK	1999	Carrier	throat swab	ND: P1.18-1,3: F1-62: ST-1221 (cc22)	3069
35891	OX9930387	UK	1999	Carrier	throat swab	ND: P1.18-1,3: F4-1: ST-22 (cc22)	2982
35892	OX9930617	UK	1999	Carrier	throat swab	ND: P1.18-1,3: F4-1: ST-22 (cc22)	2982
35893	OX9930530	UK	1999	Carrier	throat swab	ND: P1.18-1,3: F4-1: ST-22 (cc22)	3120
35894	OX9930296	UK	1999	Carrier	throat swab	ND: P1.18-9,25-9: F5-5: ST-198 (cc198)	3017
35895	OX9931193	UK	1999	Carrier	throat swab	ND: P1.21-7,16: F5-36: ST-1421 (cc1157)	2933
35896	OX9932322	UK	1999	Carrier	throat swab	ND: P1.21-7,16: F5-36: ST-1421 (cc1157)	2933
35897	OX9930886	UK	1999	Carrier	throat swab	ND: P1.18-1,3: F4-1: ST-1426 (cc22)	2415
35898	OX9931800	UK	1999	Carrier	throat swab	ND: P1.19-1,3-5: F5-1: ST-269 (cc269)	*i*
35899	OX0050049	UK	2000	Carrier	throat swab	ND: P1.22,9: F5-12: ST-275 (cc269)	3011
35900	OX0050081	UK	2000	Carrier	throat swab	ND: P1.18-4,25: F4-1: ST-1136 (cc1136)	3014
35901	OX0050092	UK	2000	Carrier	throat swab	ND: P1.18-1,3: F4-1: ST-1426 (cc22)	2415
35902	OX0050158	UK	2000	Carrier	throat swab	ND: P1.22,14: F5-9: ST-162 (cc162)	3030
35903	OX0050185	UK	2000	Carrier	throat swab	ND: P1.18-1,3-16: F5-7: ST-862 (cc103)	3111
35904	OX0050256	UK	2000	Carrier	throat swab	ND: P1.22-1,14-3: F1-20: ST-35 (cc35)	2548

id	isolate	country	year	Disease	Source	strain designation	rST
35905	OX0050268	UK	2000	carrier	throat swab	ND: P1.5-1,2-2: F3-6: ST-254 (cc254)	3141
35906	OX0050325	UK	2000	Carrier	throat swab	ND: P1.22,14-32: F3-9: ST-213 (cc213)	2979
35907	OX0050360	UK	2000	Carrier	throat swab	ND: P1.7,16-2: F1-5: ST-74 (cc32)	3092
35908	OX0050377	UK	2000	Carrier	throat swab	ND: P1.18-1,3: F1-62: ST-1221 (cc22)	3069
35909	OX0050393	UK	2000	Carrier	throat swab	ND: P1.19-1,3-5: F5-1: ST-269 (cc269)	2986
35910	OX0050468	UK	2000	Carrier	throat swab	ND: P1.21-7,16: F5-36: ST-1421 (cc1157)	2933
35911	OX0050574	UK	2000	Carrier	throat swab	ND: P1.18-1,3: F4-1: ST-1673 (cc22)	*i*
35912	OX0050607	UK	2000	Carrier	throat swab	ND: P1.19-1,15-11: F5-1: ST-269 (cc269)	*i*
35913	OX0050776	UK	2000	Carrier	throat swab	ND: P1.18-1,3: F1-5: ST-22 (cc22)	3194
35914	OX0050781	UK	2000	Carrier	throat swab	ND: P1.18-1,3: F4-1: ST-2638 (cc22)	*i*
35915	OX0050852	UK	2000	Carrier	throat swab	ND: P1.5,2: F1-7: ST-60 (cc60)	3179
35916	OX0050858	UK	2000	Carrier	throat swab	ND: P1.21-7,16: F5-36: ST-1157 (cc1157)	2933
35917	OX0050902	UK	2000	Carrier	throat swab	ND: P1.22,14: F5-5: ST-213 (cc213)	2993
35918	OX0050946	UK	2000	Carrier	throat swab	ND: P1.18-1,3: F4-1: ST-22 (cc22)	2982
35919	OX0050965	UK	2000	Carrier	throat swab	ND: P1.18-1,3: F4-1: ST-22 (cc22)	2982
35920	OX0051080	UK	2000	Carrier	throat swab	ND: P1.7,30: F1-2: ST-53 (cc53)	2984
35921	OX0051099	UK	2000	Carrier	throat swab	ND: P1.22,14-32: F3-9: ST-213 (cc213)	2979
35922	OX0051106	UK	2000	Carrier	throat swab	ND: P1.7,30: F1-2: ST-53 (cc53)	2975
35923	OX0051114	UK	2000	Carrier	throat swab	ND: P1.22,14: F5-5: ST-1644 (cc213)	3010
35925	OX0051153	UK	2000	Carrier	throat swab	ND: P1.18-9,25-9: F5-5: ST-198 (cc198)	3017
35926	OX0051189	UK	2000	Carrier	throat swab	ND: P1.19-10,13-2: F5-5: ST-461 (cc461)	3001
35927	OX0051290	UK	2000	Carrier	throat swab	ND: P1.22,14: F5-9: ST-162 (cc162)	2429
35928	OX0051317	UK	2000	Carrier	throat swab	ND: P1.22,14: F5-9: ST-755 (cc162)	2429
35929	OX0051410	UK	2000	Carrier	throat swab	ND: P1.19,15-1: F1-7: ST-60 (cc60)	3089
35930	OX0051415	UK	2000	Carrier	throat swab	ND: P1.18-1,3: F4-1: ST-22 (cc22)	3120
35931	OX0052051	UK	2000	Carrier	throat swab	ND: P1.18-1,3: F4-1: ST-184 (cc22)	2415
35932	OX0052158	UK	2000	Carrier	throat swab	ND: P1.21-7,16: F5-36: ST-1421 (cc1157)	2933
35933	OX0052813	UK	2000	Carrier	throat swab	ND: P1.7,30: F1-2: ST-53 (cc53)	2984

id	isolate	country	year	Disease	Source	strain designation	rST
35934	OX0052881	UK	2000	Carrier	throat swab	ND: P1.19-1,3-5: F5-1: ST-269 (cc269)	2986
35935	OX01060129	UK	2001	carrier	throat swab	ND: P1.ND,ND: F-ND: ST-60 (cc60)	3044
35936	OX01060268	UK	2001	Carrier	throat swab	ND: P1.21-7,16: F5-36: ST-1157 (cc1157)	2933
35937	OX01060273	UK	2001	Carrier	throat swab	ND: P1.5,2: F1-7: ST-60 (cc60)	3179
35938	OX01060284	UK	2001	Carrier	throat swab	ND: P1.21,4: F1-5: ST-2196 ()	3249
35939	OX01060373	UK	2001	Carrier	throat swab	ND: P1.22,14: F5-5: ST-213 (cc213)	2993
35940	OX01060380	UK	2001	Carrier	throat swab	ND: P1.7,16-2: F1-5: ST-74 (cc32)	3092
35941	OX01060426	UK	2001	Carrier	throat swab	ND: P1.22,14-32: F3-9: ST-213 (cc213)	2979
35942	OX01060508	UK	2001	Carrier	throat swab	ND: P1.18-1,3: F4-1: ST-184 (cc22)	2415
35943	OX01060686	UK	2001	Carrier	throat swab	ND: P1.22,14-32: F3-9: ST-213 (cc213)	2979
35944	OX01060731	UK	2001	Carrier	throat swab	ND: P1.19-1,30-1: F5-1: ST-2091 (cc269)	3004
35945	OX01060795	UK	2001	Carrier	throat swab	ND: P1.5-1,10-26: F5-5: ST-1653 ()	3276
35946	OX01060836	UK	2001	Carrier	throat swab	ND: P1.19-10,13-2: F5-5: ST-461 (cc461)	3001
35947	OX01060852	UK	2001	Carrier	throat swab	ND: P1.17-4,9: F5-5: ST-823 (cc198)	2996
35948	OX01061080	UK	2001	Carrier	throat swab	ND: P1.22,9: F5-12: ST-275 (cc269)	3011
35949	OX01061099	UK	2001	Carrier	throat swab	ND: P1.22-1,14-3: F1-20: ST-35 (cc35)	2548
35950	OX01061117	UK	2001	Carrier	throat swab	ND: P1.18-4,25-49: F4-1: ST-1136 (cc1136)	3014
35951	OX01061128	UK	2001	Carrier	throat swab	ND: P1.22,14: F5-9: ST-162 (cc162)	3030
35952	OX01061141	UK	2001	Carrier	throat swab	ND: P1.18-1,3-16: F5-7: ST-862 (cc103)	3111
35953	OX01061379	UK	2001	Carrier	throat swab	ND: P1.18-1,3: F1-5: ST-22 (cc22)	
35954	OX01061911	UK	2001	Carrier	throat swab	ND: P1.7,30: F1-2: ST-53 (cc53)	3104
35955	OX01061948	UK	2001	Carrier	throat swab	ND: P1.19-1,3-5: F5-1: ST-269 (cc269)	6973
36226	OX9932316	UK	1999	Carrier	throat swab	B: P1.7-2,4: F1-5: ST-41 (cc41/44)	3114

i - incomplete profile; deleted – the PorA loop region(s) is missing

2. Isolate culture

The genomic DNA from 432 *Neisseria meningitidis* isolates was prepared from archived stocks stored in Brain Heart Infusion broth (BHI) plus 10% glycerol and frozen at -80°C. One microliter of frozen culture was incubated on labelled Columbia horse blood agar (Oxoid) plates at 37°C in an atmosphere of 5% CO_2 for 24 hours. A single colony was picked for sub-culture and incubated for 24 hours at 37°C in a 5% CO_2 atmosphere on a fresh, labelled, Columbia horse blood agar plate.

3. DNA extraction

Genomic DNA was prepared using the Wizard® Genomic DNA Purification Kit (Promega), adapting the first and last steps of the manufacturer's instructions for isolating genomic DNA from Gram-negative bacteria for use with non-broth cultures. Approximately 2 µl of plate grown subcultured cells, 18-24 hours old, were gently suspended directly to 600 µl of nuclei lysis solution in labelled, 1.5ml screw top microcentrifuge tube. Suspended cells were incubated at 80°C for 5 minutes to lyse the cells and then cooled to room temperature. 3 µl of RNase Solution was added to the cooled cell lysate; tubes were capped and inverted 6-8 times to mix thoroughly. Samples were then incubated at 37°C for 60 minutes and then cooled to room temperature. 200 µl of Protein Precipitation Solution was added to the RNase-treated cell lysate, capped and vortexed at high speed for 20 seconds to mix the Protein Precipitation Solution with the cell lysate. Once thoroughly mixed the samples were incubated on ice for 5-7 minutes. Chilled and precipitated cell lysate samples were then centrifuged at 13,000x*g* for 3 minutes. Tubes were then carefully removed from the centrifuge and checked for pelleted proteins; any flocculent tubes were spun again for an additional 3 minutes at 13,000x*g*.

The supernatant, containing the DNA, was transferred to a new, labelled, flip-top 1.5 ml DNase-free microcentrifuge tube containing 600 µl of room temperature isopropanol. Tubes were closed and gently mixed by inverting 10-12 times until thread-like strands of DNA formed a visible mass. Samples were then centrifuged at 13,000x*g* for 2 minutes. Tubes were carefully

removed from the centrifuge, and the supernatant gently poured off onto clean, absorbent paper and allowed to drain upside-down for 30-60 seconds. 600 µl of 70% ethanol at room temperature was added to the tubes containing the DNA pellets; tubes were closed and inverted 3-4 times to wash the DNA pellet. The samples were again centrifuged at 13,000x*g* for 2 minutes. Once carefully removed from the centrifuge the ethanol was aspirated from the tube and tubes were left open, but covered with a clean tissue, for 10-15 minutes to allow the residual ethanol to evaporate. To each tube 200 µl of DNA Rehydration Solution was added, and the DNA was left to rehydrate in closed tubes at room temperature for 4-6 hours, periodically mixing the solution by gently tapping the tube. After 4-6 hours, the samples were transferred to the refrigerator at 8^{o}C until quantification and purity checks were done, generally within 2-3 days.

4. DNA concentration and purity checks

The initial concentration and purity (260/280 ratio) of each DNA extract was estimated using the NanoDrop™ 1000 Spectrophotometer and 2 µl of rehydrated DNA sample at room temperature. Nucleic acids have an absorbance maximum of 260nm; and proteins, phenol or other contaminants absorb strongly at or near 280nm. A ratio of 1.8 is accepted as "pure" for DNA, and if the ratio is lower, this then indicates the presence of a contaminant in the DNA extraction. Based on the NanoDrop™ DNA-50 reading obtained 100 ng/µl dilutions were prepared in new, labelled, DNase free microcentrifuge tubes in a volume of 200 µl volume for each sample using the Promega DNA Rehydration Solution. NanoDrop measurements were not used in preparing the final concentration of whole genome DNA extractions. The NanoDrop spectrophotometry indirectly infers DNA concentration based on the absorbance of light at 260nm however other molecules such as RNA, single-stranded DNA and some proteins also absorb light at 260nm. Therefore, the smallest amount of contamination in the DNA extraction will influence the reading; making the NanoDrop prone to overestimation of DNA concentration. Final dilutions were prepared and quantified using the Qubit™ 1.0 Fluorometer Quantitation Platform and the

BR Quant-iT™ assay to confirm purity and a dsDNA concentration of between 75 ng/µl and 100 ng/µl for each sample.

Thin-wall, clear 0.5ml Qubit™ tubes were labelled and placed in a 96-well tube rack. A 1:200 working solution of dsDNA BR reagent was prepared and a 195 µl aliquot was added to each sample tube while 190 µl was added to two standard tubes. 10 µl of each standard was added into each of the labelled standards tubes, and 5 µl of each sample was added to the sample tubes. Tubes were capped and gently mixed by plate vortex for 2-3 minutes to avoid creating bubbles. Any bubbles found were gently tapped out of solution or removed with a 2 µl pipette tip. All tubes were incubated at room temperature for 2 minutes. The Qubit™ Fluorometer was calibrated using the two standards, which are a high and low concentration respectively. Once calibrated each sample was read and the values recorded. Any concentration adjustments were made, re-checked and recorded. DNA preparations were then loaded into sterile DNase-free 96-well skirted sequencing plates in 150 µl volumes and frozen at -20^{o}C before being packaged and sent for sequencing; while 50 µl was retained. The extracted stock DNA was also frozen at -20^{o}C and placed in the DNA archive.

5. Whole genome sequencing procedures

The DNA library preparation protocol involved fragmenting the genomic DNA, attaching specific adapter sequences for paired-end sequencing, and then removal of the adaptor sequences during sequence QA/C. The fragmentation of DNA was done using acoustic shearing technologies, such as the Covaris® or EpiSonic™ acoustic platforms [30, 31]. All projects used the Illumina® sequencing platform.

Procedures were performed by The Sanger Institute at the Wellcome Trust Genome Campus, Hinxton, UK. Standard Illumina® multiplex paired-end libraries were generated using 1µg of genomic DNA sheared to between 200 and 300 bp using a Covaris E210 acoustic shearing device. DNA was end-repaired using a mixture of T4 polymerase, T4 nucleotide kinase and Klenow polymerase in the presence of rATP at 20^{o}C for 30 minutes. A 3’ non-templated adenosine

residue was added by incubation with Klenow exo- and dATP at 37°C for 60 minutes. A-tailed DNA was ligated to the Illumina® multiplexing adapter oligo (Multiplexing Sample Preparation Oligonucleotide Kit, PE-400-1001) by incubation in the presence of T4 DNA ligase at 20°C for 15 minutes.

Adapter ligated DNA was amplified by PCR using Phusion® DNA polymerase (New England BioLabs) and primers from the Illumina® multiplexing sample preparation oligonucleotide kit. This created up to 12 libraries each with individual 8 bp sequence tags, with cycling as follows: 98°C 2 minute hold; 14 cycles of 98°C, 10 seconds; 65°C, 30 seconds; 72°C, 30 seconds; 1 cycle of 72°C, 5 minutes.

Before and after each of these steps DNA was simultaneously cleaned up and size selected using a 1:1 (sample:beads) ratio of AMpure beads (Beckman Coulter Genomics). All steps were performed in microtitre plates using either multichannel pipettes and or a Beckman FXp automated liquid handler.

Libraries were then pooled in equimolar ratio; with volumes of each being calculated from the outcome of quantitative PCR using the Kapa Library Quant Kit KK4835 (KAPA Biosystems). Libraries were denatured using 2M NaOH and diluted to a loading concentration of 8 pM. Twelve tagged, paired-end library aliquots were run per flowcell lane; every eighth lane contained the control genome, phiX 174. A strip tube containing the libraries was placed on to an Illumina® cBot (formerly 'cluster station') along with the flowcell and set of clustering reagents. A standard Illumina® clustering protocol was used with an additional QC step after cluster amplification; the flowcell was stained with SYBR green and then viewed with a fluorescence microscope. Passing flowcells were returned to the cBot and sequenced using the Illumina® Genome Analyser II or the Illumina® HighSeq sequencing systems.

6. *De novo* genome assembly from short-read data

Using the Illumina® 8 bp sequence tags isolates were deconvoluted on the basis of the unique tag sequences. Once the short-read sequences for each genome were extracted the paired-end files

for each library were converted to a .FASTQ file using the shuffleReads_fastq.pl script found in the VELVET package. These procedures were performed by the sequencing centre as part of the quality check process. The short-read .FASTQ files were then posted to a dedicated ftp:// site by the centre and downloaded to the Maiden Group server, at the University of Oxford for assembly. Sequences were then assembled using the *de novo* assembly algorithm, VELVET [32, 33].

The assembly parameters were optimised using a wrapper script (VELVEToptimiser.pl) written by a member of the VELVET users community (Simon Gladman, CSIRO) and supplied with the VELVET software package. The script allows a range of hash values to be assessed to produce the best assembly based on user defined parameters. The minimum sequence output size was set to 100 bp with scaffolding turned off; all other program settings were set as default and no read trimming was performed. Once generated, there was no further manipulation of the assembled contiguous sequences, also referred to as contigs. Assembled genome files were formatted in the .FASTA format and upload to the *Neisseria* sequence database and linked to their respective isolate record.

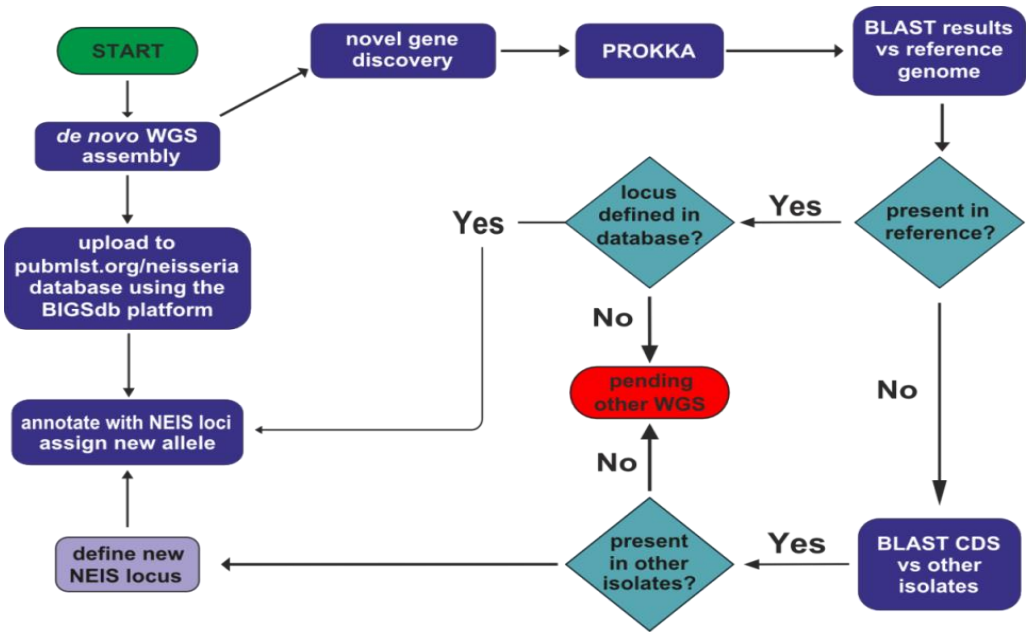
7. Population genome annotation

A database record was created for each isolate in the *Neisseria* PubMLST isolate database and included provenance, clinical and phenotype data if known. Assembled genomes were uploaded into an instance of the Bacterial Isolate Genome Sequence Database (BIGSdb) in .FASTA file format and linked to their respective isolate record [34]. Previously published reference genomes for FAM18, Z2491, G2136 and H44/76 were also uploaded to the *Neisseria* PubMLST.org website and linked to their respective isolate record. Once isolate records were populated with data and linked to their respective sequence bin (which can include whole genome data, as well as single sequence reads or antigenic peptide data) the isolate genomes were curated and analysed.

7.1. Automatic and semi-automatic annotation and curation of genomes

A population gene-by-gene annotation flowchart was developed as a guide for the curation and annotation of the genome assemblies (Figure 1). Each genome was queried against the *Neisseria* PubMLST.org sequence definition database using the BLAST algorithm to identify the relative position of loci within individual sequence contigs. The loci, when found, were annotated with a NEIS (abbreviation for *Neisseria* species gene) locus tag identifier and an allele assigned if a reference was defined within the sequence definition database. The procedure is a semi-automatic process that required a manual step to tag loci and assign alleles. The annotation of genomes in Chapter Three and Chapter Four utilized the updated BIGSdb autotagger and auto definer software to scan and annotate the genomes automatically. The auto features identify loci, known reference alleles, and new alleles greater than or equal to 98% sequence identity, automatically tag the loci and assign an allele number. This annotation and curation update decreases the manual processing of data but does not preclude manual intervention to assure data quality.

Figure 1. Population level gene-by-gene annotation flowchart. The method of gene-by-gene annotation is a general approach to the curation of large sets of genome sequences. It uses the BIGSdb tools and the PROKKA gene discovery software. New ORFs were screened against the sequence definition database and reference genomes to eliminate genes already known. Genomes were screened for the new loci to determine presence, frequency and variance. Diagram was adapted from Harrison *et al*. 2015 [35].



7.2. Manual curation and annotation of genomes

Unassigned or new alleles with less than 98% sequence identity were manually curated by alignment against nearest matches, and checked for start and stop codons using MEGA5 alignment tools [36]. Full coding sequences were assigned a new allele number for sequences with greater than 80% sequence identity. The Kyoto Encyclopedia of Genes and Genomes (KEGG), and the Pathogen Resource Integration Centre (PATRIC) website, were used to check the identification and function of those alleles with less than 80% sequence identity [37, 38].

7.3. Gene finding, Chapter Four

Individual contigs, with little or no annotation, were analysed using PROKKA (Prokaryotic genome annotation software) to discover potential novel genes and candidates were investigated using the online KEGG database and the PATRIC website [37-39]. PROKKA is a command-line software tool to annotate draft prokaryotic genomes. Protein coding genes are annotated in two stages. The program first identifies the coordinates of candidate genes using Prodigal, a gene finding algorithm [40]. The candidate gene was automatically searched, in a hierarchical manner, against the Pfam and UniProtKB online databases before domain-specific databases, protein family models were assessed, and a PROKKA annotation label was assigned to the coding region identified [41]. The coding regions of the NZ-05/33 reference genome were translated and used to differentiate the regions not present in this genome. The new coding sequences were mapped to the genome sequences using the BIGSdb BLAST search function and their prevalence in the lineage was determined.

8. Reference to *de novo* sequence validation

Genome coverage and assembly were assessed by re-sequencing four published reference genomes: FAM18, Z2491, G2136 and H44/76 [42-44]. The isolates were processed step-by-step alongside the experimental datasets as described in sections 2.1 to 2.4 above. The new draft

genome assembly of each published reference was assigned a version number in the *Neisseria* PubMLST.org sequence database and associated with its isolate record.

The BIGSdb Genome Comparator tool, implemented within the PubMLST.org website, was employed to compare the genome coverage and sequence variation of the *de novo* assembled short-read data genome to the finished reference genome. The coding sequences within the reference genome sequence bin were extracted and compared against the assembled draft genome using BLAST. Loci of the re-sequenced genomes were assessed as being allele 1, representing identity to the reference genome, or allele 2, indicating that the allele sequence did not match the reference genome sequence. The loci with conflicting allele sequences were investigated and their differences described and categorized.

9. Phylogenetic analysis

Phylogenetic methods were used to estimate the relationship of the isolates, as well as the meningococcal lineages, to each other and infer the population structure of the species. Phylogenetic networks were employed as a way of addressing events such as recombination, gene duplication and horizontal gene transfer. As a result, networks provide modelling of non-tree-like evolutionary representations of the data.

The BIGSdb Genome Comparator tool creates formatted DNA sequence output files (.aln, .FASTA, .nex), including an allelic based distance matrix, which can be used with different phylogenetic programs and analysis packages. The distance based method Neighbor-Net, as implemented in the BIGSdb software, SplitsTree, and maximum-likelihood options available in MEGA-CC were used for constructing phylogenetic networks and trees respectively [45-47]. An estimate of mean evolutionary divergence was calculated using the Kimura 2-parameter model with same rate among sites using a bootstrap of 500 replicates to correct for multiple hits and calculate standard error estimates. This model takes into account transitional and transversional substitution rates while assuming that each nucleotide frequency is the same and that

substitution rates do not vary among sites. Ambiguous pairwise positions were removed from the analysis.

Bayesian evolutionary analysis sampling trees (BEAST) was run to test evolutionary hypotheses using ribosomal protein sequences [48]. Using a range of temporally diverse times the meningococcal Lineage 3 was tested to estimate a rate and a divergence time.

10. Statistical Analysis

10.1. *p*-distance estimate

The evolutionary distance between a given pair of sequences was estimated using the MEGA sequence alignment program. This distance is the proportion of nucleotide sites, or amino acids, at which two sequences being compared are different. It was obtained by dividing the number of differences by the total number compared. The calculation does not make any correction for multiple substitutions at the same site, substitution rate biases, or differences in evolutionary rates among sites.

10.2. Chi-square 2-tailed test

To compute the statistical significance of related pairs observed in the long term carriage dataset the two-tailed chi-squared distribution was used to determine if deviations of the estimated parameter (presence of a clonal complex in a region versus presence of a related pair found in that clonal complex) were significant.

10.3. Kimura 2-parameter model

A substitution model of DNA sequence evolution was used to describe the relative rate of change within a lineage and between genomes. The model corrects for multiple hits, taking into account transitional (purine <-> purine or pyrimidine <-> pyrimidine) and transversional (purine <-> pyrimidine) substitution rates, while assuming that the four nucleotide frequencies are the same, and that rates of substitution do not vary among sites.

11. Troubleshooting

11.1. Sanger sequencing checks

PCR and Sanger dideoxy sequencing were used as needed for resolution of MLST and antigen typing loci differences. Testing used a reserved lot of DNA from the extract sent for Illumina® sequencing. Reserved DNA was amplified and sequenced using MLST, PorA VR1, PorA VR2, or fHbp specific primers using previously published primers and methods [26, 49-51]; there were no FetA VR allele conflicts. Trace files were assembled using pregap and gap4, part of the Staden sequence assembly package, and compared to the Illumina® derived sequence using the MEGA5 alignment tools [36, 52].

11.2. Mapped assembly regions

Conflicts in Illumina® derived sequence for typing loci that were not resolved by a second Sanger dideoxy sequencing were reassessed using the Bowtie short-read alignment program [53, 54]. The sequence region of a reference genome containing the target gene of interest plus 500 bp at both ends was extracted from the BIGSdb and used as the reference segment and converted to a .ebwt file. The Illumina® short-reads were converted to .SAM files and mapped against the reference segment using a randomized alignment order to avoid mapping bias. Alignments were allowed no more than two mismatches. Assembled .SAM files were visualized using the Tablet graphical package [22, 23]. The depth of coverage and polymorphic sites were identified and evaluated based on depth of coverage and prevalence of the nucleotide in a given position. Where applicable a consensus sequence was exported as a .FASTA file and uploaded to the sequence bin for the appropriate genome, the locus tagged and an allele assigned.

Chapter 3 Proof of Concept, the *de novo* assembly of short-read whole genome sequencing data

1. Introduction

While it has become clear that a species can not be adequately represented by a single genome, some researchers contend that microbial genomics will reach a point of diminishing returns; that new genomes will not serve the purpose for which they are being sequenced, to define genetic variability [55-57]. Our awareness of genomic variation in the *Neisseria meningitidis* population and the fact it is a predominantly a non-pathogenic organism, which is not represented in most genomes available for study, argues against this viewpoint [35, 44, 58]. Whole genome sequencing technology has not only increased the production of new genome sequences across the microbial domain, but those studied have only begun to illuminate the range of genomic diversity contained within each bacterial species being examined. Using the capacity of next-generation sequencing to make quickly available an extensively sampled demographic of the *N. meningitidis* population addresses the representational population gap, and provides a unique insight into the complete genome of any given isolate.

In the past decade there has been a demand for cheaper and faster sequencing methods and the demand has driven the development of many new technologies. The traditional Sanger sequencing, or first generation sequencing, has been overtaken by massively parallel sequencing technologies that facilitate high-throughput sequencing. The major competing second-generation platforms included the Illumina®, SOLiD, and Roche 454 System. However, at the time of writing the 454 platform was no longer being developed. The main challenges facing second generation sequencing technologies are for effective sample preparation, increased detection sensitivity, and cost savings; all of which are inherent in the choice of biochemistry and molecular biology methods. Technologically speaking, research and development of nucleotide sequencing over the past two decades has focused on diminishing these challenges while at the same time efficiently optimizing the production of sequence. These efforts have produced commercially available systems, available since 2005, that are capable of whole genome

sequencing both economically and faster than the chain termination methodology developed by Sanger. The application of this technology goes beyond genome sequencing, such as expression analysis, non-coding RNA characterization, and metagenomics; and each type of experiment has a tolerance for the shortcomings and advantages of each platform. The main application, however, remains the complete genetic description of an organism and how the phenotype is concomitant to the genotype.

1.1. Second generation sequencing

The technologies available and considered, for the genome sequencing in Chapters Three to Five, included the Roche 454 Genome Sequencer and the Illumina® platform. Each of these has its particular strengths and weaknesses, and these contributed to the decision for choosing the appropriate platform. Both platforms implement a common workflow, use cyclical-array sequencing by synthesis and share three key method improvements over Sanger sequencing and shotgun procedures. The bacterial cloning of DNA fragments is not required, all reactions are performed in parallel, and base reads are detected directly after incorporation. Considerations also included: (i) low-cost per isolate genome production; (ii) a flexible and sustainable platform; (iii) read length accuracy; and (iv) the availability and ease of use of post processing packages.

The major advantage of the 454 platform was the relatively long read length when compared to Illumina. However, this did not outweigh the major 454 limitations. The cost of running a 454 platform was around $60 USD per megabase with a low throughput per run, and the base signal detection system limits homopolymer repeats. The pyrosequencing chemistry did not incorporate a termination moiety to prevent multiple consecutive incorporations at a given cycle. The length of the tract must be inferred by signal intensity alone, and was, therefore, prone to higher error rates than those caused by the discrimination of incorporation versus non-incorporation [59]. Consequently, the dominant errors are insertion-deletion, rather than a substitution. This type of error is problematic for

sequencing phase variable genes and general regulatory mechanisms in prokaryotes [60]; both of which are found in meningococcal genomes. This technology is no longer being developed and the biotechnology company, 454 Life Sciences, will cease production of the platform in 2016.

The Illumina® platform supports and is compatible with most DNA preparation and library building protocols and offers the highest throughput of the two platforms. Illumina® sequencing chemistries are the limiting factor for this platform; however, continual developments of protocols and reagents have increased read lengths and have progressed from 54 base reads in 2009 to 250 base reads in 2015. The reason for the shorter read lengths is the incomplete cleavage of the fluorescent labels or terminating moieties that in turn cause signal decay and dephasing. That means the dominant error type is nucleotide substitution during the base calling. These error rates are less than 2.0%, and can be reduced to 0.1% when using paired-end sequencing and quality metrics [59]. In addition, the cost is significantly less, $2 USD per megabase, and in this regard, its use for large bacterial population studies is more economically feasible; and as costs continue to reduce it is a financially sustainable platform.

1.2. Assembly of genomes from second generation sequencing data

One of the biggest challenges post whole genome sequencing is coping with the amount of sequence data produced. Much of the data is in a short-read format and assembling at best a single contiguous genome, or at the very least a set of contiguous sequences is time-consuming. Several bioinformatics tools have been developed to deal with the copious amounts of sequence data being generated. They fall into one of two broad categories, *de novo* assembly and reference mapping (also known as alignment mapping).

1.2.1. Reference mapping

The mapping of short-read sequence to a reference sequence uses, in general, the Burrow-Wheeler transformation that facilitates linear-time alignment to a larger reference sequence. These include programs such as MAQ, BWA, Bowtie and many others [53, 61, 62]. Mapping has been effectively used to investigate closely related isolates. Nevertheless, this approach suffers from some weaknesses [63]. It is necessary to have a high-quality reference sequence to make the comparison, which is not always available, and each sequence polymorphism must be confirmed independently. It is also important to choose parameters that will maximize the number of reads that 'match' while minimizing the number discarded due to either no match or matches of n+1 across the assembly. Consequently, the process is computationally intensive and cannot detect variation in, or even the presence of, genes that are not present in the reference genome. Further, the density of sequence polymorphism in all but a minority of bacteria is such that this approach is impractical for the study of isolates that have not already been demonstrated to be monophyletic. Despite its potential issues, reference mapping can, however, be effectively used to reassemble short-read data to investigate missing loci or poorly assembled draft genomes.

1.2.2. *De novo* assembly

De novo sequence assembly is a process whereby individual sequence reads are merged to reform the genome into one continuous or several disconnected contiguous sequences, or contigs. The two main algorithm types employed for this are known as greedy methods and graph-based methods [64, 65]. Software packages such as PHRAP, Phusion, and TIGR use greedy methods and were designed for long sequence reads (1kb or more). The method constructs sequences incrementally by picking the highest scoring overlap, merging the overlapping fragments, and adding the resulting new

fragment to the pool of sequences. This process is repeated until no more merges can be carried out [66-68]. The seed-and-extend version of the greedy method was specifically designed for short-read assemblies and includes the programs SHARCGS, SSAKE and QSRA [69-71]. The CELERA and Edena programs, aimed at assembling long and short reads respectively, are examples of greedy seed-and-extend methods [72, 73]. The greedy methods are computationally intense and for high throughput genome sequence assembly the speed and time for assembly are sub-optimal. These assembly options were not considered and will not be further discussed.

There are two graph–based methods of assembly: the overlap-layout-consensus (OLC) and the sequence by hybridization (SBH). The graph methods both start by pre-processing the sequence reads to find pair-wise overlaps that are used to create an unweighted string-graph. The string-graph, also known as a path, ultimately corresponds to a contiguous sequence assembly. The underlying graph of OLC method uses nodes (sequence reads) and edges (the overlaps) to determine a simple path that links together all nodes. This approach requires overlap scoring and can be computationally intense as it scales with the square of the number of reads. The increased computational requirements mean they are not well suited for short-read sequence assembly that produces millions of sequence reads. Algorithms such as VELVET, ABySS, SOAP denovo, use SBH method to generate consensus sequences from the short sequence reads [32, 74, 75]. More recently SPAdes was introduced which combines modified EULER-SR and modified VELVET design [76, 77]. The SBH approach employed dually represents overlaps by nodes and the reads with an overlap by edges to the corresponding node. What this amounts to is a partitioning of reads into overlapping sub-strings (sequences) and building a de Bruijn graph that is linear even with very high coverage. While these assemble regions of genomic complexity poorly, especially tracts of repetitive DNA sequence that are substantially longer than the average read lengths, the approach can generate multiple contiguous sequences of

varying length for each genome. The sequence data assembled corresponds to the majority of the genome, especially when 'paired-end' sequencing strategies are used, 'high-quality draft' bacterial genomes can be assembled effectively in this way [78-83]. This method represents an alternative and more broadly applicable approach for high-throughput, population studies.

2. A Proof of Concept: quality *de novo* assembly from short-read sequencing data

For many years research on the meningococcus has been at the forefront of work on the population biology of bacteria and pathogens in particular [84]. While the issues surrounding second generation sequencing and assembly will remain a concern for a while longer, steps can be taken to describe, assess and benchmark the use of these methods as viable tools. In the published work, Bratcher HB, Corton C, Jolley KA, Parkhill J, Maiden MC. 'A gene-by-gene population genomics platform: *de novo* assembly, annotation and genealogical analysis of 108 representative *Neisseria meningitidis* genomes.' BMC Genomics 2014, 15:1138, a two-year investigation into the use of short-read genome sequencing and *de novo* genome assembly. HBB designed the study, carried out the analysis and wrote the manuscript; CC assembled the draft genomes; KAJ designed and developed the BIGSdb platform used for the analysis; JP collaborated in the sequencing of the genomes; MCJM conceived the approach of high-throughput whole genome gene-by-gene analysis, approved the study outline, manuscript and contributed to the final editing of the manuscript.

BMC Genomics

METHODOLOGY ARTICLE Open Access

A gene-by-gene population genomics platform: *de novo* assembly, annotation and genealogical analysis of 108 representative *Neisseria meningitidis* genomes

Holly B Bratcher[1], Craig Corton[2], Keith A Jolley[1], Julian Parkhill[2] and Martin CJ Maiden[1*]

Abstract

Background: Highly parallel, 'second generation' sequencing technologies have rapidly expanded the number of bacterial whole genome sequences available for study, permitting the emergence of the discipline of population genomics. Most of these data are publically available as unassembled short-read sequence files that require extensive processing before they can be used for analysis. The provision of data in a uniform format, which can be easily assessed for quality, linked to provenance and phenotype and used for analysis, is therefore necessary.

Results: The performance of *de novo* short-read assembly followed by automatic annotation using the pubMLST.org *Neisseria* database was assessed and evaluated for 108 diverse, representative, and well-characterised *Neisseria meningitidis* isolates. High-quality sequences were obtained for >99% of known meningococcal genes among the *de novo* assembled genomes and four resequenced genomes and less than 1% of reassembled genes had sequence discrepancies or misassembled sequences. A core genome of 1600 loci, present in at least 95% of the population, was determined using the Genome Comparator tool. Genealogical relationships compatible with, but at a higher resolution than, those identified by multilocus sequence typing were obtained with core genome comparisons and ribosomal protein gene analysis which revealed a genomic structure for a number of previously described phenotypes. This unified system for cataloguing *Neisseria* genetic variation in the genome was implemented and used for multiple analyses and the data are publically available in the PubMLST *Neisseria* database.

Conclusions: The *de novo* assembly, combined with automated gene-by-gene annotation, generates high quality draft genomes in which the majority of protein-encoding genes are present with high accuracy. The approach catalogues diversity efficiently, permits analyses of a single genome or multiple genome comparisons, and is a practical approach to interpreting WGS data for large bacterial population samples. The method generates novel insights into the biology of the meningococcus and improves our understanding of the whole population structure, not just disease causing lineages.

Keywords: *Neisseria meningitidis*, *de novo* assembly, BIGSdb, Gene-by-gene analysis, cgMLST, rMLST, rST, Bacterial population genomics

* Correspondence: martin.maiden@zoo.ox.ac.uk
[1]Department of Zoology, University of Oxford, Oxford, UK
Full list of author information is available at the end of the article

Background

The widespread application of parallel high-throughput 'next generation' sequencing (NGS) technologies has made whole genome sequence (WGS) data available for tens of thousands of bacterial isolates [1]. Increasingly, these data are publicly available only as depositions in short-read sequence archives: in December 2013 the European Bioinformatics Institute (EBI) Sequence Read Archive (SRA), contained more than 100,000 bacterial WGS records, over 90% of which comprised millions of short sequence reads each of fewer than 200 bases in length. These data represent a major resource for studies of bacterial diversity, evolution and function; however, as the throughput of genome finishing and annotation technologies has not kept pace with sequence determination, the genomes have to be reassembled to be interpreted. Typically, this is done either by mapping to a reference sequence or by *de novo* assembly to generate draft genomes comprising multiple contiguous sequences (contigs).

The approach of mapping short-read sequences to a reference sequence has been effectively used to analyse WGS data from closely related isolates in numerous studies [2-9], especially by using the data obtained to reconstruct genealogies based on phylogenetic trees. This approach has a number of limitations, including: the necessity for a high-quality reference sequence with which to make the comparison; variation in sequence not present in the reference cannot be detected; the approach is poorly scalable; analyses typically have to be re-run as new genomes are obtained; and finally, the density of sequence polymorphisms in the majority of bacterial populations is such that this approach is not feasible for the study of isolates that are not genetically closely related. The use of *de novo* assembly methods represents an alternative, more broadly applicable approach, with assemblers based on de Brujin graphing being widely used as they deal effectively with large volumes of data [10,11] and can assemble short-read sequences of fewer than 100 bases in length into contigs that contain the majority of the genome. Further, when paired-end sequencing strategies are employed, 'high quality draft' bacterial genomes can be assembled [2-9,12-14]. Once they have been assembled, these sequences can be annotated by comparisons to known genes or genome databases [15], using an approach similar to that used in multilocus sequence typing (MLST), which has been widely employed for sequence-based analyses at the population scale since 1998 [16]. The Bacterial Isolate Genome Sequence Database (BIGSdb) platform provides this functionality for WGS data [17].

Neisseria meningitidis, the meningococcus, is a pathogen of global significance and an informative model organism for investigating the relationship between genotype and phenotype, as it is highly diverse phenotypically and genotypically [18]. Due to the importance of the disease, the most studied meningococcal phenotype is the propensity to invade, although most episodes of meningococcal infection result in asymptomatic carriage, which typically occurs in 10-20% of the human population [19,20]. Only a very small number of infections result in devastating and rapidly progressing disease, in the form of septicaemia, meningitis, or both. For reasons that are incompletely understood, some meningococcal genotypes are much more likely to cause invasive disease than others. Nucleotide sequence-based typing, especially MLST and antigen sequence typing (AGST), have established that these genotypes correspond to certain genealogies, known as the 'hyperinvasive lineages' [21]. There are a number of factors known to contribute to the hyperinvasive phenotype, particularly the possession of certain capsular polysaccharides, but species-level comparisons suggest that the majority of the pan-genome is widely shared among invasive and non-invasive genotypes. This has led to the conclusion that the ability to cause invasive disease is both polygenic and different among hyperinvasive lineages [22-24], but the determinants associated with particular lineages remain poorly defined. Comparative WGS of meningococcal isolate collections that include representative disease and carriage isolates have the potential to define the genetic differences which determine the hyper invasive phenotypes.

Here, WGS data collected by NGS technology were investigated with *de novo* assembly and population annotation to characterize 108 diverse meningococcal genomes, including the major hyperinvasive lineages observed worldwide over the last 60 years. The draft genomes were analysed for accuracy and coverage using the BIGSdb platform [17] which enabled comparison with 24 antigen and MLST typing loci previously characterised with Sanger sequencing and four finished reference genomes, cross-validating these technologies. These data established the robustness and reliability of using *de novo* draft genomes for a population-wide level of analysis for meningococcus genomes and presented a WGS description of the major hyperinvasive lineages, providing insights into their structure, evolution, and function.

Results

Genome assembly

Short-read sequences were assembled into draft genomes using Velvet [25] and VelvetOptimiser [26] programs, using 54 or 76 base read files. The sum total length of assembled contigs ranged from 1,975,180 bp to 2,211,536 bp and had a G + C content between 51-52%, consistent with previously finished meningococcal genomes (see Additional file 1: Table S1). Assemblies consisted of 291 to 407 contigs, with a mean of 367. The average N50, a value that represents the length at which contigs of equal or longer length contain at least 50% of the assembled sequence,

across all genomes was 19,495 bp. This statistic provided an indication of the total genome coverage; however, it was not a measure of genome assembly quality. Overall, a higher k-mer setting for the assembly was associated with the higher N50 values and, within the bounds of the read length, assembled repeat regions within the genome that were under 100 bp in length.

All the *de novo* assemblies consisted of contigs terminating at repetitive sequence regions longer than the read length of 54 or 76 bases and these termination regions contained a higher read depth than the preceding regions. A change from Taq polymerase to Phusion® high-fidelity DNA polymerase affected the assembly statistics (Table 1: groups I and J), and increased the average longest contig length by 41% (Table 1: groups J and K). Improved sequencing chemistry that generated read lengths of 76 bases resulted in a 20% decrease in the average number of contigs per genome (Table 1: group K). The combined use of the new DNA polymerases and improved chemistry resulted in a decreased number of contigs and a reduction of incomplete coding sequence (CDS) of approximately 61% (Table 1: group K).

Sequencing accuracy

The WGS assemblies were compared to previously determined dideoxy (Sanger) sequence results at the MLST [16], eMLST [27,28], *porA VR1* and *VR2* [29], *fetA* [30] and *fHbp* [31] loci found throughout the genome (Figure 1). There were thirty-four sequence discrepancies (1.1% of CDS) between the Illumina *de novo* assembled and Sanger sequenced alleles, from 20 of the 108 genomes (Table 2). The number and distribution of sequence changes found in the resequencing experiments enabled the likely reasons for the discrepancies to be identified. In the majority of cases these could be attributed to either editing or labelling problems in the original Sanger sequencing experiments with only four instances that were a direct consequence of the assembly of the short-read sequences by the Velvet algorithm. The four MLST profiles affected by trace file editing errors maintained their original clonal complex assignment; however, their sequence type (ST) was amended to a new designation as a consequence of this work. In summary, the errors in the original Sanger experiments were due to: eleven trace file editing errors; 19 samples mislabelled during Sanger sequencing; and four occurrences of Velvet mis-assembly caused by short tandem repeat (STR) regions.

The four draft sequences for which finished genome sequences were available (H44/76, FAM18, Z2491, and G2136) were compared to the published closed sequences using the BIGSdb Genome Comparator tool. Sequence discrepancies were found between all four resequenced draft genomes and their respective finished reference genome. The H44/76 and G2136 reference genomes, created with Roche 454 technology and finished using capillary sequencing, had sequence differences in thirty hypothetical proteins, thirty-five annotated CDS, nine pseudogenes and five putative proteins, a total of 79 loci for these two published genomes (see Additional file 2: Table S2, sections B-E). FAM18 and Z2491 reference genomes,

Table 1 Velvet assembly statistics of 108* genomes analyzed at 1605 core meningococcal loci

Multiplex group (number of libraries)	Sequence read length	Most common k-mer	Average contig count	Average N50	Average longest contig	Average genome length	Average number of core loci identified	Average number of incomplete loci found
A (12)	54	39	374	18450	66014	2088067	1603	45
B (12)	54	35	407	16252	63032	2073517	1603	39
C (12)	54	39	339	20375	76751	2087187	1602	43
D (12)	54	41	322	22022	79870	2086109	1604	34
E (12)	54	37	396	15813	59228	2108812	1602	43
F (11)	54	41	345	18364	66689	2085872	1603	40
G (12)	54	39	369	16595	61439	2107242	1602	46
H (12)	54	37	329	20557	77125	2095768	1603	46
I (11)	54	39	386	13842	53816	2085635	1604	54
J (7)	54	41	403	22586	104539	2109587	1600	37
K (7)	76	63	295	31378	122718	2143841	1604	26
Average *de novo* Assembly Statistics	**n/a**	**39**	**360**	**19658**	**75566**	**2097421**	**1603**	**41**

*The total number of genomes analysed is 120 and includes: 5 genomes sequenced a second time using 54 base reads (J) and 7 genomes sequenced a second time using 76 base reads (K).

Two genomes failed to sequence in their original groups (F and I), these genomes were subsequently rerun in group J.

The increase in read length used for multiplex group K produced larger than expected assembly improvements. A significant drop in the number of contigs and a corresponding increase in the N50 value were achieved with the relatively small 22 base read increase. Therefore, additional increased base read lengths should continue to increase the coverage of long repeat regions and decrease the number of contigs per assembly.

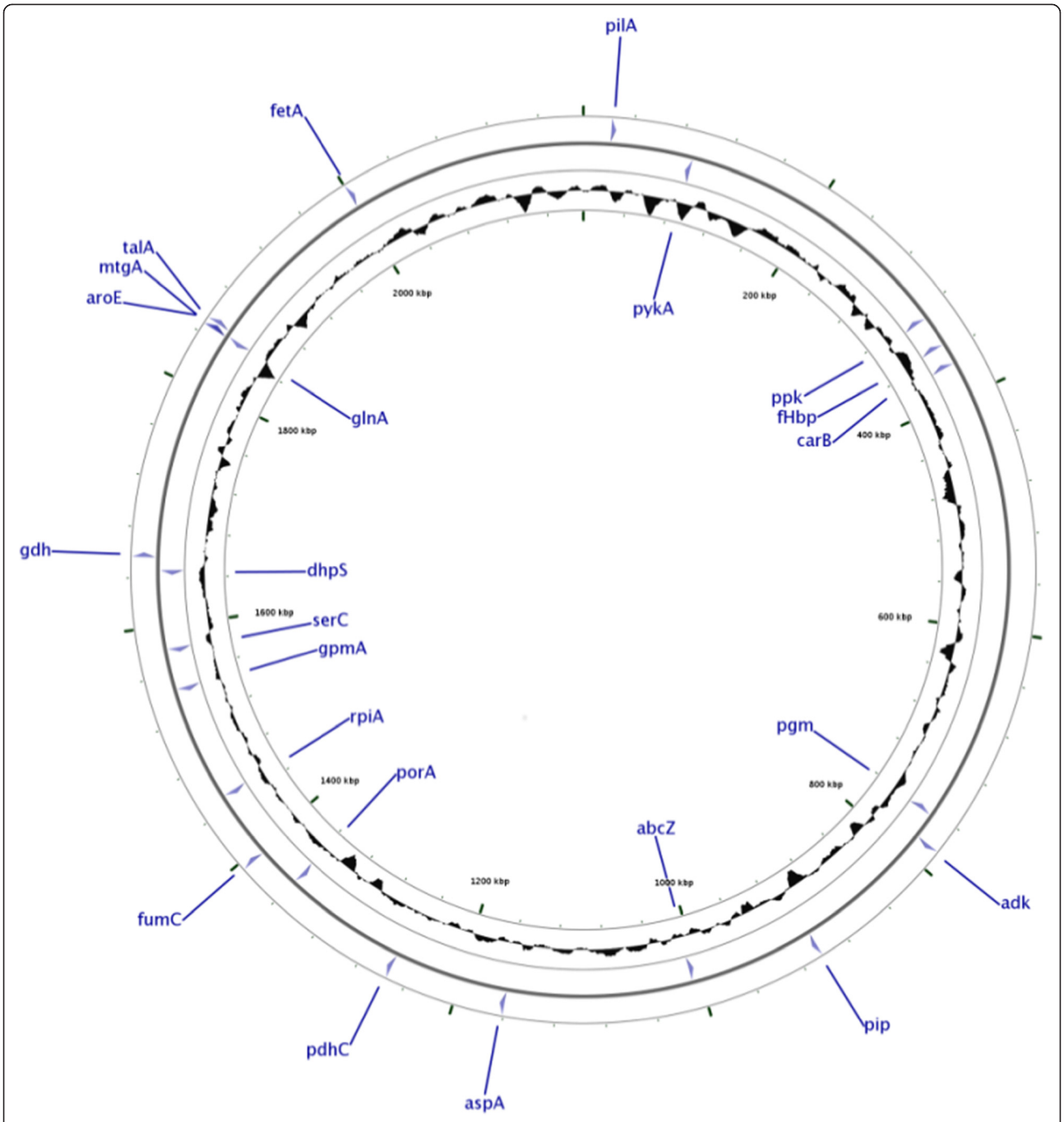


Figure 1 Location of eMLST and antigen genes within the meningococcal genome. CGView map of the *Neisseria meningitidis* reference genome, FAM18, showing the placement of the conventional and extended MLST loci and the 3 antigen genes (4 typing fragments) used to assess sequence accuracy of the *de novo* high-throughput assembly method across the genome.

obtained using ABI 3700 and a combination of ABI373 and 377 respectively, had sequence discrepancies among twenty-two annotated CDS, ten pseudogenes, five putative protein sequences and fourteen hypothetical proteins; totalling 51 loci of the published CDS sequences for these genomes. The majority of these CDS affected (69.1%) had a single nucleotide change each and the remaining 30% had two or more nucleotide changes. The differences were categorized as non-synonymous or synonymous amino acid changes (see Additional file 3: Table S3). Differences caused by assembly failures (24 loci) or paralogous loci (23 loci pairs) contained cross identified

Table 2 Comparison of Sanger derived MLST and AGST loci to their respective *de novo* assembled genome

Typing locus	Original Sanger derived allele	Illumina derived allele	Retested Sanger derived allele	Number of bases, likely cause of discrepancy
MLST				
[f]*adk*	1	10	10	9, mislabelled
[f]*gdh*	8	34	34	1, editing error
[f]*pdhC*	1	547	547	1, editing error
	14	60	60	1, editing error
eMLST				
[f]*pykA*	9	5	5	35, mislabelled
[f]*ppk*	3	17	17	2, editing error
	12	1	1	19, mislabelled
[f]*pip*	2	1	1	6, mislabelled
	4	19	19	2, editing error
[f]*dhpS*	11	87	87	10, mislabelled
	10	86	86	1, editing error
	33	42	42	1, editing error
[f]*aspA*	8	78	78	6, mislabelled
[f]*gpm*	7	11	11	14, mislabelled
[f]*rpiA*	1	18	18	2, editing error
[f]*serC*	4	56	56	2, editing error
	29	56	56	20, mislabelled
	4	5	5	8, mislabelled
	8	56	56	8, mislabelled
[f]*talA*	7	20	20	3, editing error
	2	6	6	21, mislabelled
Antigens				
PorA VR1	7	18-1	18-1	30, mislabelled
PorA VR2	1-1	1-2	1-2	3, mislabelled
	16	3	3	30, mislabelled
	15	15-1	15-1	1, editing error
	14-1	14	14	3, repeat sequence*
fHpb	25	5	5	198, mislabelled
	39	16	16	5, mislabelled
	24	14	14	196, mislabelled
	5	22	22	5, mislabelled
	35	32	32	51, mislabelled

*repeat sequence compression during assembly; this occurred in 4 different isolates.

reads (see Additional file 2: Table S2, sections E & F). Paralogous gene cross-identification occurred most often in CDS annotated as hypothetical proteins, a total of ten. These, plus six additional paralogous loci, were manually curated and defined using up- and down-stream sequence in order to enable the BIGSdb scanning function to correctly distinguish the divergent regions of the paralogous genes without manual curation. A list containing the identification of all CDS with sequence differences, and those loci missing in the draft genomes was generated (see Additional file 2: Table S2, section A-E).

The BIGSdb Genome Comparator tool was used to assess genome coverage of the resequenced reference genomes (Table 3a and b). The Z2491 draft genome contained 1872 of the 1876 (99.8%) CDS present in the reference genome, of which 51 (2.7%) were partial sequences, that is the locus was found at the end of a contig and therefore incompletely assembled. The four loci not identified in the draft genome

Table 3 Re-sequenced genome comparisons sequence differences identified among four re-sequenced genomes and their respective finished sequence

a.			Missing Sequence		Failed assembly of repeat sequence tracts		Paralogous cross identification		Sequencing Discrepancy	
Isolate	**CDS count**	**Total number of bases**	**CDS count**	**total number of bases**	**CDS count**	**number of bases[§]**	**CDS count**	**total number of bases[§]**	**CDS count**	**Number of bases affected**
Z2491	1876	1693839	4 (0.2%)	5408	2 (0.2%)	3039	5 (0.3%)	4737	12 (0.6%)	32
FAM18	1914	1767562	9 (0.5%)	21895	6 (0.3%)	9070	8 (0.4%)	6990	9 (0.5%)	24
G2136	1904	1718346	7 (0.4%)	14484	10 (0.5%)	13125	12 (0.6%)	4974	25 (1.3%)	90
H44/76	1975	1784201	8 (0.4%)	16716	7 (0.4%)	14458	17 (0.9%)	4824	25 (1.3%)	76

b.				
Isolate	**Loci Present**	**Loci with identical sequence match**	**Loci with nucleotide sequence discrepancy**	**Loci that are present but incomplete**
Z2491	1872 (99.8%)	1801 (96.2%)	19 (1.0%)	51 (2.7%)
FAM18	1905 (99.5%)	1775 (93.2%)	23 (1.2%)	107 (5.6%)
G2136	1897 (99.6%)	1757 (92.6%)	47 (2.5%)	93 (4.9%)
H44/76	1967 (99.2%)	1821 (92.6%)	49 (2.5%)	97 (4.9%)

§ For each CDS that had either a failed assembly or paralogous cross-identification error the entire CDS length was counted as affected.
Sequence differences were identified using the BIGSdb Genome Comparator tool. All transposase CDS were removed from the analysis. The Z2491 and FAM18 were originally sequenced and finished using ABI373 and 377 or ABI 3700 methods in 2000 and 2007 respectively, and the H44/76 and G2136 genomes were originally sequenced and finished in 2011 using Roche 454 FLX and capillary-based sequencing.

belonged to four non-contiguous hypothetical proteins: NMA0440, NMA1192, NMA1307 and NMA1860, totalling 5408 nucleotides. Comparison of the FAM18 draft to the reference genome identified 1905 of 1914 (99.5%) published CDS and 107 (5.6%) were also incomplete. The nine CDS (0.5%), 21,895 bases, not found in the draft genome included three *mafB2* genes (NMC0597, NMC1790, NMC2084), a putative *frpC* pseudogene (NMC1345) and an unidentified pseudogene (NMC0296) in addition to genes encoding: a single MafB protein fragment (NMC2090), the iron-regulated FrpC protein (NMC0527), a pilus secretin (NMC0408), and putative cell-surface protein (NMC1668). The BIGSdb Genome Comparator tool identified 1967 of 1975 published CDS in the H44/76 draft genome (99.2%) and 1897 of 1904 published CDS of the G2136 draft genome (99.6%). H44/76 resequenced missing 16,716 bases that included genes encoding: three TspB family proteins (NMBH4476_0598, NMBH4476_0681, NMBH4476_1698), two iron-regulated proteins FrpA and FrpC (NMBH4476_0805, NMBH4476_1605), NlpC/p60 family protein (NMBH4476_1938), a putative Caudovirus prohead protease (NMBH4476_0570) and a hypothetical protein gene (NMBH4476_1682); and the resequenced G2136 genome was missing genes for three hypothetical proteins (NMBG2136_0443, NMBG2136_0446 and NMBG2136_0522), a membrane protein (NMBG2136_1025), a fimbral protein precursor (NMBG2136_0028) and two FrpC iron-regulated proteins (NMBG2136_0523 and NMBG2136_1306), covering 14,484 bases. All four resequenced genomes were also mapped to their respective finished genomes to look for the missing loci. In all instances the loci were present in the short-read sequence data, however read depth was low (2-5×) and therefore could not be assembled by the parameters set for the Velvet assembly.

While it is possible to revise the assembly parameters and recover some of the missing data in the assemblies, this would potentially be at a cost to the overall quality of the assembly by swapping specificity and sensitivity and could in fact reduce the N50 value, therefore this option was not implemented for this analysis. Technically, the foundations resulting in the underrepresentation of these regions in the subsequent sequence reads have many sources: for example GC bias affects the stability of the DNA strand which could influence the read ability or modify the probability of a fragmentation. It has been shown that optimized or PCR-free protocols reduce GC bias affects [32-35] and if these genomes were resequenced using a PCR-free approach it is possible the overall genome coverage would increase.

Gene-by-gene annotation

All of the draft genome assemblies were annotated using a gene-by-gene approach using the BIGSdb platform as described previously [17,36]. Each genome was scanned against defined loci contained in the PubMLST *Neisseria* sequence definition database using the default parameters (70% minimum identity; 50% minimum alignment; and a BLASTN word size of 15). Alleles previously identified were assigned an allele number automatically, in a process referred to as 'tagging', and new alleles were manually curated and submitted to the sequence definition database for allele number assignment. The genome data were subsequently rescanned to assign the new alleles to the

respective genome in which it was found. Partially assembled loci, those found at the end of a contig, were tagged as present in the genome but flagged as incomplete. The average number of incomplete coding sequences (CDS) found per genome was 41 (see Additional file 1: Table S1). BIGSdb also identified sixteen paralogous CDS pairs, these included six recognized CDS (including two ribosomal protein genes), ten hypothetical genes and one putative lipoprotein (see Additional file 2: Table S2, section F).

The meningococcal core genome

Comparison of the four finished genomes identified 1760 CDS (89.1%) that were present in all genomes. The list was refined by determining gene presence of the 1760 CDS in two additional finished genomes [37]. The *de novo* assembled draft genomes were then also compared to identify CDS present in 95% of the genomes, to account for the draft nature of the genomes and genes missing rarely in isolates. All genomes were unique in terms of gene content, but 1605 CDS were present in at least 95% of the isolates. These were categorised as 'core loci' for this dataset and were used to search the functional pathway database KEGG [38,39] to determine the function of the gene product of each CDS (Figure 2). Only 37% of the core loci (597 in total) had an enzyme commission number (EC) assigned. Those with an assigned EC indicated the presence of two environmental processing pathways, four genetic information processing pathways and 12 metabolic functional pathways. Almost three-quarters

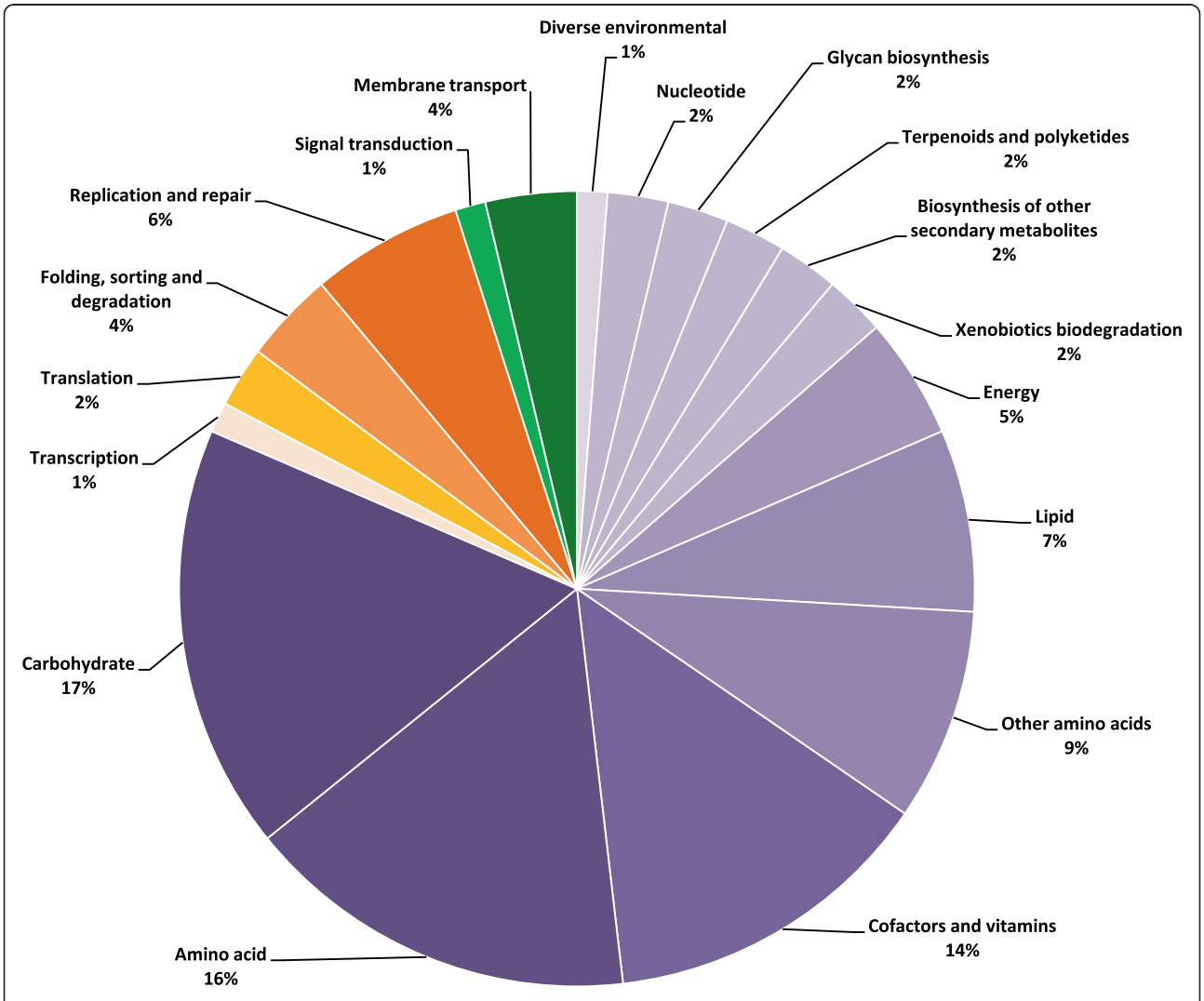


Figure 2 KEGG functional and informational processing pathways identified in the meningococcal genome. Loci from the core gene list were used to search the pathway database KEGG for functional and informational pathways. A total 537 loci of the 1605 core genes have assigned Enzyme Commission numbers (EC). The figure shows the breakdown of the genes in to three main groups and specific associated pathways: orange – genetic information processing (4 pathways, 20%), purple – metabolism (12 pathways, 73%), green – environmental information processing (2 pathways, 6%).

(72%) of the EC identified genes were involved in metabolic pathways, 20% were genetic information processing pathways and 6% were identified as being involved in environmental information processing.

Genealogical analyses

Distance matrices based on the number of locus differences were calculated among the 108 isolates and 20 previously published genomes [37] with the Genome Comparator tool and represented with NeighborNet graphs for the core loci (cgMLST) and the ribosomal protein (rMLST) genes [40] (Figure 3a and b). The rMLST and cgMLST schemes clustered 101/128 (71%) of the isolates into ten distinct lineages that corresponded to the major invasive clonal complexes identified by seven-locus MLST [27] and were consistent with previously defined clades [37].

The higher resolution of rMLST and cgMLST, as opposed to the seven- or twenty- locus MLST, also resolved the substructure characteristic of lineage 3 (ST-41/44 complex). This lineage sub-structure is captured in MLST by the designation of two central genotypes that are differentially associated with invasive disease, and at the sequence type level share five of the seven MLST alleles [27,41]. Analysis of this lineage also showed that isolates associated with the ST-41 belonged to a well-defined monophyletic lineage, while the ST-44 associated isolates were a more diverse but distinct lineage. Further exploration of this complex is necessary to more fully define the relationships within this clade and the variable pathogenic nature associated with each group. The association of capsule loci with the lineage 11 (ST-11 complex), and in lineage 8 (ST-8 complex), at the cgMLST level (1605 core genes) shows the serotype B and C associated genomes on different branches, and only lineage 11 (ST-11 complex) maintains this separation at the rMLST level (53 ribosomal genes). The remaining lineages did not have sufficient numbers to clearly differentiate capsule associations; and additional studies with larger strain collections will be required to make these associations more distinctly.

Four sets of lineage specific draft genomes, thirty-four in total, were assessed for genome coverage using one of four reference genome annotations and the BIGSdb Genome Comparator tool. Each of the four sets of *de novo* assembled genomes contained over 98% of the CDS defined by their closest reference genome. Seven isolates in the collection belong to lineage 5 (ST-32 complex) and were compared to the H44/76 reference genome. All 1976 of the H44/76 CDS were identified across the seven *de novo* assembled genomes. There was an average of 1951 CDS (98.7%) identified per genome. Seven isolates belonging to lineage 8 (ST-8 complex) were compared to the G2136 reference genome. All 1911 G2136 CDS were identified across the *de novo* assembled genomes, with an average of 1865 CDS (97.6%) found per genome. A further ten isolates belonged to lineage 11 (ST-11 complex) and were compared to the FAM18 genome sequence. The comparison identified 1912 (99.8%) of the 1915 FAM18 genome CDS across the ten genomes, with an average CDS count of 1879 CDS (98.1%) per genome. Ten isolates also belonged to the lineage 4 (ST-4 complex) and were compared to the Z2491 genome, identifying 1936 (99.9%) of 1937 CDSs across all ten genomes. Each genome had an average of 1899 CDS (98%) per genome and the only CDS not found in all ten of the lineage 4 genomes was the coenzyme A gene, *coaD*.

Discussion

Exhaustive comparison of bacterial genomes, including all sources of genetic variation (i.e. individual sequence polymorphisms, insertions, deletions, and rearrangements at all scales) requires complete, closed ('finished') genomes. The majority, if not all, of short-read WGS data generated to date with NGS technology are incapable of meeting this ideal without extensive additional data combined with manual assembly and curation [42,43]. There are many questions in bacterial biology, however, which can be adequately addressed with population genomic approaches that employ subsets of the genome [44], such as MLST (Figure 1), rMLST (Figure 4) and cgMLST; and for these analyses NGS datasets provide a rich source of information [15]. For such analyses to be robustly conducted, however, it is necessary to establish an analysis paradigm that interprets data consistently within known parameters of completeness and accuracy [45]. Here we demonstrate how bioinformatics tools that are freely available and widely understood can be combined to interrogate NGS data using the example of the diverse human pathogen *Neisseria meningitidis* [17]. The data and analyses are easily accessed through the PubMLST *Neisseria* website [http://pubmlst.org/neisseria/].

Although the sequence read lengths employed here were relatively short (54-76 bp) [42] and the meningococcus has a complex genome comprising many short tandem repeats (STR) and homopolymeric tracts [46-48], the Velvet algorithm was consistently capable of assembling the majority of protein coding sequences (over 1850 complete loci per genome) to extremely high levels of accuracy. Indeed, where comparable data were available for genes previously used for sequence based-typing, the majority of the discrepancies were due to errors in the editing or labelling of the specimens used in the original Sanger sequences, and the remaining, the result of STR sequence compression during assembly [49]. Once these errors had been taken in to account, the two approaches were in complete agreement. There was also very good agreement with complete reference genomes, although this depended on the read length of the short-read

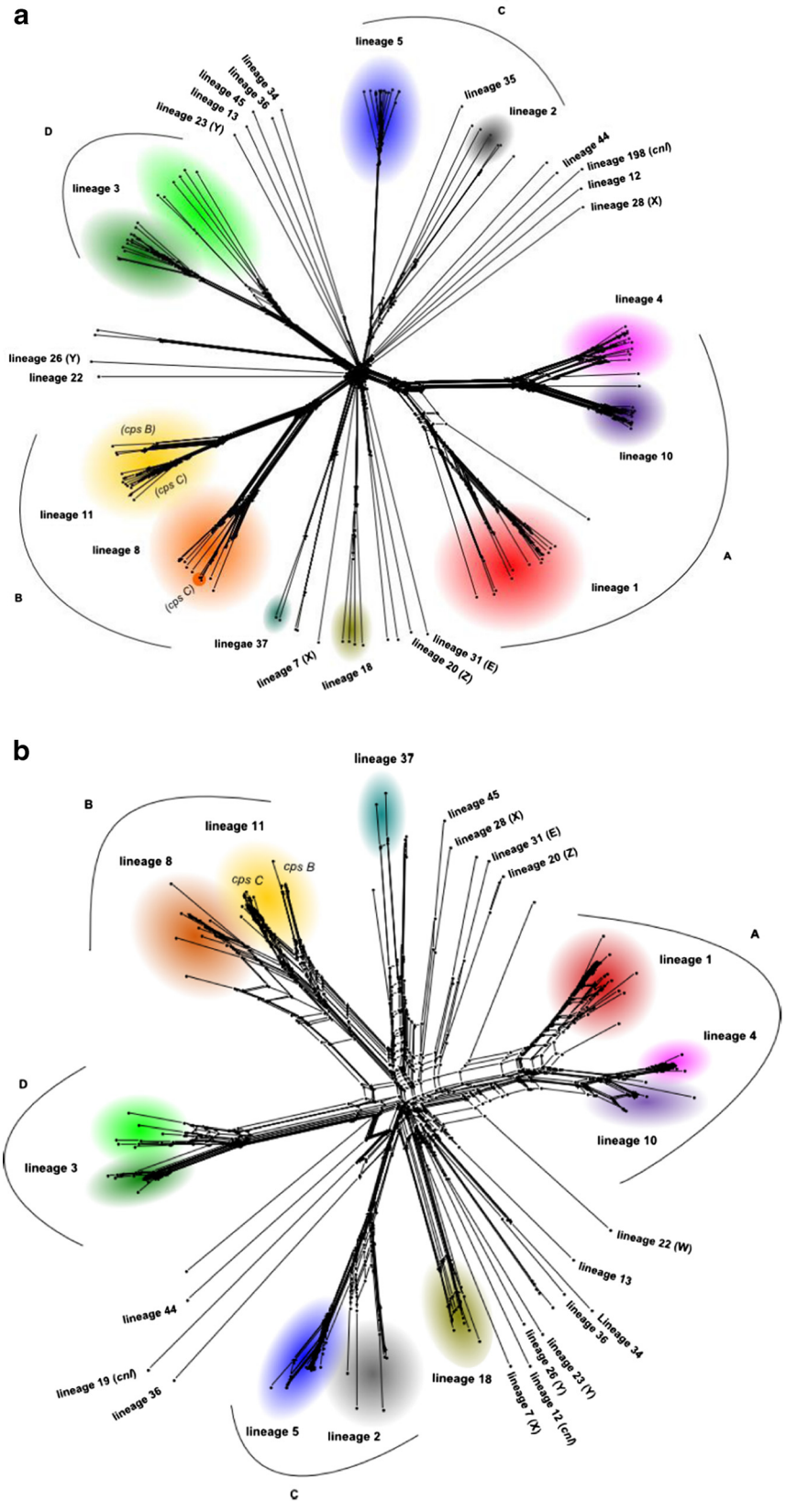


Figure 3 (See legend on next page.)

(See figure on previous page.)
Figure 3 128 representative *Neisseria meningitidis* genomes from the 20th and 21st Century. The relationships between meningococcal isolates are represented by two datasets in which (a.) 1605 core meningococcal loci (cgMLST) or (b.) 53 ribosomal protein genes (rMLST), a subset of the 1605 core loci, are used. In both trees major phylogenetic groups are noted A-D. For cross-compatible identification, and where there are 2 or more strains per lineage are present, the major MLST derived clonal complexes (cc) are identified by colour: red – ST-1 cc, purple – ST-5 cc, pink – ST-4 cc, teal – ST-37 cc, yellow – ST-11 cc, orange – ST-8 cc, green – ST-41/44 cc, blue – ST-32 cc, grey – ST-269 cc, olive – ST-18 cc. Capsular types other than A, B, or C are noted in parentheses, accept for Lineage 11 which are labelled (cps B and cps C). Unlabelled nodes are undefined lineages and currently do not have a clonal complex association; a full list of lineage and associated clonal complex nomenclature can be found in Table 4.

sequence data, with substantial improvement as read length increased. Read lengths of 100 bp, which are now routinely available, would reduce the missing data substantially [44,50]. Data quality was also determined by the details of the chemistry and procedures used [51,52], showing that NGS data are optimally useful when this information is deposited with them. Some coverage effects were seen, with sequences near the origin of replication consistently sequenced to a higher depth [53], than others but the genome of each assembly was adequately covered.

The BIGSdb platform accommodates sequence data derived from a particular isolate ranging from a single gene through multiple genes and contigs up to and including complete genomes [17]. The Genome Comparator tool can either use the annotations from a reference genome, which were used to compare the reference genomes with the assembled genomes, or sets of loci defined in the PubMLST sequence definition database, for which it maintains a complete catalogue of diversity described to date [15]. To enable consistent referencing each complete defined locus in the PubMLST *Neisseria* database, which can be any identifiable sequence string, was identified with a unique and arbitrary 'NEIS' number, which can be associated with other designations such as conventional Demeric gene names [54]. Additional loci that represent gene fragments used in typing schemes and peptide loci representing typing antigen variable regions [31,55], are also indexed within the database. The BIGSdb 'autotagger' function identified and automatically annotated an average of 1899 CDS from each assembled genome, with only a small number of paralogous loci (no more than 20) in the pangenome. The currently identified paralogous loci require additional manual annotation, they have been found to vary between the *Neisseria* species and may vary among meningococcal lineages. In conclusion, the approach can be used to analyse large numbers of WGS datasets consistently and is generally applicable for use across the bacterial domain.

Ultimately the PubMLST *Neisseria* database can be expanded, through a process of iterative gene discovery, to become a catalogue of the meningococcal 'pan genome' i.e. all of the genes present in the species or genus [56,57]. This database will develop over time by a process of community annotation but, by definition, the members of the meningococcal 'core genome', i.e. genes present in all meningococci, will already be present. Because every bacterial isolate is potentially an unrepresentative mutant and due to the imperfect nature of NGS assemblies, the core genome cannot be simply defined as the genes present in all isolates; however, the estimate of a core genome comprising 1605 genes generated here is in good agreement with other estimates (1532-1706) which were based on substantially fewer genomes [23,37,58,59]. A total 37% of the meningococcal core genes were assigned an EC number at the time of writing, indicating the magnitude of the annotation task which NGS data generates. While the membership of the core genome will be refined over time, it is unlikely to be very different from that proposed here. An updated list of meningococcal core genes will be maintained in the database.

The genealogies reconstructed with the NeighborNet algorithm using Genome Comparator data for the cgMLST and rMLST were consistent with those previously generated with MLST and a variety of other approaches [40,60,61]. The ribosomal genes (rMLST) and core genome (cgMLST) data provide more resolution, demonstrating that the six major hyperinvasive lineages included in this dataset cluster in to a number of larger groups [62]. Some lineages are more closely related to each other although the star phylogeny demonstrates a highly diverse and recombining population from which invasive lineages have emerged independently on several occasions [63]. As suggested from multilocus enzyme electrophoresis (MLEE) and other data [64], the serogroup A-associated lineages 1, 4 and 10 (ST-1, ST-4 and ST-5 complexes respectively) likely share a common ancestor [65], as do: lineage 8 (ST-8 complex) and lineage 11 (ST-11 complex); and lineage 5 (ST-32 complex) and lineage 2 (ST-269 complex). Lineage 3 (ST-41/44 complex) is a diverse lineage comprising both more and less invasive types. These data confirm that the invasive lineages are defined by sequence variation in the core genome, although certain members of the accessory genome, for example the capsule [66], the meningococcal disease associated island phage [67,68], and restriction modification systems [37,69] are differentially distributed among lineages.

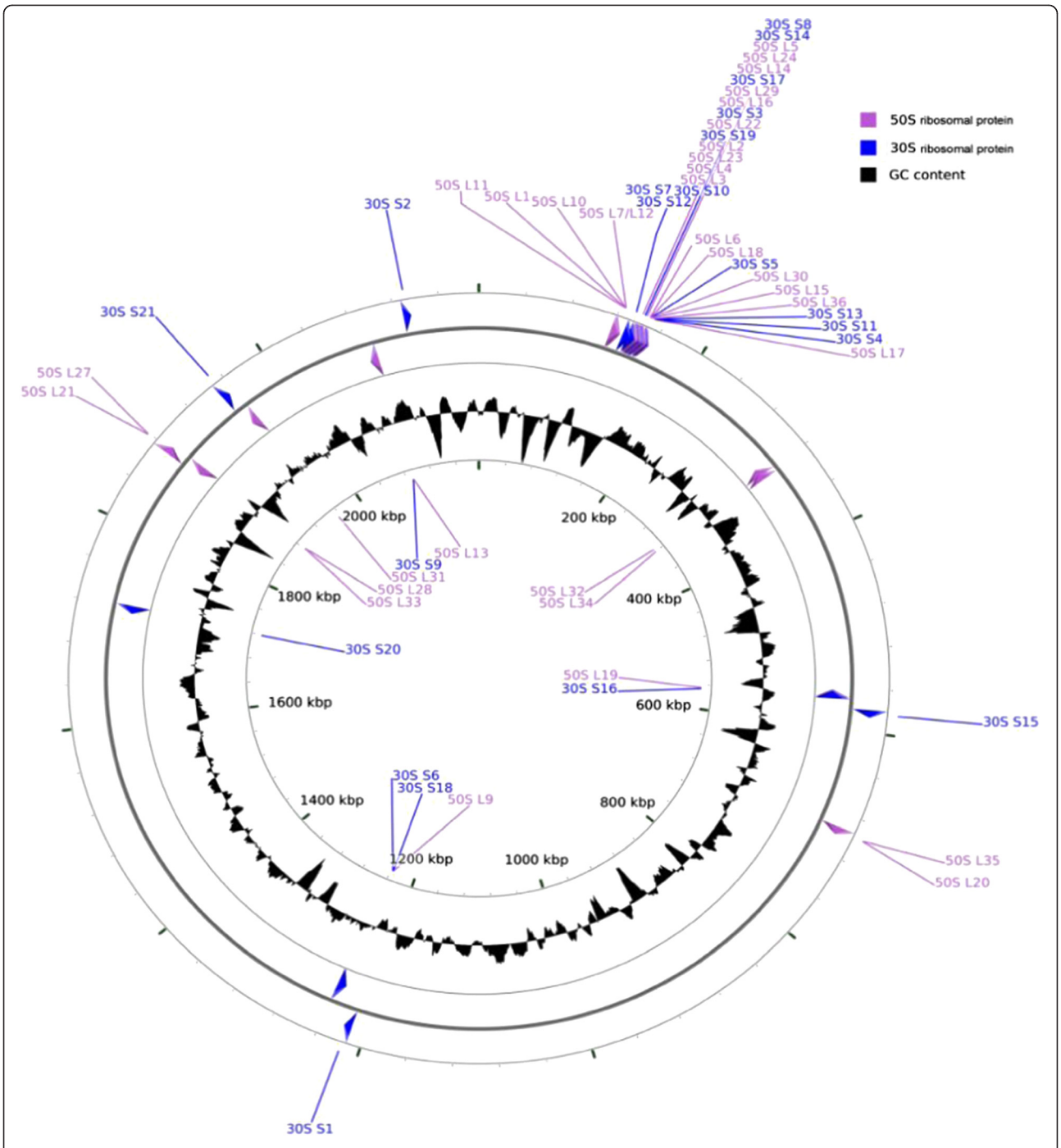


Figure 4 Location of rMLST genes within the meningococcal genome.

To reflect the increased resolution of whole genome typing we propose the use of a lineage nomenclature (Table 4) to distinguish groupings obtained by rMLST and cgMLST from the clonal complex association identified by MLST. This nomenclature also allows for the designations of sub-lineages which our data set, and others not described here, define additional prevalent biological and phenotypic associations such as the ET-15 mutants of lineage 11. The proposal was presented to a satellite sub-group meeting of the XIX International Pathogenic *Neisseria* Conference in October 2014, which included submitters, curators and users and the proposal is under consideration for adoption by the PubMLST Management Committee.

Table 4 Proposed Whole Genome Lineage Nomenclature

WGS nomenclature	MLST nomenclature
Lineage 11 ^	ST-11 cc
Lineage 3 ^	ST-41/44 cc
Lineage 23 ^	ST-23 cc
Lineage 1	ST-1 cc
Lineage 2	ST-269 cc
Lineage 4	ST-4 cc
Lineage 5	ST-32 cc
Lineage 6	ST-60 cc
Lineage 7	ST-750 cc
Lineage 8	ST-8 cc
Lineage 9	ST-92 cc
Lineage 10	ST-5 cc
Lineage 12	ST-53 cc
Lineage 13	ST-213 cc
Lineage 14	ST-174 cc
Lineage 15	ST-1157 cc
Lineage 16	ST-116 cc
Lineage 17	ST-175 cc
Lineage 18	ST-18 cc
Lineage 19	ST-198 cc
Lineage 20	ST-103 cc
Lineage 21	ST-212 cc
Lineage 22	ST-22 cc
Lineage 24	ST-106 cc
Lineage 25	ST-162 cc
Lineage 26	ST-167 cc
Lineage 27	ST-178 cc
Lineage 28	ST-181 cc
Lineage 29	ST-226 cc
Lineage 30	ST-231 cc
Lineage 31	ST-254 cc
Lineage 32	ST-282 cc
Lineage 33	ST-292 cc
Lineage 34	ST-334 cc
Lineage 35	ST-35 cc
Lineage 36	ST-364 cc
Lineage 37	ST-37 cc
Lineage 38	ST-376 cc
Lineage 39	ST-461 cc
Lineage 40	ST-549 cc
Lineage 41	ST-865 cc
Lineage 42	ST-1117 cc
Lineage 43	ST-1136 cc
Lineage 44	ST-4821 cc
Lineage 45	ST-4240/6688 cc

^ **Distinct** sub-lineages present; proposal to use decimal based (i.e. 11.1, 11.2, etc.) system for defined sub-lineages.
To simplify and differentiate between MLST typing and whole genome based typing we propose a lineage nomenclature that is associated with defined clonal complex (cc). The data includes all PubMLST *Neisseria* database isolates.

Conclusions

WGS data has the potential to unify studies of bacteria by providing comprehensive descriptions of genomic variation. To achieve this it is necessary to: (i) make the data available in a comprehensible way, along with information describing its completeness and accuracy; and (ii) link them to provenance and phenotype information, which describes the source of the sample and its properties, as well as the known properties of the genes identified and the deduced product. These datasets will grow in completeness and accuracy over time; however, it is also necessary for these data to be presented in a stable context, enabling even incomplete information to be explored. The approach described and validated here for the meningococcus is one way of achieving this, which employs generic, freely accessible and widely used tools. The use of the web interface within the PubMLST *Neisseria* database enables a process of community annotation whereby different members of the community can participate in the maintenance and improvement of sequence annotation and interpretation.

Methods

Bacterial strains and genomic DNA extraction

Genomic DNA from 108 diverse *Neisseria meningitidis* isolates was prepared from archive stocks which have been extensively characterized and previously reported [16,27,30,70]; this data set includes the 107 MLST global reference collection isolates and FAM18 [16,47]. Cultures were incubated on Columbia horse-blood ager (Oxoid) at 37°C in an atmosphere of 5% CO2 for 24 hours, subcultured and genomic DNA extracted using the Wizard® Genomic DNA Purification Kit (Promega).

Illumina sequencing

Standard Illumina multiplex libraries, grouped A-K, were generated. Adapter ligated DNA was amplified by PCR using Taq or Phusion® DNA polymerase and primers from the Illumina multiplexing sample preparation oligonucleotide kit, creating up to 12 libraries per group. Before and after each of these steps DNA was simultaneously cleaned up and size selected using a 1:1 (sample:beads) ratio of Ampure beads (Beckman Coulter Genomics). Libraries

were pooled in equimolar ratio and a maximum of twelve tagged, paired-end library aliquots were run per flowcell lane; every eighth lane contained the control genome, phiX 174. A standard Illumina clustering protocol was used with an additional QC step after cluster amplification. Passing flowcells were sequenced using the Illumina Genome Analyser II platform. Sequence reads have been deposited in the European Nucleotide Archive [EMBL: ERS006904 to ERS007010].

Genome assembly

Short-read sequences were assembled using the VelvetOptimiser *de novo* short-read assembly program optimisation script using the default parameters [25,26]. Once generated, there was no further manipulation of the assembled draft genome sequences.

Method analysis

For each step of the process where variation or patterning, not associated with or inherent in the genome biology could be introduced, non-biological run nodes were recorded. These included notations of: date; technician; reagent lot used; manual and robotic library preparation methods including plate lane; and sequencing steps specifically noting chemistry changes, flowcell lane, number of samples per multiplex group, and the machine used.

BIGSdb genome annotation and locus tagging

The sequence definition database was seeded using the core loci identified in finished *Neisseria meningitidis* genome annotations. The locus tag identifiers, 'NEIS' followed by an integer, was adopted in order to allow automated accessioning of loci as they are identified and added to the database. The NEIS, (short for '*Neisseria* genus') loci list was determined using the genome annotations of FAM18, H44/76, G2136, Z2491 and MC58 and represent, notionally, the pan-genome of the meningococcus. This included the ribosomal protein loci, a sub set of the core loci which are also orthologous across all bacterial species [40]. The NEIS identifiers are linked to an alias table that contains additional locus nomenclature associated with each locus which is searchable and therefore cross compatible with various annotations; such as specific finished genome locus tags, KEGG EC or common name. The number of loci contained in the list of the NEIS locus identifiers is not static and will change as loci are curated and added to the database over time.

The draft genome sequences were queried within BIGSdb using BLAST against the sequence definition database to identify defined allelic variation. Alleles were automatically annotated and assigned with the appropriate allele number for those loci for which definitions exist, in a process referred to as 'tagging' while new alleles were manually curated and assigned a new allele accession number. For the gene sequences with frame shift mutations, internal stop codons, etc., the sequence was assigned an allele designation and flagged as having an internal stop codon. Any gene sequences with missing data, i.e. those at the ends of contigs, were flagged as incomplete and not assigned an allele number. Once identified the locus allelic variant was linked to the isolate metadata.

Reference to *de novo* genome comparisons

Assembled draft genome sequences were compared to their reference genome using the BIGSdb Genome Comparator tool and assessed using the finished genome CDS sequence annotation. Genes from each genome were also compared to previously typed loci, including conventional and extended MLST loci, three antigen loci, *PorA VR1* and *VR2*, *FetA VR*, and *fHbp*, a surface antigen being explored as a vaccine candidate.

Sanger sequencing

Sanger sequencing was performed for resolution of typing loci conflicts found between Illumina and Sanger derived sequences using a reserved sample of the DNA used for Illumina sequencing. Reserved Illumina DNA was amplified and sequenced using previously published methods and primers for conventional MLST, eMLST, *PorA VR1*, *PorA VR2*, or *fHbp* loci [16,27,29,31]. Sanger trace files were assembled using the Staden sequence assembly package [71], and compared to the Illumina derived sequence using the MEGA5 alignment tools [72].

Bowtie and tablet

Read depth and sequence conflicts were checked by remapping using the Bowtie short-read aligner [73]. For target sequence assessment the contig containing the typing loci was extracted from the sequence bin and used as the reference segment and the FAM18, Z2491, H44/74 and G2136 finished genomes were used for read mapping their resequenced genomes respectively. Briefly, the short-reads were converted to SAM files and mapped against the reference segment using a randomized alignment order to avoid mapping bias. Aligned .SAM files were visualized using the Tablet software package [74]. Read depth and conflicting nucleotides of interest were identified and investigated.

Data access

Assembled contigs and annotation information can be accessed at PubMLST *Neisseria* database [http://pubmlst.org/neisseria/] using the query search, project '107 global collection'. Sequence reads have also been deposited in the European Nucleotide Archive (ENA) EMBL: ERS006904 to ERS007010 inclusive.

Additional files

References to the data sets supporting the results of this article are included within the article and its 3 additional .pdf files.

Additional file 1: Table S1. Velvet *de novo* assembly output statistics.

Additional file 2: Table S2. Loci with sequence discrepancies between the reference and resequenced genome.

Additional file 3: Table S3. Sequence discrepancy categories.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HBB designed the study, carried out the analysis and wrote the manuscript. CC assembled the draft genomes. KAJ designed and develops the BIGSdb platform. JP collaborated in the sequencing of the genomes. MCJM conceived the approach of high-throughput, whole genome, gene-by-gene analysis using *de novo* draft genomes and co-wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Theresa Feltwell, John Burton, Michael Quail, and Stephen D Bently of The Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK for sequencing coordination, assistance and technical consultations. MCJM is a Wellcome Trust Senior Fellow in Basic Biomedical Sciences. This publication made use of the Neisseria Multi Locus Sequence Typing website (http://pubmlst.org/neisseria/) developed by Keith Jolley and sited at the University of Oxford [17]. The development of this site has been funded by the Wellcome Trust and European Union.

Author details

[1]Department of Zoology, University of Oxford, Oxford, UK. [2]The Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK.

Received: 2 October 2014 Accepted: 4 December 2014
Published: 18 December 2014

References

1. Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R: **Microbiology in the post-genomic era.** *Nat Rev Microbiol* 2008, **6**(6):419–430.
2. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing.** *Genome Res* 2007, **17**(11):1697–1706.
3. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB: **ALLPATHS: De novo assembly of whole-genome shotgun microreads.** *Genome Res* 2008, **18**(5):810–820.
4. Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J: **De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer.** *Genome Res* 2008, **18**(5):802–809.
5. Farrer RA, Kemen E, Jones JD, Studholme DJ: ***De novo* assembly of the *Pseudomonas syringae pv. syringae* B728a genome using Illumina/Solexa short sequence reads.** *FEMS Microbiol Lett* 2009, **291**(1):103–111.
6. Nishito Y, Osana Y, Hachiya T, Popendorf K, Toyoda A, Fujiyama A, Itaya M, Sakakibara Y: **Whole genome assembly of a natto production strain *Bacillus subtilis* natto from very short read data.** *BMC Genomics* 2010, **11**:243.
7. Nagarajan H, Butler JE, Klimes A, Qiu Y, Zengler K, Ward J, Young ND, Methe BA, Palsson BO, Lovley DR, Barrett CL: **De Novo Assembly of the Complete Genome of an Enhanced Electricity-Producing Variant of Geobacter sulfurreducens Using Only Short Reads.** *PLoS One* 2010, **5**(6):e10922.
8. Silva A, Schneider MPC, Cerdeira L, Barbosa MS, Ramos RTJ, Carneiro AR, Santos R, Lima M, D'Afonseca V, Almeida SS, Santos AR, Soares SC, Pinto AC, Ali A, Dorella FA, Rocha F, de Abreu VAC, Trost E, Tauch A, Shpigel N, Miyoshi A, Azevedo V: **Complete Genome Sequence of *Corynebacterium pseudotuberculosis* I19, a Strain Isolated from a Cow in Israel with Bovine Mastitis.** *J Bacteriol* 2011, **193**(1):323–324.
9. Cerdeira LT, Carneiro AR, Ramos RTJ, de Almeida SS, D'Afonseca V, Schneider MPC, Baumbach J, Tauch A, McCulloch JA, Azevedo VAC, Silva A: **Rapid hybrid de novo assembly of a microbial genome using only short reads: Corynebacterium pseudotuberculosis I19 as a case study.** *J Microbiol Meth* 2011, **86**(2):218–223.
10. Flicek P, Birney E: **Sense from sequence reads: methods for alignment and assembly (vol 6, pg S6, 2009).** *Nat Methods* 2010, **7**(6):479–479.
11. Ronen R, Boucher C, Chitsaz H, Pevzner P: **SEQuel: improving the accuracy of genome assemblies.** *Bioinformatics* 2012, **28**(12):i188–196.
12. Chain PSG, Grafham DV, Fulton RS, FitzGerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, Cole JR, Ding Y, Dugan S, Field D, Garrity GM, Gibbs R, Graves T, Han CS, Harrison SH, Highlander S, Hugenholtz P, Khouri HM, Kodira CD, Kolker E, Kyrpides NC, Lang D, Lapidus A, Malfatti SA, Markowitz V, Metha T, *et al*: **Genome Project Standards in a New Era of Sequencing.** *Science* 2009, **326**(5950):236–237.
13. Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR, Chisholm SW: **Whole Genome Amplification and De novo Assembly of Single Bacterial Cells.** *PLoS One* 2009, **4**(9):e6864.
14. Earl AM, Eppinger M, Fricke WF, Rosovitz MJ, Rasko DA, Daugherty S, Losick R, Kolter R, Ravel J: **Whole-Genome Sequences of *Bacillus subtilis* and Close Relatives.** *J Bacteriol* 2012, **194**(9):2378–2379.
15. Maiden MC, van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND: **MLST revisited: the gene-by-gene approach to bacterial genomics.** *Nat Rev Microbiol* 2013, **11**(10):728–736.
16. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.** *Proc Natl Acad Sci USA* 1998, **95**(6):3140–3145.
17. Jolley KA, Maiden MC: **BIGSdb: Scalable analysis of bacterial genome variation at the population level.** *BMC Bioinformatics* 2010, **11**(1):595.
18. Caugant DA: **Population genetics and molecular epidemiology of *Neisseria meningitidis*.** *APMIS* 1998, **106**(5):505–525.
19. Yazdankhah SP, Caugant DA: ***Neisseria meningitidis*: an overview of the carriage state.** *J Med Microbiol* 2004, **53**(Pt 9):821–832.
20. Neal KR: **Changing carriage rate of *Neisseria meningitidis* among university students during the first week of term: cross sectional study.** *BMJ* 2000, **320**(7238):846–849.
21. Caugant DA, Maiden MC: **Meningococcal carriage and disease - population biology and evolution.** *Vaccine* 2009, **27**(Suppl 2):B64–70.
22. Marri PR, Paniscus M, Weyand NJ, Rendon MA, Calton CM, Hernandez DR, Higashi DL, Sodergren E, Weinstock GM, Rounsley SD, So M: **Genome sequencing reveals widespread virulence gene exchange among human *Neisseria* species.** *PLoS One* 2010, **5**(7):e11835.
23. Schoen C, Blom J, Claus H, Schramm-Gluck A, Brandt P, Muller T, Goesmann A, Joseph B, Konietzny S, Kurzai O, Schmitt C, Friedrich T, Linke B, Vogel U, Frosch M: **Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in *Neisseria meningitidis*.** *Proc Natl Acad Sci USA* 2008, **105**(9):3473–3478.
24. Joseph B, Schneiker-Bekel S, Schramm-Gluck A, Blom J, Claus H, Linke B, Schwarz RF, Becker A, Goesmann A, Frosch M, Schoen C: **Comparative genome biology of a serogroup B carriage and disease strain supports a polygenic nature of meningococcal virulence.** *J Bacteriol* 2010, **192**(20):5363–5377.
25. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**(5):821–829.
26. Zerbino D: **Using the Velvet *de novo* Assembler for Short-Read Sequencing Technologies.** *Curr Protoc Bioinformatics* 2010, **11**(5):1–12.
27. Didelot X, Urwin R, Maiden MC, Falush D: **Genealogical typing of *Neisseria meningitidis*.** *Microbiology* 2009, **155**(10):3176–3186.
28. Holmes EC, Urwin R, Maiden MCJ: **The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*.** *Mol Biol Evol* 1999, **16**(6):741–749.
29. Russell JE, Jolley KA, Feavers IM, Maiden MC, Suker J: **PorA variable regions of *Neisseria meningitidis*.** *Emerg Infect Dis* 2004, **10**(4):674–678.
30. Thompson EAL, Feavers IM, Maiden MCJ: **Antigenic diversity of meningococcal enterobactin receptor FetA, a vaccine component.** *Microbiology* 2003, **149**(Pt 7):1849–1858.
31. Brehony C, Wilson DJ, Maiden MC: **Variation of the factor H-binding protein of *Neisseria meningitidis*.** *Microbiology* 2009, **155**:4155–4169.
32. Benjamini Y, Speed TP: **Summarizing and correcting the GC content bias in high-throughput sequencing.** *Nucleic Acids Res* 2012, **40**(10):e72.

33. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.** *Nucleic Acids Res* 2008, **36**(16):e105.
34. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ: **Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G + C)-biased genomes.** *Nat Methods* 2009, **6**(4):291–295.
35. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A: **Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.** *Genome Biol* 2011, **12**(2):R18.
36. Jolley KA, Hill DM, Bratcher HB, Harrison OB, Feavers IM, Parkhill J, Maiden MC: **Resolution of a meningococcal disease outbreak from whole genome sequence data with rapid web-based analysis methods.** *J Clin Microbiol* 2012, **50**(9):3046–3053.
37. Budroni S, Siena E, Hotopp JCD, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR, Daugherty SC, Angiuoli SV, Covacci A, Pizza M, Rappuoli R, Moxon ER, Tettelin H, Medini D: ***Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination.** *Proc Natl Acad Sci USA* 2011, **108**(11):4494–4499.
38. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 1999, **27**(1):29–34.
39. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 2000, **28**(1):27–30.
40. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony CM, Colles FM, Wimalarathna HM, Harrison OB, Sheppard SK, Cody AJ, Maiden MC: **Ribosomal Multi-Locus Sequence Typing: universal characterization of bacteria from domain to strain.** *Microbiology* 2012, **158**:1005–1015.
41. Jolley KA, Maiden MC: **Using MLST to study bacterial variation: prospects in the genomic era.** *Future Microbiol* 2014, **9**:623–630.
42. Loman NJ, Constantinidou C, Chan JZM, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ: **High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity.** *Nat Rev Microbiol* 2012, **10**(9):599–606.
43. Aury JM, Cruaud C, Barbe V, Rogier O, Mangenot S, Samson G, Poulain J, Anthouard V, Scarpelli C, Artiguenave F, Wincker P: **High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies.** *BMC Genomics* 2008, **9**:603.
44. Reuter S, Ellington MJ, Cartwright EJ, Koser CU, Torok ME, Gouliouris T, Harris SR, Brown NM, Holden MT, Quail M, Parkhill J, Smith GP, Bentley SD, Peacock SJ: **Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology.** *JAMA Intern Med* 2013, **173**(15):1397–1404.
45. Bratcher HB, Bennett JS, Maiden MCJ: **Evolutionary and genomic insights into meningococcal biology.** *Future Microbiol* 2012, **7**(7):873–885.
46. Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR, Morelli G, Basham D, Brown D, Chillingworth T, Davies RM, Davis P, Devlin K, Feltwell T, Hamlin N, Holroyd S, Jagels K, Leather S, Moule S, Mungall K, Quail MA, Rajandream MA, Rutherford KM, Simmonds M, Skelton J, Whitehead S, Spratt BG, Barrell BG: **Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491.** *Nature* 2000, **404**(6777):502–506.
47. Bentley SD, Vernikos GS, Snyder LA, Churcher C, Arrowsmith C, Chillingworth T, Cronin A, Davis PH, Holroyd NE, Jagels K, Maddison M, Moule S, Rabbinowitsch E, Sharp S, Unwin L, Whitehead S, Quail MA, Achtman M, Barrell B, Saunders NJ, Parkhill J: **Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18.** *PLoS Genet* 2007, **3**(2):e23.
48. Tettelin H, Saunders NJ, Heidelberg J, Jeffries AC, Nelson KE, Eisen JA, Ketchum KA, Hood DW, Peden JF, Dodson RJ, Nelson WC, Gwinn ML, DeBoy R, Peterson JD, Hickey EK, Haft DH, Salzberg SL, White O, Fleischmann RD, Dougherty BA, Mason T, Ciecko A, Parksey DS, Blair E, Cittone H, Clark EB, Cotton MD, Utterback TR, Khouri H, Qin H, *et al*: **Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58.** *Science* 2000, **287**(5459):1809–1815.
49. Wetzel J, Kingsford C, Pop M: **Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies.** *BMC Bioinformatics* 2011, **12**:95.
50. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions.** *Nature Reviews Genetics* 2012, **13**(2):36.
51. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ: **A large genome center's improvements to the Illumina sequencing system.** *Nat Methods* 2008, **5**(12):1005–1010.
52. Quail MA, Swerdlow H, Turner DJ: **Improved protocols for the illumina genome analyzer sequencing system.** *Curr Protoc Hum Genet* 2009, **18:** doi:10.1002/0471142905.hg1802s62.
53. Paszkiewicz K, Studholme DJ: **De novo assembly of short sequence reads.** *Brief Bioinform* 2010, **11**(5):457–472.
54. Demerec M, Adelberg EA, Clark AJ, Hartman PE: **A proposal for a uniform nomenclature in bacterial genetics.** *Genetics* 1966, **54**(1):61–76.
55. Bambini S, De Chiara M, Muzzi A, Mora M, Lucidarme J, Brehony C, Borrow R, Masignani V, Comanducci M, Maiden MC, Rappuoli R, Pizza M, Jolley KA: **Neisseria adhesin A variation and revised nomenclature scheme.** *Clin Vaccine Immunol* 2014, **21**:966–971.
56. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome.** *Curr Opin Genet Dev* 2005, **15**(6):589–594.
57. Bennett JS, Bentley SD, Vernikos GS, Quail MA, Cherevach I, White B, Parkhill J, Maiden MCJ: **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* 2010, **11**:652.
58. Joseph B, Schwarz RF, Linke B, Blom J, Becker A, Claus H, Goesmann A, Frosch M, Muller T, Vogel U, Schoen C: **Virulence evolution of the human pathogen *Neisseria meningitidis* by recombination in the core and accessory genome.** *PLoS One* 2011, **6**(4):e18441.
59. Hotopp JC, Grifantini R, Kumar N, Tzeng YL, Fouts D, Frigimelica E, Draghi M, Giuliani MM, Rappuoli R, Stephens DS, Grandi G, Tettelin H: **Comparative genomics of *Neisseria meningitidis*: core genome, islands of horizontal transfer and pathogen-specific genes.** *Microbiology* 2006, **152**(Pt 12):3733–3749.
60. Brehony C, Jolley KA, Maiden MC: **Multilocus sequence typing for global surveillance of meningococcal disease.** *FEMS Microbiol Rev* 2007, **31**(1):15–26.
61. Brehony C, Trotter CL, Ramsay ME, Chandra M, Jolley KA, van der Ende A, Carion F, Berthelsen L, Hoffmann S, Harðardóttir H, Vazquez J, Murphy K, Toropainen M, Caniça M, Ferreira E, Diggle M, Edwards G, Taha M-K, Stefanelli P, Kriz P, Gray S, Fox A, Jacobsson S, Claus H, Vogel U, Tzanakaki G, Heuberger S, Caugant DA, Frosch M, Maiden MCJ: **Differential age distribution of disease-associated meningococcal lineages-Implications for vaccine development.** *Clin Vaccine Immunol* 2014, **21**(6):847–853.
62. Watkins ER, Maiden MC: **Persistence of hyperinvasive meningococcal strain types during global spread as recorded in the PubMLST database.** *PLoS ONE* 2012, **7**(9):e45349.
63. Didelot X, Falush D: **Inference of bacterial microevolution using multilocus sequence data.** *Genetics* 2007, **175**(3):1251–1266.
64. Caugant DA, Mocca LF, Frasch CE, Frøholm LO, Zollinger WD, Selander RK: **Genetic structure of *Neisseria meningitidis* populations in relation to serogroup, serotype, and outer membrane protein pattern.** *J Bacteriol* 1987, **169**(6):2781–2792.
65. Olyhoek T, Crowe BA, Achtman M: **Clonal population structure of *Neisseria meningitidis* serogroup A isolated from epidemics and pandemics between 1915 and 1983.** *Rev Infect Dis* 1987, **9**:665–682.
66. Harrison OB, Claus H, Jiang Y, Bennett JS, Bratcher HB, Jolley KA, Corton C, Care R, Poolman JT, Zollinger WD, Frasch CE, Stephens DS, Feavers I, Frosch M, Parkhill J, Vogel U, Quail MA, Bentley SD, Maiden MCJ: **Description and nomenclature of *Neisseria meningitidis* capsule locus.** *Emerg Infect Dis* 2013, **19**(4):566–573.
67. Bille E, Ure R, Gray SJ, Kaczmarski EB, McCarthy ND, Nassif X, Maiden MC, Tinsley CR: **Association of a bacteriophage with meningococcal disease in young adults.** *PLoS ONE* 2008, **3**(12):e3885.
68. Bille E, Zahar JR, Perrin A, Morelle S, Kriz P, Jolley KA, Maiden MC, Dervin C, Nassif X, Tinsley CR: **A chromosomally integrated bacteriophage in invasive meningococci.** *J Exp Med* 2005, **201**(12):1905–1913.
69. Claus H, Friedrich A, Frosch M, Vogel U: **Differential distribution of novel restriction-modification systems in clonal lineages of *Neisseria meningitidis*.** *J Bacteriol* 2000, **182**(5):1296–1303.
70. Urwin R, Russell JE, Thompson EA, Holmes EC, Feavers IM, Maiden MC: **Distribution of Surface Protein Variants among Hyperinvasive Meningococci: Implications for Vaccine Design.** *Infect Immun* 2004, **72**(10):5955–5962.
71. Staden R: **The Staden sequence analysis package.** *Mol Biotechnol* 1996, **5**:233–241.

72. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28**(10):2731–2739.
73. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**(3):R25.
74. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D: **Tablet–next generation sequence assembly visualization.** *Bioinformatics* 2010, **26**(3):401–402.

doi:10.1186/1471-2164-15-1138
Cite this article as: Bratcher *et al.*: A gene-by-gene population genomics platform: *de novo* assembly, annotation and genealogical analysis of 108 representative *Neisseria meningitidis* genomes. *BMC Genomics* 2014 **15**:1138.

Additional Table 1

Velvet *de novo* assembly output statistics of 120 genomes sorted by multiplex group and assessed for 1605 core loci. 'Good' assemblies had a long k-mer value with (i) a large maximum contig length, (ii) a large N50 contig size, (iii) the smallest number of contigs and (iv) a total assembly length that closely matched that of known reference genomes.

multiplex group	Isolate	pubMLST isolate id	kmer size	number of contigs	N50 value	longest contig	total length	# core genes found	number of incomplete
A	F4698	120	33	395	12,600	49,668	2,060,501	1605	20
	F6124	128	31	455	11,285	45,224	2,056,609	1603	90
	H44/76	237	37	338	20,701	57,598	2,099,963	1605	0
	500	343	39	304	21,394	97,812	2,070,954	1605	30
	BZ 10	398	33	346	18,670	50,095	2,052,180	1605	51
	BZ 169	408	37	329	22,242	94,444	2,142,210	1605	33
	BZ 198	409	35	329	23,777	77,381	2,092,617	1602	36
	NG PB24	434	39	321	25,396	78,292	2,114,016	1600	40
	NG P20	436	39	305	21,082	97,812	2,066,372	1605	33
	8680	441	31	375	16,653	48,493	2,088,107	1605	50
	3906	443	35	415	15,772	54,108	2,070,442	1592	58
	1000	446	31	578	11,824	41,239	2,142,832	1600	98
B	153	238	33	360	14,696	56,581	2,060,468	1600	5
	297-0	442	31	551	9,690	39,840	1,975,180	1595	10
	14/1455	451	33	396	12,729	56,942	2,062,931	1605	0
	G2136	638	29	350	16,159	88,069	2,032,480	1603	2
	B6116/77	639	35	342	17,161	87,501	2,062,986	1605	0
	312 901	642	39	289	20,949	69,056	2,053,253	1604	1
	AK50	647	35	376	19,451	58,724	2,142,828	1604	1
	M-101/93	648	35	433	16,908	54,910	2,140,588	1603	2
	M40/94	650	37	309	26,096	87,419	2,134,770	1605	0
	88/03415	654	27	701	7,921	30,184	2,072,899	1605	0
	860060	657	39	375	18,036	68,810	2,109,112	1602	3
	890326	658	33	407	15,225	58,343	2,034,704	1599	6
C	2059001	387	43	284	25,837	99,092	2,078,511	1605	30
	SB25	640	31	456	10,361	30,130	2,044,661	1601	111
	94/155	641	33	310	18,055	91,556	2,047,907	1604	56
	204/92	645	35	418	16,437	72,963	2,128,705	1605	57
	400	646	41	327	25,497	105,386	2,120,654	1604	32
	50/94	649	39	348	22,754	77,352	2,108,496	1605	41
	N45/96	652	39	353	21,906	67,724	2,120,144	1605	36
	E32	655	37	346	21,399	86,854	2,123,723	1603	42
	E26	656	39	264	21,788	82,564	2,038,548	1590	22
	A22	659	43	326	23,475	64,764	2,141,924	1603	26
	71/94	660	39	311	18,826	87,551	2,045,624	1599	29
	860800	661	39	323	18,170	55,079	2,047,351	1597	28
D	7891	7	41	291	23,459	74,420	2,068,305	1604	21
	20	34	41	318	19,311	54,815	2,082,218	1605	28
	26	35	39	294	24,570	74,067	2,070,452	1605	32

	S5611	67	41	337	22,332	103,457	2,104,463	1605	39
	154	239	41	274	22,576	63,682	2,062,368	1605	27
	D8	316	37	291	23,361	112,831	2,070,336	1601	35
	NG 3/88	417	41	361	21,693	63,914	2,131,612	1603	39
	Z2491	613	41	293	23,311	99,084	2,079,826	1605	38
	AK22	643	41	341	20,443	91,975	2,077,842	1603	38
	L93/4286	644	41	326	22,290	82,476	2,080,870	1605	37
	931905	651	35	384	17,823	47,488	2,090,848	1604	34
	91/40	653	41	349	23,097	90,226	2,114,167	1603	34
E	243	52	39	317	17,283	77,632	2,069,890	1605	38
	CN100	90	35	316	19,064	73,069	2,130,298	1605	45
	BZ 147	403	37	474	12,070	42,975	2,102,833	1604	64
	DK 353	412	39	359	16,773	54,630	2,081,831	1601	51
	EG 328	413	37	443	18,002	58,857	2,195,602	1602	46
	EG 011	416	31	404	15,108	36,790	2,112,591	1600	53
	NG 6/88	419	37	396	14,421	69,438	2,063,651	1593	41
	NG G40	425	35	414	16,664	48,685	2,116,622	1599	32
	NG 080	430	35	397	14,987	61,702	2,109,324	1605	51
	528	445	35	525	11,108	62,642	2,113,911	1597	25
	371	466	41	347	17,599	69,495	2,088,208	1605	30
	79126	494	37	360	16,676	54,826	2,120,981	1605	42
F	6748	10	43	356	19,249	51,927	2,096,380	1605	19
	80049	299	35	339	15,944	53,762	2,135,885	1603	36
	D1	314	39	311	19,986	69,585	2,067,345	1605	41
	90/18311	391	41	310	20,580	54,431	2,067,022	1603	40
	BZ 163	407	39	306	18,208	69,617	2,112,365	1604	37
	NG F26	420	35	323	19,217	87,452	2,058,744	1604	40
	NG H41	422	39	434	14,287	69,665	2,087,496	1598	68
	NG E28	426	39	386	17,364	61,923	2,115,078	1596	49
	BRAZ10	468	41	343	18,376	54,425	2,067,672	1605	40
	MA-5756	507	41	340	17,733	49,614	2,067,675	1605	43
	92001	597	41	302	19,796	50,174	2,080,136	1605	28
G	A4/M1027	1	39	364	15,404	50,093	2,069,108	1603	66
	139M	13	35	436	11,994	62,369	2,135,812	1605	31
	H1964	210	35	314	15,831	59,026	2,070,317	1605	43
	BZ 232	410	39	383	19,322	62,152	2,131,285	1598	28
	DK 24	411	39	383	19,327	75,997	2,160,627	1600	36
	EG 327	414	37	406	13,373	36,942	2,079,945	1600	41
	NG H15	421	39	393	15,387	57,400	2,097,789	1605	58
	NG E30	427	37	323	16,946	51,511	2,053,239	1603	43
	NG H36	428	41	342	19,230	66,445	2,149,056	1604	49
	SWZ107	444	39	381	16,136	58,145	2,124,785	1593	41
	690	467	35	296	19,054	74,942	2,065,096	1604	40
	79128	492	37	406	17,138	82,248	2,149,843	1601	70
H	S3131	19	41	264	24,737	96,136	2,074,237	1605	35
	S4355	24	41	319	22,947	89,288	2,089,442	1605	36

	255	46	39	302	20,445	74,202	2,067,387	1605	43
	254	64	37	358	15,763	69,510	2,069,498	1604	56
	IAL2229	84	37	272	21,515	96,122	2,064,262	1605	32
	196/87	340	37	348	20,035	77,445	2,144,814	1605	66
	38VI	349	37	303	21,991	97,811	2,064,093	1605	43
	NG 4/88	418	37	322	21,855	59,704	2,058,894	1596	45
	NG H38	423	37	369	18,573	60,656	2,129,552	1601	55
	NG E31	424	39	418	18,401	83,563	2,146,433	1598	40
	106	488	37	309	22,754	77,691	2,085,979	1605	34
	322/85	493	33	366	17,670	43,368	2,154,620	1604	69
I	120M	2	33	463	10,581	33,563	2,074,516	1605	72
	129	11	35	347	14,410	79,861	2,083,935	1605	42
	10	31	39	354	14,054	53,317	2,070,182	1605	51
	393	61	35	397	14,162	47,836	2,081,413	1604	55
	11-004	82	39	348	13,854	47,478	2,068,319	1604	48
	1014	160	39	363	14,158	51,115	2,067,144	1605	50
	F1576	344	37	335	15,602	38,728	2,057,731	1605	47
	BZ 83	400	35	373	14,417	56,792	2,092,840	1603	50
	BZ 133	401	39	374	17,758	63,829	2,106,158	1605	42
	EG 329	415	35	439	12,338	45,158	2,136,007	1604	69
	NG144/82	431	37	451	10,926	74,300	2,103,741	1604	68
J	FAM18	4	41	466	16,865	69,539	2,067,451	1605	70
	139M	13	43	406	23,609	113,337	2,155,870	1605	31
	M597	369	41	344	21,063	111,183	2,068,928	1604	42
	DK 24	411	43	397	23,918	96,694	2,172,284	1600	36
	NG G40	425	37	437	23,891	103,915	2,120,214	1599	32
	528	445	41	444	23,521	100,197	2,139,911	1597	25
	E26	656	41	324	26,136	136,907	2,042,452	1590	22
K	6748	10	61	296	40,825	175,414	2,144,958	1605	19
	139M	13	57	294	26,154	149,689	2,141,819	1605	31
	F4698	120	63	243	29,907	81,537	2,132,393	1605	20
	BZ 169	408	63	269	36,513	99,534	2,211,536	1605	33
	297-0	442	63	289	29,365	102,428	2,065,716	1595	27
	14/1455	451	67	374	20,999	92,289	2,137,543	1605	24
	88/03415	654	63	297	35,880	158,137	2,172,921	1605	30
	mean assembly values for all multiplex groups		**n/a**	**361**	**19,103**	**72,140**	**2,095,268**	**1603**	**38**

Additional table 2 Loci with sequence discrepancies between the reference and re-sequenced genome

Comparison of finished reference to de novo assembled genome and the allelic variation found and grouped into one of five categories (A-F). Annotated loci of H44/76, Z2491, FAM18, and G2136 finished genomes were compared to their corresponding de novo resequenced genome using the BIGSdb Genome Comparator Tool. Differences were investigated using MEGA5 and Bowtie and reciprocal BLAST searches were done to identify the paralogous matches. Section (A) lists the loci in each de novo assembled genome that were unidentified and section (F) lists loci that had sequence differences but were not caused by sequencing errors, but instead paralogous loci cross-identification by BIGSdb scanning program. Section (B-D) lists the loci that have credible sequencing errors while section (E) lists loci that failed to assemble fully due to repeat sequence tracts. Loci with a reference to resequenced discrepancy are in **bold**; orthologs of each locus identified are also listed for each genome, with potential orthologs in parentheses.

Finished genome (potential) orthologous locus tag identifier					
G2136	**H44/76**	**Z2491**	**FAM18**	**BIGSdb**	Annotation [#]
A. Loci missing in the re-sequenced, *de novo* assembled genomes					
NMBG2136_1893	**NMBH4476_1938**	NMA0440 *	NMC1978 *	**NEIS1978**	(nlpC) p60 family protein
NMBG2136_2006	NMBH4476_2050	NMA0324	**NMC2084**	**NEIS2084**	adhesin (mafB2)
NMBG2136_0596	NMBH4476_0371	NMA2113	**NMC1790**	nca	adhesin (mafB2)
NMBG2136_1726	NMBH4476_0383	NMA2113	**NMC0597**	**NEIS0597**	adhesin (mafB2)
NMBG2136_1619	NMBH4476_1199	NMA2005	**NMC1668**	**NEIS1668**	cell-surface protein
NMBG2136_0443	NMBH4476_1677	(NMA0688)	NMC0450	nca	conserved hypothetical protein
NMBG2136_0446	NMBH4476_1682	NMA0690	NMC0450	nca	conserved hypothetical protein
NMBG2136_0028	NMBH4476_0024	NMA0272	NMC0003	nca	fimbrial protein precursor
-	-	**NMA0532**	-	nca	hypothetical protein
NMBG2136_0966	NMBH4476_1083	**NMA1192**	NMC0972	nca	hypothetical protein
NMBG2136_0966	NMBH4476_1083	**NMA1307**	NMC0972	nca	hypothetical protein
NMBG2136_0522	NMBH4476_0489	NMA0786	NMC1653	nca	hypothetical protein
(NMBG2136_0437)	**NMBH4476_1682**	NMA0690	NMC0452	nca	hypothetical protein
NMBG2136_0966	NMBH4476_1083	**NMA1860**	NMC0972	nca	hypothetical protein
NMBG2136_0523	NMBH4476_0805	NMA1626	**NMC0527**	nca	iron-regulated protein (frpC)
NMBG2136_0523	**NMBH4476_1605**	NMA1626	NMC0527	nca	Iron-regulated protein (frpA)
NMBG2136_1306	**NMBH4476_0805**	NMA1626	NMC0527	nca	Iron-regulated protein (frpC)
NMBG2136_1025	NMBH4476_1097	-	-	nca	membrane protein
-	-	-	**NMC0296**	nca	pseudogene
-	-	-	**NMC1345**	nca	pseudogene
-	-	**NMA0787**	-	nca	pseudogene
NMBG2136_2006	NMBH4476_2050	NMA0324	**NMC2090**	**NEIS2090**	pseudogene (mafb protein fragment)
NMBG2136_1527	**NMBH4476_0570**	NMA1914	NMC1573	nca	putative Caudovirus prohead protease (phage associated protein)
NMBG2136_0019	**NMBH4476_0598**	NMA1797	NMC1715	nca	tspB family protein
NMBG2136_0019	**NMBH4476_0681**	NMA1797	NMC0025	nca	tspB family protein
NMBG2136_0019	**NMBH4476_1698**	NMA1797	NMC1866	nca	tspB protein
NMBG2136_0407	NMBH4476_1758 *	NMA0650	**NMC0408**	**NEIS0408**	type IV pilus secretin (pilQ)
B. Non-synonymous amino acid change					
NMBG2136_1781	NMBH4476_0312	NMA2170	NMC1854	**NEIS1854**	7-cyano-7-deazaguanine reductase (queF)
NMBG2136_0481	NMBH4476_1645	**NMA0724**	NMC0484	**NEIS0484**	chromosome segregation protein
NMBG2136_1411	NMBH4476_0699	NMA1729	NMC1458	**NEIS1458**	conserved hypothetical protein
NMBG2136_1570	NMBH4476_0527	**NMA1952**	NMC1611	**NEIS1611**	conserved hypothetical protein
NMBG2136_0979	**NMBH4476_1172**	-	-	nca	hypothetical protein
NMBG2136_1921	**NMBH4476_1964**	NMA0418	NMC2001	**NEIS2001**	hypothetical protein
NMBG2136_0794	NMBH4476_1326	**NMA1073**	NMC0801	**NEIS0801**	hypothetical protein
NMBG2136_0798	NMBH4476_1319	**NMA1083**	NMC0806	nca	hypothetical protein
NMBG2136_0912	NMBH4476_1232	**NMA1135**	NMC0917	**NEIS0917**	hypothetical protein
-	**NMBH4476_0323**	-	-	nca	hypothetical protein
NMBG2136_0274	NMBH4476_0268	NMA2214 *	NMC0832	nca	hypothetical protein
NMBG2136_0303	NMBH4476_1858 *	**NMA0537**	NMC0306	**NEIS0306**	hypothetical protein / integral membrane protein

NMBG2136_1513	NMBH4476_0584	**NMA1897**	NMC1557	**NEIS1557**	initiation factor IF2 (infB)
NMBG2136_1372	**NMBH4476_0741**	NMA1692	**NMC1418**	**NEIS1418**	lipoprotein (nlpD)
-	-	**NMA1997**	-	nca	pseudogene
-	-	**NMA0266**	-	nca	pseudogene
-	-	**NMA0267**	-	nca	pseudogene
NMBG2136_0343	NMBH4476_1819	NMA0579	**NMC0342**	**NEIS0342**	putative prolyl endopeptidase
NMBG2136_1183	NMBH4476_0936	NMA1483	**NMC1208**	**NEIS1208**	putative transmembrane transport protein
NMBG2136_0941	**NMBH4476_0480**	NMA1167	NMC1661	**NEIS0661**	replication initiation factor
NMBG2136_0668	NMBH4476_1468	NMA0929	**NMC0672**	**NEIS0672**	threonyl-tRNA synthetase (thrS)
NMBG2136_1246	NMBH4476_1774	NMA1558	**NMC1282**	**NEIS1282**	TonB-dependent receptor protein
NMBG2136_1758	**NMBH4476_0336**	NMA2146	**NMC1829**	**NEIS1829**	tspA antigen protein
NMBG2136_0019	NMBH4476_0681	**NMA1797**	NMC0283	**NEIS0283**	tspB protein (part of MDA island)
NMBG2136_1767	NMBH4476_0327	NMA2156	NMC1839	**NEIS1839**	type 4 prepilin-like proteins leader peptide-processing enzyme (pilD)
C. Synonymous amino acid change					
NMBG2136_1779	NMBH4476_0314	NMA2168	NMC1852	**NEIS1852**	antibacterial fatty acid resistance protein B
NMBG2136_0798	NMBH4476_1326	NMA1073	NMC0801	**NEIS0801**	conserved hypothetical protein
NMBG2136_0441	NMBH4476_1679	NMA0690	NMC0448	nca	conserved hypothetical protein
NMBG2136_1032	NMBH4476_1090	NMA1881	NMC1048	nca	DNA transposition protein (gpB)
NMBG2136_2041	NMBH4476_2092	NMA0241	**NMC2132**	**NEIS2132**	electron transfer flavoprotein alpha-subunit (etfA)
NMBG2136_0784	**NMBH4476_1336**	NMA1063	NMC0852	**NEIS0792**	GTP-binding protein
NMBG2136_0274 *	NMBH4476_0268	**NMA2214**	NMC0832	nca	hypothetical protein
NMBG2136_0605	**NMBH4476_1531**	-	NMC0606	nca	hypothetical protein
NMBG2136_0443	**NMBH4476_1677**	NMA0688	NMC0450	nca	hypothetical protein
NMBG2136_0787	NMBH4476_1333	**NMA1066**	NMC0795	**NEIS0795**	hypothetical protein
NMBG2136_0798	NMBH4476_1314	NMA1073	**NMC0806**	nca	hypothetical protein
NMBG2136_0794	NMBH4476_1319	**NMA1078**	NMC0805	nca	hypothetical protein
NMBG2136_1836	NMBH4476_1885	**NMA0505**	NMC1918	**NEIS1918**	inner membrane transport protein
NMBG2136_1817	**NMBH4476_1867**	NMA0527	NMC1900	**NEIS1900**	lacto-N-neotetraose biosynthesis glycosyl transferase (lgtE)
NMBG2136_2028	**NMBH4476_2080**	NMA0225	NMC2118	**NEIS2118**	phospholipase, patatin family
-	-	-	**NMC0785**	nca	pseudogene
NMBG2136_1630	NMBH4476_0463	NMA2015	NMC1679	**NEIS1679**	putative 2-oxoglutarate/malate transporter
NMBG2136_1820	(NMBH4476_1869)	(NMA0524)	(NMC1902)	nca	putative glycosyl transferase
-	-	**NMA1058**	-	nca	putative glycosyl transferase (pseudogene)
D. Synonymous and non-synonymous amino acid changes					
NMBG2136_1903 ^	NMBH4476_1948	NMA0430	NMC1987	**NEIS1987**	ATP-dependent RNA helicase (hrpA)
NMBG2136_0602	NMBH4476_1533	NMA2115	NMC0599	**NEIS0599**	conserved hypothetical protein
NMBG2136_0948	**NMBH4476_1198**	NMA1174	NMC0957	**NEIS0957**	Flavodoxin-like fold family protein
NMBG2136_0869	NMBH4476_1488	NMA0905	NMC0870	nca	host specificity protein J / IgA-specific serine endopeptidase
NMBG2136_1893	NMBH4476_1938 *	NMA0440 *	**NMC1978**	**NEIS1978**	(nlpC) p60 family protein
NMBG2136_0586	NMBH4476_1542	-	NMC0586	**NEIS0586**	hypothetical protein
NMBG2136_0589	NMBH4476_1541	-	NMC0591	**NEIS0591**	hypothetical protein
NMBG2136_0272	**NMBH4476_0266**	NMA2214	NMC1767	**NEIS0785**	hypothetical protein
NMBG2136_0274 *	NMBH4476_0268	**NMA2214**	NMC0832	nca	hypothetical protein
NMBG2136_0623	**NMBH4476_1513 ^**	NMA0876	NMC0625	**NEIS0625**	hypothetical protein
NMBG2136_0368	**NMBH4476_1793**	NMA0607	NMC0369	nca	hypothetical protein
NMBG2136_0543 ^	NMBH4476_1585	NMA0810	NMC0548	**NEIS0548**	hypothetical protein
NMBG2136_0662 ^	NMBH4476_1474	NMA0921	NMC0666	**NEIS0666**	hypothetical protein
NMBG2136_0820	**NMBH4476_1292 ^**	NMA1641	NMC1340	nca	hypothetical protein
NMBG2136_1296	**NMBH4476_0860 ^**	NMA0458	NMC1340	nca	hypothetical protein

NMBG2136_0303	**NMBH4476_1858 ^**	NMA0537 *	NMC0306	**NEIS0306**	hypothetical protein
NMBG2136_1304 ^	NMBH4476_0811	NMA1626	NMC0527	nca	iron-regulated protein (frpA)
NMBG2136_1817	**NMBH4476_1868**	NMA0525	NMC1901	**NEIS1901**	lacto-N-neotetraose biosynthesis glycosyl transferase (lgtB)
NMBG2136_0361	**NMBH4476_1801**	NMA0600	NMC0361	**NEIS0361**	modulator of drug activity B
NMBG2136_1917	**NMBH4476_1961 ^**	NMA0422	NMC1997	**NEIS1997**	pantetheine-phosphate adenylyltransferase (coaD)
-	**NMBH4476_0626**	-	-	nca	pseudogene
-	**NMBH4476_0496 ^**	-	-	nca	pseudogene
-	**NMBH4476_1855 ^**	-	-	nca	pseudogene
NMBG2136_0779 ^	-	-	-	nca	pseudogene
-	-	**NMA0315**	-	nca	putative MafB alternative C-terminus (fragment)
-	**-**	**NMA0319**	**-**	nca	putative MafB alternative C-terminus (fragment)
-	**-**	**NMA0273A**	**-**	nca	putative membrane protein (pseudogene)
NMBG2136_1188	NMBH4476_0932 *	NMA1491	**NMC1214**	**NEIS1214**	transcription-repair coupling factor (mfd)
E. Assembly failure due to long repeat tract					
NMBG2136_1893	NMBH4476_1938 *	**NMA0440**	NMC1978 *	**NEIS1978**	(nlpC) p60 family protein
NMBG2136_1714	NMBH4476_0383	NMA2099	NMC1778	**NEIS1778**	aldose 1-epimerase
NMBG2136_1785	NMBH4476_0308	NMA2174	**NMC1858**	**NEIS1858**	conserved hypothetical protein
NMBG2136_0927	NMBH4476_1217	NMA1150	NMC0932	**NEIS0932**	dihydrolipoyllysine-residue succinyltransferase (sucB)
NMBG2136_1820	NMBH4476_1869	**NMA0832**	**NMC0568**	**NEIS0568**	glycosyl transferase
NMBG2136_2007	(NMBH4476_2058)	-	(NMC2091)	nca	hypothetical protein
NMBG2136_1361	NMBH4476_0752	(NMA0291)	(NMC0021)	nca	hypothetical protein
-	-	NMA1723	**NMC1452**	**NEIS1452**	hypothetical protein
NMBG2136_0520	NMBH4476_0489	**NMA0784**	NMC1653	**NEIS1653**	membrane protein
NMBG2136_0282	**NMBH4476_0276**	NMA2206	**NMC0276**	**NEIS0276**	peptidyl-prolyl cis-trans isomerase (surA) / rotamase
NMBG2136_1610	NMBH4476_0484	**NMA1996**	NMC1658	**NEIS1658**	periplasmic type I secretion system protein | hemoysin D
NMBG2136_0010	**NMBH4476_0010**	NMA0257	NMC2148	**NEIS2148**	phosphoglycerate kinase
NMBG2136_0563	-	-	-	nca	pseudogene
NMBG2136_1786	-	-	-	nca	pseudogene
-	**NMBH4476_0307**	-	-	nca	pseudogene
-	**NMBH4476_0843**	-	-	nca	pseudogene
-	**NMBH4476_1566**	-	-	nca	pseudogene
NMBG2136_1404	**NMBH4476_0706**	NMA1733	NMC1462	**NEIS1462**	putative lipoprotein
NMBG2136_1188	**NMBH4476_0932**	NMA1491	NMC1214 *	**NEIS1214**	transcription-repair coupling factor (mfd)
NMBG2136_0407	**NMBH4476_1758**	NMA0650	NMC0408 *	**NEIS0408**	type IV pilus secretin (pilQ)
F. Locus is paralogous in the genome					
NMBG2136_0915; NMBG2136_1845	**NMBH4476_1229; NMBH4476_1894**	**NMA1138; NMA0495**	**NMC0920; NMC1928**	**NEIS0920; NEIS1928**	50S ribosomal protein subunit L31 (rpmE)
NMBG2136_0914; NMBG2136_0160	**NMBH4476_1230; NMBH4476_0161**	**NMA1137; NMA0107**	**NMC0919; NMC0154**	**NEIS0919; NEIS0154**	50S ribosomal protein subunit L36 (rpmJ)
NMBG2136_0942; NMBG2136_1614	NMBH4476_0479; NMBH4476_1204; NMBH4476_1694	NMA1168; NMA2000	**NMC0951; NMC1662**	**NEIS0951; NEIS1662**	conserved hypothetical protein
NMBG2136_0121	NMBH4476_0121	**NMA0149; NMA0134**	NMC0116	**NEIS0116**	elongation factor TU (tufA2)
-	**NMBH4476_1709; NMBH4476_1706**	-	-	nca	hypothetical protein
NMBG2136_0604	NMBH4476_1532	NMA0036	**NMC1803; NMC0223**	nca	hypothetical protein
-	**NMBH4476_1692; NMBH4476_0736**	NMA0594	NMC1424	nca	hypothetical protein
NMBG2136_0607	NMBH4476_1538; NMBH4476_1529	NMA0856; NMA0858; NMA0857	**NMC0595; NMC0608;** (NMC0594)	**NEIS0595; NEIS0608**	hypothetical protein

NMBG2136_0967	NMBH4476_1082	**NMA1193; NMA1859**	NMC0973	**NEIS0973**	hypothetical protein
NMBG2136_1605	NMBH4476_0489	**NMA1989**	NMC1653	**NEIS1653**	hypothetical protein
NMBG2136_1730; NMBG2136_1733; NMBG2136_1729	**NMBH4476_0361; NMBH4476_0367**	NMA2118	NMC1796; NMC0600; NMC1795	**NEIS1796; NEIS0600; NEIS1795**	hypothetical protein
NMBG2136_0023	**NMBH4476_0593; NMBH4476_0686**	NMA1793	NMC0030	nca	hypothetical protein
NMBG2136_1358	NMBH4476_0755	NMA0862	**NMC1403; NMC1551**	nca	hypothetical protein; (opaA)
NMBG2136_1658	NMBH4476_0591; NMBH4476_0435	(NMA0862)	**NMC1719; NMC0903**	**NEIS1719**	Opa1800 outer membrane protein (opaJ); hypothetical protein (opaD)
NMBG2136_0787; NMBG2136_0518	**NMBH4476_1333; NMBH4476_1608**	NMA1066; NMA1066	NMC0795; NMC0524	**NEIS0795; NEIS0524**	peptidase, C39 family
NMBG2136_0795	**NMBH4476_1325; NMBH4476_1315**	NMA1047	NMC0802	nca	putative lipoprotein

hypothetical protein: a protein whose existence has been predicted, but for which there is no experimental evidence that it is expressed in vivo

putative protein: a region of genomic DNA is predicted to contain a protein-coding gene, ab initio gene finding with extrinsic evidence that the gene is functional

pseudogene: dysfunctional genes that have lost their protein-coding ability or are otherwise no longer expressed

[#] Definition of locus may vary between finished genomes depending on the gene finding algorithim and/or annotation software used; listed definition is specific to the genome listed.

* The same locus, in two separate genomes, has different sequence or assembly issue.

^ Indicates that that the sequence discrepancy was caused by an insertion/deletion.

nca NEIS locus tag identifier currently not assigned, however locus is being curated for future addition to the sequence definition database.

Additional Table 3 Sequence discrepancy categories

single nucleotide sequence change per allele resulting in	**number of alleles affected per isolate**			
	Z2491	**FAM18**	**H44/76**	**G2136**
non-synonymous amino acid change	8	5	4	3
synonymous amino acid change	3	3	5	7
multiple nucleotide sequence change per allele resulting in	**number of alleles affected per isolate**			
	Z2491	**FAM18**	**H44/76**	**G2136**
non-synonymous amino acid change	4	1	1	3
synonymous amino acid change	1	0	0	0
non & synonymous amino acid change	4	2	1	0
insertion/deletion change per allele resulting in	**number of alleles affected per isolate**			
	Z2491	**FAM18**	**H44/76**	**G2136**
single insertion/deletion	0	0	3	3
double insertion/deletion	0	0	3	3

Chapter 4 Lineage 3 genomic characterisation, the largest and most diverse meningococcal clonal complex, ST-41/44

1. Introduction

The biomass of the bacterial domain has been estimated to be the largest reservoir of biodiversity on earth [85]. The debate surrounding the bacterial species concept is ongoing. Genotypic characterization has been used to augment the classic taxonomic descriptions of morphology, physiological and metabolic features, changing the way we view the classification of bacteria. The addition of genetic information to taxonomy is, in general, an accepted reference point for classification [86]. *Neisseria meningitidi*s has benefited from the use of genetic typing and classification schemes [11, 87-89]. Two of the many challenges surrounding the bacterial species concept are the diverse phenotypic differences found within some bacterial strains and the definition of distinct but highly related groups of strains (lineages) defined together as a species. *N. meningitidis* is one such organism that epitomizes these two challenges and many others.

The meningococcus, an obligate human Gram-negative, oxidase positive diplococcus, is usually a commensal inhabitant of the oropharynx but very rarely causes disease. A small proportion of these strains are responsible for invasive meningococcal disease and are referred to as hyperinvasive [90]. The relationship among strains circulating within the meningococcal population is not entirely understood. What is known is that isolates recovered from asymptomatic carriers are genetically more diverse than those isolated from invasive disease cases [91]. Both commensal and hyperinvasive meningococci have been found to express a polysaccharide capsule - a virulence factor linked to disease potential [92]. Additional factors that enable the transition from asymptomatic carriage to virulent disease are also poorly understood.

Genetic typing of the meningococcus, first introduced in the 1980s in the form of multilocus enzyme electrophoresis (MLEE) helped to establish the basic meningococcal population structure. It also established the association of the hyperinvasive disease phenotype with

particular strain types. Unfortunately, like serological characterization, which uses the capsular polysaccharide (serogroup) and two of the major antigen proteins PorB and PorA, the MLEE characterization does not necessarily specify genotypic similarity [93-95]. Following the MLEE technique a related method, multilocus sequence typing (MLST), was developed in the late 1990s [26]. The MLST method became a routine typing procedure in both the research and the clinical setting [93]. When the MLST model is applied to a genomic level of analysis, a high resolution of the data is obtained [87, 96]. Sequence-based typing methods are ideal for their portability and provide a standardized method of investigation and characterization.

The aim of this study was to use a gene-by-gene MLST approach to investigate the population structure of *N. meningitidis* using genomes from a large, diverse group of related strains to elucidate the evolution of the group. This group of strains is commonly recognized as the ST-41/44 clonal complex (41/44cc) known to cause hyperinvasive disease but also responsible for a large population of strains carried asymptomatically [97]. In this respect, the 41/44cc reflects the larger meningococcal population as a whole. Until whole genome sequences became available, the genetic basis for the difference in phenotype within this lineage was largely unexplored. Questions about this species also necessitate additional consideration in population and evolutionary terms; including those that underlie different pathogenic potentials, such as adaptation, niche exploitation, and host response [98]. The diverse and polygenic nature of the *N. meningitidis* genome make this organism one of the best candidates for investigation using population genetic analysis. To better understand the evolution of the population structure of the 41/44cc, a basic historical context for this particular group is also beneficial.

1.1. The emergence of a new hyperinvasive antigen type

In the late twentieth century an epidemiological investigation in The Netherlands identified a serogroup B strain with a specific antigen profile that was responsible for the observed increase of invasive systemic meningococcal disease [99]. The strain, characterized at the time as PorB 4, PorA P1.4, was a new antigen profile. The main finding related to the

appearance of this new antigen type was an epidemiological shift in the age distribution from a younger age group to an older age group, adversely affecting young adults between the ages of 15 and 20 years. [100, 101]. The occurrence was unusual for a serogroup B associated strain, which until then was only known to cause sporadic cases and localized outbreaks. This particular strain continued to cause increased disease and appeared across Europe during the 1980s [100, 102, 103]. A retrospective study looking at serogroup B invasive meningococcal disease between 1975 and 1995 found the same Dutch strain emerged in the UK around 1985 [104]. It caused 5% of the observed UK disease at that time and was associated with the same age group as observed in The Netherlands. Throughout the 1980s and 1990s, disease rates caused by this particular antigen type continued to increase across most of Europe and the UK, reaching epidemic and sub-epidemic levels in certain regions [105-111]. Throughout the twentieth century the highest rates of disease were reported in the UK, the Czech Republic, Norway and Greece; additionally large outbreaks in the United States, Ireland, France, and Germany also occurred. At the same time, the Dutch antigen type appeared in New Zealand where a decade-long epidemic began [111].

Since 1999, the European Union Invasive Bacterial Infections Surveillance Network (EU-IBIS) tracked reports of meningococcal disease from twenty-seven countries [112]. Since that time the rates of invasive meningococcal disease have been declining, but the Lineage 3 strains continue to cause the majority of serogroup B disease. More recently a unique 41/44cc multi-year outbreak in an extended Irish Traveller family from 2010 to 2013, spanning three geographical regions in the Republic of Ireland, has also been documented [113]. The historically recent appearance of this antigen type and its association with invasive disease presents a unique opportunity to explore the evolution and population structure of this diverse organism.

1.2. A note on nomenclature

As briefly outlined in the introduction of this chapter, an agreement on defining a species is not only ongoing but ever-changing; as are the methods that are used to determine the classification. Likewise, how groups of meningococcal strains are referred to has also undergone changes. The characterization of isolates has been designated based on targeted typing motifs, and the nomenclature was designed to confer the maximum amount of information in the shortest possible format. The MLEE, MLST, and even the antigen type designations are somewhat static and have very limited power to resolve or address questions beyond the specific character that they were used to define. The increased use of whole genome sequence data for both clinical and research purposes also requires a standardized set of terms with which to communicate and convey information. The designation of a particular strain or group of related strains should reflect the depth of the data used, and unused, to derive the grouping and the inferences that are possible (the resolution). In short a hierarchical nomenclature that immediately tells a reader the extent of supporting data used and what information is available for interrogation. This proposal was made to the *Neisseria* community for consideration at the XIXth International Pathogenic *Neisseria* Conference (IPNC) in 2014 during the PubMLST.org user's group meeting.

A full list of suggested lineage designations and their associated clonal complex is found in Chapter Three (Bratcher *et* al 2014, BMC Genomics, Figure 4). The format uses the term 'Lineage' followed by a numeral when the derived information is from whole genome sequence data, and the 'clonal complex' nomenclature should be used when data is from seven-locus MLST only. Therefore in this chapter, except where the clonal complex designation is needed to clarify a distinction, 'Lineage 3' will be used to refer to genomically derived data or results for the clonal complex ST-41/44. In the same way, the 'clade' designation proposed by Budroni *et. al.* in 2011 and is utilized the same way here [44]. The term clade will be used to designate lineages that are more closely related to each other

and share a recent common ancestor that is intermediate to the common meningococcal ancestor.

2. Isolate collection and strain designation

The *Neisseria* PubMLST.org database was queried, first in 2011 and again in 2014, for all isolate records that were assigned to the 41/44cc and contained genomic sequence data. The typing information, sequence bin, and provenance data were downloaded and categorized. A total of 1254 records were assigned to the complex, and after filtering 116 duplicate records were removed. Two additional database queries were run to determine if: (i) any incomplete MLST profiles could potentially belong to 41/44cc; and (ii) if any ST unassigned to a clonal complex could be a 41/44cc strain. There were no incomplete MLST profiles that would have contained the required four out of seven allelic matches, even if the profile were complete. Twenty-one STs unassigned to a currently defined clonal complex could be assigned to the 41/44cc. The list of isolates was then queried to filter out isolate records that did not have genomic data (174 records) or where genomic data had poor sequence assembly coverage (50 genomes). The coverage was assessed based on two general criteria; (i) the number of complete *N. meningitidis* core gene (cgMLST) alleles found in the assembly; and (ii) the presence of a whole ribosomal protein profile (rMLST). Any genomes with over 100 incomplete cgMLST alleles or any genomes with incomplete rMLST alleles were removed from the analysis. The 174 isolate records without genomes were cultured, their DNA extracted and draft genome sequences obtained. In total 1109 isolate records and genomes were included in the study.

2.1. Available isolate record data

A full list of the isolates is found in the Appendix A, page 106. Provenance and clinical data were available for most of the isolates. The majority of the isolates were from the UK (63.9%), and the remaining were from 21 countries across the globe. The collection spanned almost three decades from 1986 to 2014, with 90.7% of isolate records listing the year of isolation. There were only 102 isolate records with age-related information. The records

contained in PubMLST.org (as of 2014) were biased towards isolates recovered from incidents of disease, 78.2%, and 1.3% did not have a disease association recorded. The isolation source of 61.6% of the isolates was recorded. The isolates were recovered from a variety of sources: throat swabs (245 isolates), CSF (89 isolates) and blood cultures (310 isolates); as well as more uncommon sites such as nasal swabs (4), joint fluid (2 isolates), sputum (1 isolate), rectal and urethral swabs (3 isolates). The remaining isolate records, 453 (40.8%), list the source of isolation as unknown. Seroagglutination results were recorded for 87% of the isolates and included 864 B, 24 C, 3 E, 2 W, 2 Y, 1 Z, and 70 isolates that were as non-groupable.

2.2. New genomes

Based on the successful use of *de novo* assembled genomes from short-read sequence data for whole genome analysis presented in Chapter Three the same methods were applied to the 174 isolates that required sequencing for this study. The isolates were retrieved from the freezer archive, cultured, and their DNA extracted as described in the Methods sections 2-4, page 14-15. The prepared DNA samples were sent to The Sanger Institute at the Wellcome Trust Genome Campus, UK for whole genome sequencing. Once the short-read sequence files were obtained and assembled into draft genomes, Methods section 6 page 17. There was no further manipulation of the assembled genome sequences. The genome assembly statistics were generated as part of the automated pipeline once assembly of the genomes was complete. The sequence assemblies were uploaded into the *Neisseria* PubMLST.org database, linked to their isolate record and typed by extracting the seven-locus MLST loci sequences from the assembled sequence contigs. This check was done before beginning the analysis to verify that the MLST profile originally obtained for each isolate was the same as the genome-derived profile, and the isolates belonged to the 41/44cc.

3. Genome analysis

The use of short-read whole genome sequencing and *de novo* assembly are both still recently developed methods. The inherent issues of each method and the potential effect on the subsequent genome assembly require ongoing analysis, including familiarization with the genomic structure and sequence anomalies of the meningococcal genome. The sequence architecture of the genome includes phase variable genes with variable homopolymeric tracts or variable number tandum repeats (VNTRs), paralogous genes, as well as insertion and mobile genetic elements. The evaluation of the assembled draft genomes was continuously done throughout the study. Each of the sub-sections in this chapter will begin with an overview or provide specific methodological details not covered in the general methods sections, and the results will follow.

3.1. Genome coverage

The database was queried for all of the selected isolates using the unique PubMLST.org isolate identification number assigned to each isolate and the BIGSdb 'sequence bin breakdown' tool to extract the assembly statistics of all 1109 genomes. The assemblies were within the established parameters of finished *N. meningitidis* reference genomes (Table 1). The average sequence length was 2,221,413 nucleotides with a GC content of 51.6%, and the assemblies contained an average of 242 contigs. There were 127 genomes assembled from 79 base reads which contained an average of 300 to 400 contigs per genome, and a second group of genomes that used libraries containing suboptimal DNA preparations and assembled 420 to 580 contigs, even though 100 base read libraries were created. The majority (88.5%) of the genome assemblies were, however, constructed from 100 to 150 base reads and contained an average of 100 to 280 contigs. Short-read lengths are the main limiting factor for assembling larger contigs for meningococcal genomes. The production of the sequence data spanned two formulations of sequencing chemistry and an equipment upgrade from the original Illumina GAII platform to the HiSeq 2000 platform.

The resulting increase in the length of the contigs assembled, and a corresponding drop in the number of contigs was observed when the platform upgrade and the two chemistry changes took place (Figure 1).

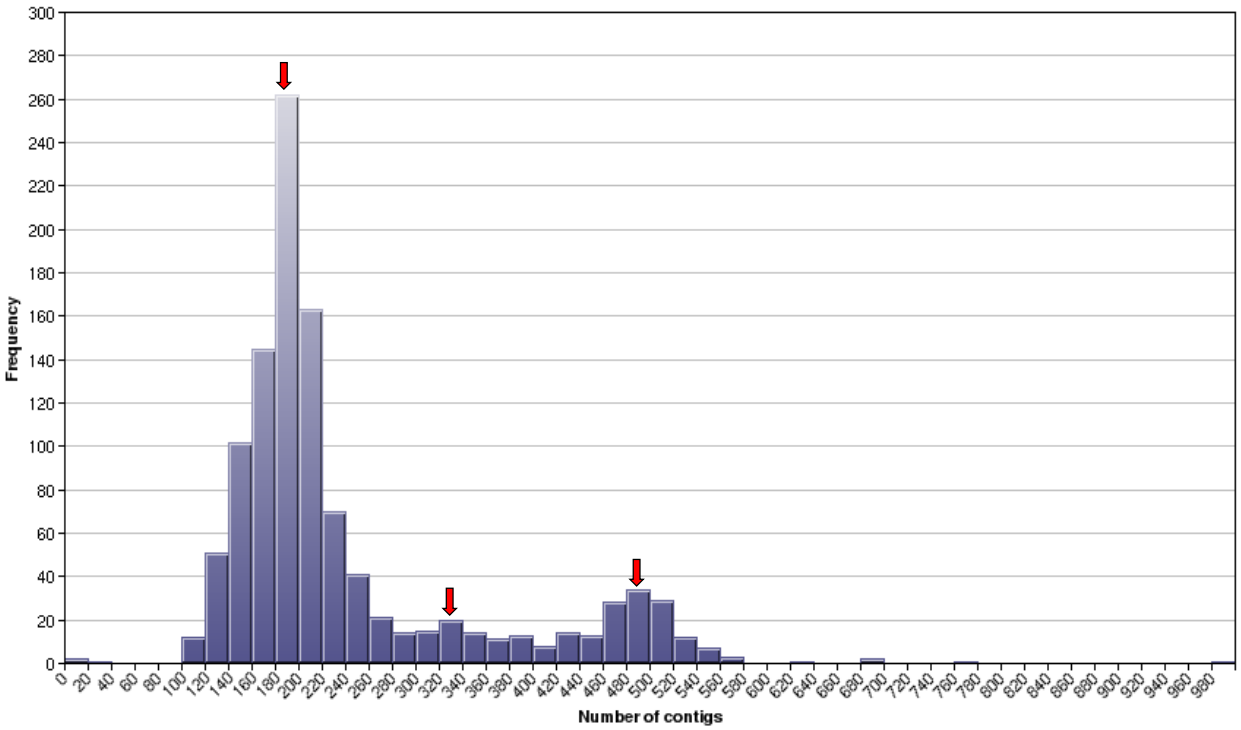
Table 1: Sequence assembly statistics

The two published reference genomes for Lineage 3 are 2.22 to 2.25 Mb in size, contain 1995 to 2037 annotated loci, and have a GC content of 51.3%. All draft genome assemblies fell within these ranges for total sequence length, GC content and expected number of loci for a draft genome. The N50 is the number of contigs whose length when summed up accounts for 50% or more of the genome. The L50 represents the length of contigs of equal or longer length that contain at least 50% of the assembled sequence across the genome.

	contig count	total length	% G+C	N50	L50
average	242	2,221,413	52	24	41,071
max*	998	6,564,763	53	76	2,248,966
min	1	2,015,543	51	1	8,332

* 21 genomes were sequenced twice, and three genomes were sequenced three times

Figure 1: Histogram showing the change in contig count of the *de novo* assembled genomes. The meningococcal genome contains several homopolymeric tracts, both simple and complex repeats. Short-read lengths are the main limiting factors for assembly of larger contigs. Over the course of this study, the sequencing chemistry was improved twice resulting in a decrease in the number of contigs. The peak centered around 190 contigs are genomes that were assembled using 100-150 base reads. The secondary peak centered around 490 contigs are assemblies that used 100 base read, however, the DNA library preparations were substandard. Nevertheless, they produced usable assemblies for whole genome analysis. The slight increase around 330 contigs are genomes sequenced using the less advanced chemistry and represent the genomes assembled using 79 base-reads.



3.2. Loci annotation and allele curation

The BIGSdb platform software includes two modules, the 'autotagger' and 'auto definer' that were used to annotate the genome sequences using the 'NEIS' locus tag identifier [114]. The NEIS number is a reference free identifier that cross references genome sequence annotations (loci) from multiple sources using a single locus tag. The locus tag identifier of the published reference genomes were recorded as an alias in the database. Each of the 1109 assemblies was scanned against the loci in the sequence definition database to detect known alleles, and to determine and assign new alleles with a sequence identity greater than

or equal to 98% to known alleles. Identified sequences were tagged with a NEIS locus tag identifier and assigned an allele number. New alleles that had a sequence identity greater than 85% (but less than 98% of a defined locus) were manually checked by extracting and aligning the new sequence to the defined locus alleles in the sequence definition database. The alignment procedure was done in MEGA6 using both nucleotide and amino acid alignments [115]. New sequence variants that had full-length coding sequences were assigned new allele numbers. The remaining identified loci that did not adhere to the locus sequence match parameters, but had at least a 70% sequence identity match, were also manually assessed and curated. The contigs containing the locus allele in question were visually checked by alignment comparison against a known locus assembly using Artemis to ensure the correct locus was being identified and annotated [116, 117]. The alleles with full coding sequences were manually curated as explained above, and also checked using the online programs KEGG and PATRIC to confirm the location and functional grouping or protein family domain [37, 38]. Locus tag identifiers and allele numbers were assigned to each new locus sequence defined if there was support provided by the online database information for the identification of the locus.

The genomes contained loci with indels, internal stop codons or altered start codons; all of which caused a frameshift in the sequence being identified. These alleles were curated as previously described, defined in the sequence definition database, and given an allele number and a flag to indicate one or more of the issues with the sequence allele. In the event that an identified sequence could not be given an allele number, the sequence was not given a locus tag identifier. During the scanning process, the BIGSdb software indicated which loci it identified as incomplete. These loci had 70% or more of the locus sequence present but are found at the end of a contig. The incomplete assembly of a locus occurs when the sequence architecture is incompatible with the short-read sequence assembly method. Sequence regions that failed to assemble contained homopolymeric tracts or VNTRs (simple or complex repeats) that were longer than the read length used to

assemble the genome. The incomplete loci were manually checked, and each locus tagged in the appropriate genome and the sequence flagged as present but incomplete. No allele was assigned to these loci; however, the BIGSdb scan provides an allele reference to the closest partial match, but this was not used in any analysis. Once the manually curated alleles were uploaded to the sequence definition database and the incomplete loci tagged, the genomes were once again searched and another round of sequence tagging was performed. This process was done iteratively gene-by-gene until all known NEIS defined loci had been curated, and as many alleles as possible had been defined and assigned.

The sequence coverage of all 1109 genomes was assessed a second time based on the level of the annotation using the BIGSdb platforms 'sequence bin breakdown' analysis option (Table 2A). The genomes were evaluated using the cgMLST scheme, described in Chapter Three, which contained 1605 loci defined as being core to the meningococcal species [96]. The expected number of core meningococcal loci tagged, >1500 loci for a draft meningococcal genome, was found in 1091 of 1109 (98.4%) of the assemblies. The expected number of tagged core genes is the minimum number of loci needed to make a genome level comparison and is based on a 75% genomic match among meningococcal genomes. Eighteen genomes had 1442 to 1521 loci tagged, however were included in the remaining analysis as a conservative representation of 'poor' draft genomes. The lower number of loci identified in these assemblies was caused by the reasons presented previously, or the loci were not found in the assembly. However, the eighteen genomes were fully annotated with allele numbers for over 99.9% of the tagged loci.

Table 2: Genome annotation level. Each genome assembly was assessed using (A) the *N. meningitidis* cgMLST (V1.0) of 1605 loci; and (B) the Lineage 3 specific core scheme 1740 loci. There are 40 cgMLST loci that were not core to Lineage 3 and 179 of the Lineage 3 loci are not core to the meningococcus as a species.

A. *Neisseria meningitidis* species core genome: cgMLST (V1.0)

	number of loci tagged	%	number of alleles designated	%	% of tagged loci with alleles designated
Average	1587	98.9	1565	97.5	98.6
Max	1603	99.9	1598	99.6	100.0
Min	1442	89.8	1441	89.8	95.4

B. *Neisseria meningitidis* Lineage 3 core genome: L3 cgMLST (V1.0)

	Loci tagged	%	Number of alleles designated	%	% of tagged loci with alleles designated
Average	1703	97.8	1680	96.5	98.7
Max	1723	98.9	1718	98.6	100.1
Min	1551	89.0	1549	88.9	95.7

3.2.1. Identification of the core Lineage 3 genes

The BIGSdb Genome Comparator tool was used to compare 1088 of the genomes, those that were identified as belonging to the 41/44cc to, to each of the published reference genomes NZ-05/33 and M01-240149 [44]. Both of these reference genomes are closed, fully annotated, typed and assigned to the 41/44cc. The comparison was done by extracting the annotated coding sequences of the reference genomes from the GenBank electronic file and comparing each of the annotated locus sequences against the assembled contigs for each of the isolate genomes. The comparison output file contained a tabulated breakdown categorizing the loci as: (i) identical loci; (ii) diverse loci; (iii) paralogous loci; and (iv) loci that were missing and/or incomplete. Additional settings selected for the analysis included alignment of all loci using MAFFT, determination of the core threshold set at 100%, and the calculation of the mean distance for each locus among all genomes. The comparison output identified the conserved loci and generated a list of loci that met the 100% core threshold and those that did not. The list of reference genome loci and their sequences were then cross-

referenced with the locus sequences in the *Neisseria* sequence definition database to obtain the corresponding NEIS locus tag identifiers.

The BIGSdb Genome Comparator tool, using the NZ-05/33 genome as a reference, identified 1410 loci as being present in all 1088 genomes, and an additional 439 loci present in 95% of the 1088 genomes. Similarly, using the M01-240149 reference genome, the core threshold of 100% was met by 1399 loci, and an additional 431 present in 95% of the 1088 genomes. For both reference genomes, the number of loci identified in the 1088 assemblies established that 92% of their genomes were shared. The two loci lists were compiled, cross-checked and filtered for problem loci. These included paralogous loci and loci that did not assemble well using short-read data, unidentified pseudogenes, and loci known or thought to be phase variable. There were 87 annotated loci contained in the reference genomes that were considered core but had no locus equivalent defined in the *Neisseria* sequence definition database. Investigation of the loci from the reference genomes not found in the sequence definition database revealed nine undefined pseudogenes and five transposase genes; all of which were excluded in any further analysis. The remaining 73 loci were identified as hypothetical proteins, and further comparison of the sequences using online database searches was done in an attempt to identify the sequence, functional grouping and/or the protein family domain. The hypothetical loci that were over 200 bases in length, with a full coding sequence, were checked using the online programs KEGG and PATRIC, no additional information was found. Locus tag identifiers were assigned to each new locus sequence defined and added to the sequence definition database as hypotheitical loci. The genomes were scanned again and tagged for each new hypothetical locus and allele numbers were assigned.

A list of Lineage 3 core genes composed of 1740 loci was identified using the two reference genomes and the methods described above. A lineage-specific core genome scheme was created in the *Neisseria* PubMLST.org database, L3 cgMLST. The scheme

was then used to curate all 1109 genomes to ensure all defined loci were identified and annotated, and to facilitate a genome-wide comparison of the genomes belonging to Lineage 3 (Table 2B). In total 1391 core loci were present in all 1109 genomes and 349 additional loci were present in at least 95% of the 1109 Lineage 3 genomes.

3.2.2. Identification of the accessory genes

When the BIGSdb Genome Comparator tool was used to determine which loci were core using the reference genomes, it also identified the loci that were present in <95% of the genomes. The majority of loci (78.4%) could be defined in six categories: 137 hypothetical proteins; 68 gonococcal genetic island genes; 45 conjugative plasmid genes; 31 VirB T4SS genes; 16 Maf family proteins; and 15 phage-related genes. Among the remaining loci the genes encoding the *hpuA/B*; the T and B cell-stimulating protein B (*tspB*); the FrpC operon proteins; the opacity family proteins; and the translation elongation factors were found. As previously described, the hypothetical loci with full coding sequences over 200 bases in length were checked using KEGG and PATRIC in an attempt to identify the sequence, functional grouping and/or the protein family domain. Locus tag identifiers and allele numbers were assigned to each new locus sequence defined if the locus was found by BLAST search in ≥1.0% of the Lineage 3 genomes. These loci were classified as belonging to the Lineage 3 accessory genome and were distributed throughout the genomes of the lineage.

3.2.3. Identification and annotation of novel genes

The isolate record information associated with the 1109 Lineage 3 genomes was surveyed to identify the uncommon variations in the isolates or their provenance. The survey was done to obtain a subset of genomes in which to look for potential novel genes. The resulting list included 20 non-serogroup B isolates, including seven *cnl* isolates, and 21 isolates not recovered from throat swabs, blood or CSF. The genomes were analyzed using PROKKA gene finding software. The coding regions of the NZ-05/33

reference genome were translated and used during the PROKKA scan and annotation to remove coding regions already annotated. The open reading frames most commonly found were annotated by PROKKA as hypothetical proteins; at least 50 were identified in each genome. The PROKKA program consistently found and annotated in all 41 genomes a single locus encoding a novel hemagglutinin protein and three separate hemagglutinin/hemolysin-like proteins that were not defined in the *Neisseria* sequence definition database. PROKKA also putatively identified six unevenly distributed loci among the genomes. These included an ABC transporter, a DNA-binding transcriptional regulator, a SMMI1/KNR4 family protein, the DNA-damage-inducible protein D, the ribonuclease YeeF, a ScaI restriction endonuclease, and the HTH-type transcriptional regulator, YybR.

4. Lineage 3 population structure and variation

4.1. Core allelic variation

A comparison of the Lineage 3 using the BIGSdb Genome Comparator tool was run using the L3 cgMLST scheme. The threshold for including loci in the core scheme was ≥95% presence to address the draft (or incomplete) nature of the genomes being used for the study. The loci and the scheme consequently contain a level of variation that was associated with this threshold parameter (Figure 2A). Analysis of the comparison output report showed that 80% (1391/1740) of the loci were present in all genomes (Figure 2B). The distance estimates, also referred to as *p*-distance, of the 1391 loci ranged from 0.000 to 0.966. There were 46 highly conserved genes, *p*-distance estimate 0, encoding proteins for 14 hypothetical proteins; nine genes involved in genetic information processing, transcription/translation and DNA replication, repair and recombination; eight ribosomal protein genes; four genes involved in metabolism; three membrane-associated proteins; one iron acquisition protein; and one gene encoding an oxidoreductases enzyme. The majority of the core loci, 798 (55.4%), had a *p*-distance estimate between 0.001 and 0.009 while 455 loci (31.6%) had distance estimates

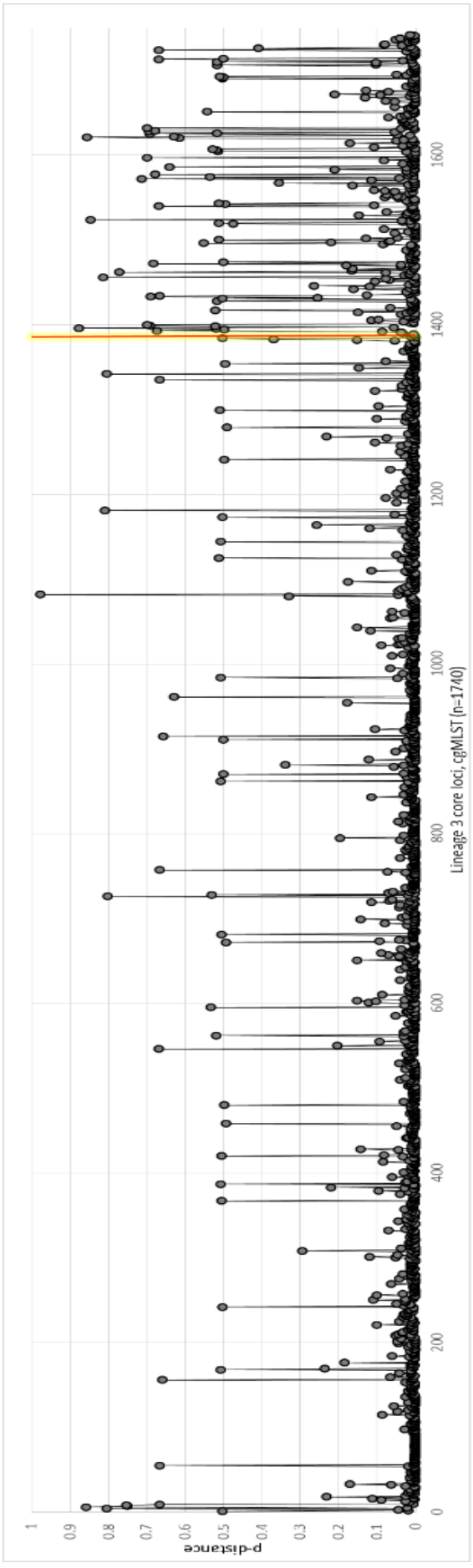
between 0.010 and 0.049. A much smaller number of loci, 75 (5.2%), had *p*-distance estimates between 0.050 and 0.499; and those loci that were highly variable, *p*-distance estimates between 0.500 and 0.966, included 67 loci (4.6%). The gene with the highest *p*-distance, 0.966, was transferrin-binding protein B (*tbpB*) NEIS1691, a surface-exposed lipoprotein and part of the iron acquisition pathway. Further investigation of this locus revealed the lineage contains both isotypes of this gene. The isotypes are differentiated by length, specifically the number of indels. Isotype 2 (the longer sequence) of the *tbpB* gene is most prevalent; however 13.5% of the Lineage 3 genomes contain isotype 1, the shorter sequence. No correlation was found between the isotype, isolate provenance (including invasiveness), serogroup, antigen type, or rST. The highest *p*-distance estimates were categorized as: hypothetical proteins (38.8%), ribosomal protein genes (29.9%), metabolism and biosynthesis pathway genes (13.4%), and genes encoding metabolic enzymes (7.5%). The remaining loci included the genes encoding a TonB-dependent receptor - an iron acquisition protein (NEIS0387); an outer membrane lipoprotein (NEIS0010) and two phage-related proteins (NEIS0970 and NEIS2463). One locus, NEIS1193, is a type III restriction-modification system endonuclease protein that encodes a hydrolase enzyme that catalyzes the endonucleolytic cleavage of DNA to give specific double-stranded fragments with terminal 5'-phosphates [118]. The enzyme acts with site-specific adenine-specific DNA-methyltransferase (NEIS2379) as part of a complex [119]. The NEIS2379 locus was highly conserved. However, it was not found in 33 genomes. The provenance of these 33 isolate genomes had no associations. The genomes that did contain the NEIS2379 locus (98.1%) were allele 1. Currently, there are only fourteen alleles in the *Neisseria* sequence definition database for NEIS2379, and nine were found among the Lineage 3 genomes.

The categorization of the loci present in at least 95% of the genomes was more diverse but dominated (41.3%) by hypothetical proteins. The list also included membrane-associated proteins, pilins and adhesins (35 loci), genetic information and processing, including glycosylation and restriction-modification systems (30 loci), and phage-related

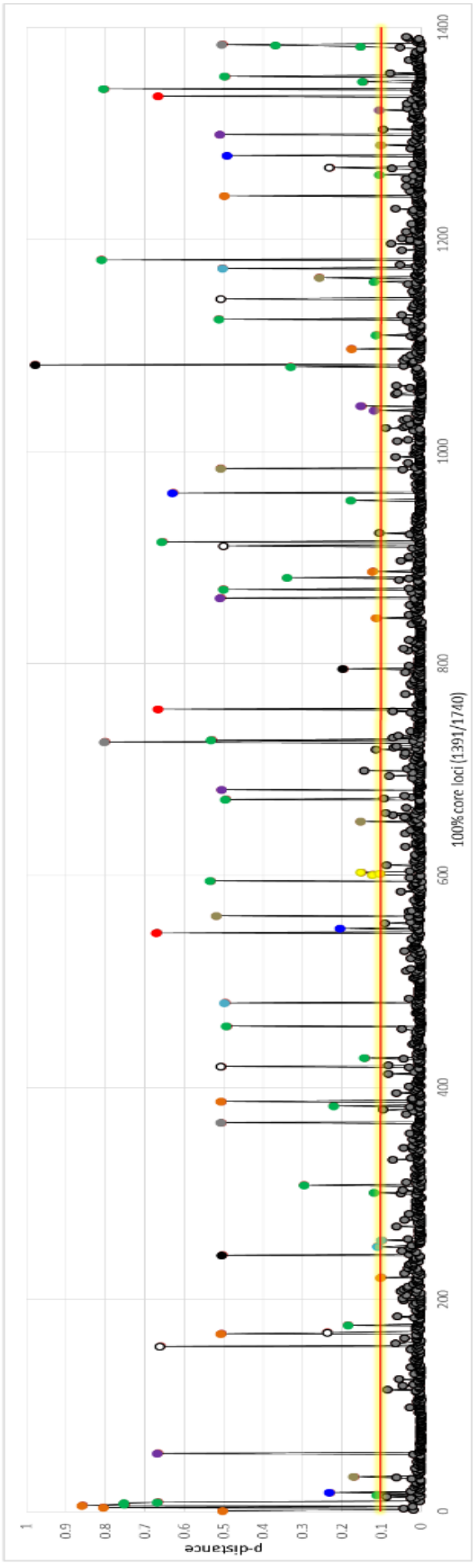
proteins (22 loci). Other significant loci included seven genes encoding iron and two zinc acquisition proteins. When separated from the loci found in all of the genomes, these loci had a higher mean *p*-distance estimate, 0.112, compared to loci present in all genomes (*p*-distance value of 0.032).

Figure 2 (on page 47): Variation, by *p*-distance, among the Lineage 3 core gene scheme (L3 cgMLST). Scatter plot with L3 cgMLST loci on the X axis and *p*-distance values on the Y axis. (A.) 1740 L3 cgMLST loci sorted into those loci present in 100% of genomes, n=1391, estimated *p*-distance 0.032, left of the yellow bar; and loci present in ≥95% but <100% of the genomes, n=349, estimated *p*-distance 0.112, right of the yellow bar. (B.) The 1391 loci present in all Lineage 3 genomes; those loci with a *p*-distance >0.100 are colour coded. Light green: hypothetical protein, red: ribosomal protein gene, yellow: phage-related, black: iron acquisition or zinc transport, teal: adhesion or pilin associated gene, orange: metabolism pathway, purple: genetic information and processing, blue: environmental information processing, light brown: membrane associated, white: enzyme or defensin.

2A.


p-distance
Lineage 3 core loci, cgMLST (n=1740)

2B.


p-distance
100% core loci (1391/1740)

4.2. Ribosomal sequence typing

4.2.1. Background

The ribosomal protein genes are vital for the protein biosynthetic mechanism; they are found in the three domains of life and are highly conserved within each domain [120]. The ribosomal multilocus sequence typing (rMSLT) database is a multispecies database that indexes the variation of the ribosomal protein encoding genes across the bacterial domain [121]. The rMLST database, like the MLST database, indexes curated reference sequences to identify gene variants and assign each variant an allele number. The combination of the ribosomal protein gene alleles is then used to summarize a ribosomal protein sequence type (rST). The data, in the form of allelic profiles or concatenated sequences, can be used to model the phylogenetic relationships among strains, species, genus and higher taxonomic groups [87, 121-124]. The variation and universal presence of the ribosomal protein genes, and their reproduction of population structure within and between species, permits a computationally non-intensive way of visualizing large collections of whole genome data [35, 96]. The rMLST database uses the generic 'BACT' locus tag identifier, short for the 'bacterial domain'. For continuity, though, the ribosomal protein genes will be referred to by their gene name and/or their 'NEIS' locus tag identifier as found in the *Neisseria* PubMLST.org database. A full list of BACT and NEIS equivalent locus tag identifiers is listed in Appendix C, page 133.

4.2.2. Lineage 3 rMLST

The meningococcal genome contains 56 ribosomal protein genes unevenly distributed around the genome. The main grouping of loci, 57.1%, were centered downstream of the origin of replication within a sequence region spanning about 30,000 nucleotides. The remaining 24 loci were scattered around the genome singly or in groups of two to four loci on both the forward and reverse strands. The groupings are conserved and the pattern appeared among all the genomes. The meningococcal ribosomal protein

gene complement contains paralogous isotypes of the 50S L3 (*rpmE1/rpmE2*) and the 50S L36 (*rpmJ1/rpmJ2*) ribosomal protein genes. All four loci are defined in the *Neisseria* pubMLST.org sequence definition database, NEIS1928/NEIS0920 and NEIS0154/NEIS0919 respectively. The remaining 52 ribosomal protein genes are not known to have paralogs within the meningococcal genome.

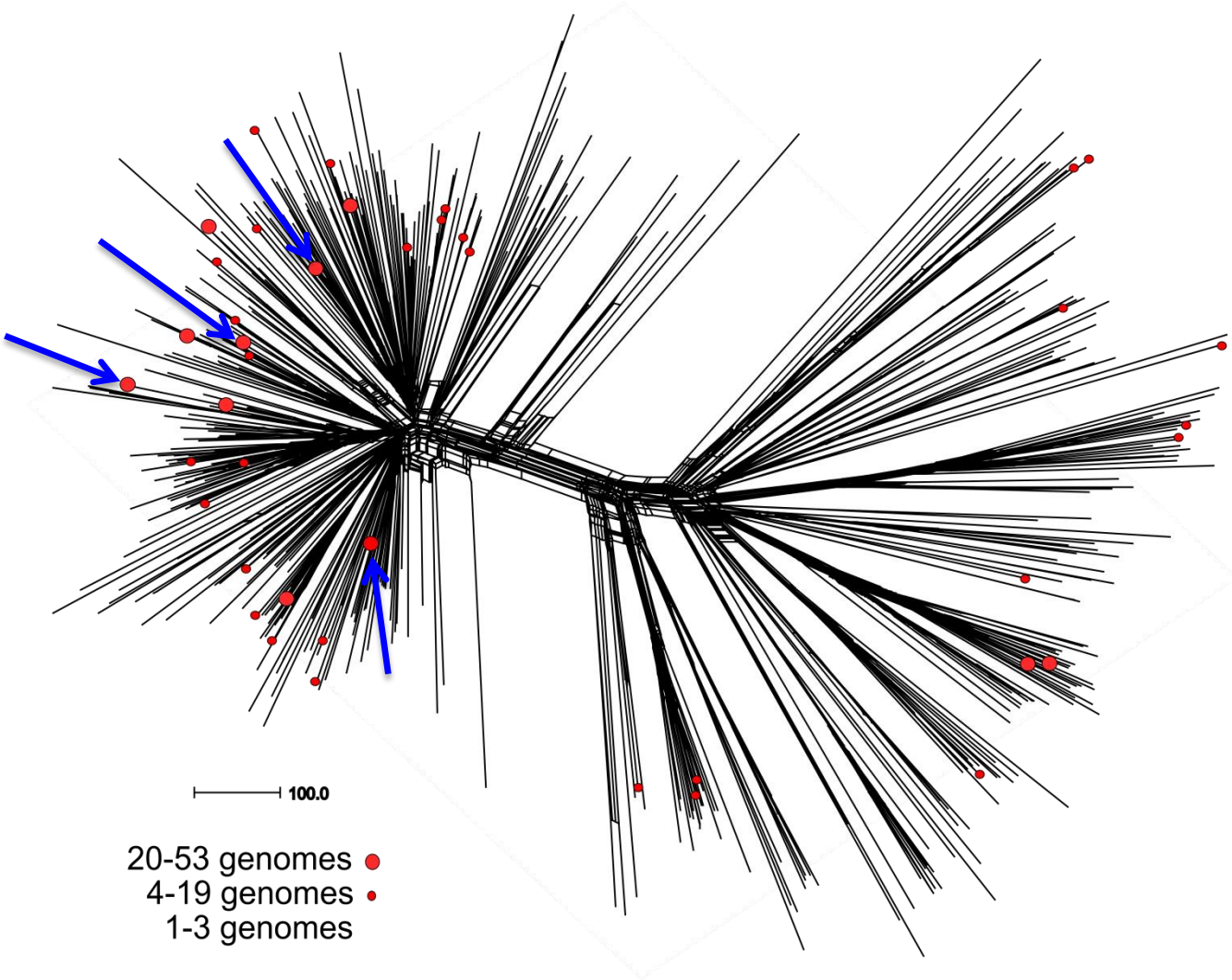
The *rpmE1/rpmE2* and *rpmJ1/rpmJ2* genes were identified and their alleles compared to look for any common allelic profiles or any patterns of occurrence. The 1109 Lineage 3 genomes contained 45 unique combinations of the *rpmE1/rpmE2* and *rpmJ1/rpmJ2* genes. The most commonly observed allele combination 1/1:1/1 was found in 891 (80.4%) of the genomes. The next most common combination 1/2:1/1 was observed 59 (5.3%) times in the dataset. The remaining 43 combinations were divided among the 159 genomes, and 20 of the 43 combinations occurred as a single representative of the allelic combination. In total 14 of the combinations were found in at least four isolates. Carriage and disease association was assessed for these 14 combinations. The most common profile, 1/1:1/1, was present in 11.6% of isolates recovered from instances of carriage while 88.4% were recovered from incidents of disease. Two profiles had higher carriage frequencies. The allelic combination 16/1:1/7 had a 60% carriage association (15/24 isolates) and the second combination, 17/1:3/1, was only found in carriage associated isolates, four in total. There was no clear association between invasive disease and the *rpmE1/rpmE2* or *rpmJ1/rpmJ2* allelic combinations, and an analysis of the isolate provenance did not reveal any temporal or phylogeographic correlation.

Twenty-two of the isolates had an incomplete rMLST profile, and no rST was assigned; these isolate genomes were removed from the rMLST analysis. Overall 579 different rSTs were present; this included the rST profiles from 21 isolates unassigned to a clonal complex. It should also be noted at this point that 17 of the 21 unassigned

isolates had rSTs that were also found in isolates assigned to the 41/44cc, and only four were single examples of a unique rST.

There are practical limitations when rendering graphical networks of the size required for genomic population analysis. The constraints are due to the nonlinear production of data points which is related to the increase in the number of pairwise comparisons made. The rendering of a dataset this large (1109 genomes) was not practical. To alter the data set without losing a substantial amount of variation the number of genomes was reduced by removing three project datasets. All MRF library isolate genomes from the epidemiological year 2013/2014 (137 genomes), all New Zealand genomes recovered from blood culture (91 in total), and 40 disease isolates from a private project were removed. The reduction in the number of strains reduced the data set to 841 genomes. The lower number of genomes, in turn, reduced the computational graphics load but still maintained the observed population structure without losing excessive amounts of data. When the 268 genomes were removed from the dataset, it also removed 149 rMLST profiles, 88 of which were unique; reducing the rSTs in the analysis by 15.2% to 491. The 491 rSTs were compared using the BIGSdb Genome Comparator tool and a distance matrix was created and visualized using a SplitsTree Neighbor-Net (Figure 3). The left half of the network contains 75% of the unique rSTs that are present ten or more times in the database, in particular it associated the four most common rSTs in the data set: rST-2442 (n=53), rST-2435 (n=50), rST-2424 (n=50), and rST-2448 (n=34) in three separate groups with five of the eight next most observed rST profiles in the database, rST-2432 (n=19), rST-2560 (n=14), rST-2476 (n=13), rST-2484 (n=), and rST-2990 (n=11).

Figure 3: Lineage 3 rMLST population structure. The network represents the 491 rSTs of the Lineage 3 data subset. The large red circles represent the most common rST in the data set; in particular rST-2442 (n=53), rST-2435 (n=50), rST-2424 (n=50), and rST-2448 (n=34) which represent 16.9% of the genomes, these are indicated by the blue arrows. The small red circles indicate rSTs shared by 4 to 19 genomes, and the unlabelled branches are rST profiles that are represented only 1 to 3 time in the dataset. The figure was created using the distance matrix output of BIGSdb Genome Comparator, visualized in SplitsTree, and annotated using InkScape.



4.2.3. Lineage 3 BEAST analysis using rMLST

The collection of Lineage 3 meningococcal genome sequences, in particular, those taken at different time points, potentially provide enough evolutionary information to allow the inference of specific temporal and spatial events in the past. The BEAST (Bayesian Evolutionary Analysis Sampling Trees) software is a cross-platform program for the analysis of molecular sequences using the Markov chain Monte Carlo (MCMC) statistical method [48]. The MCMC method is applied in BEAST to multi-dimensional datasets to estimate probability distributions for parameters of interest, while setting prior distributions based on known or hypothesized information, for example dates of sampling or a known rate of evolution. This approach allowed for testing particular evolutionary hypotheses and reconstructing rooted phylogenies.

The evolutionary rate for bacteria varies widely by species and by the type of measurement assay. The temporal frame of the dataset was relatively short (30 years); however, because of the size of the dataset there was potentially enough information to infer patterns. The 53 ribosomal protein gene sequences of Lineage 3 were chosen to investigate the temporal dimension of this group of meningococcal strains and to assess whether or not the lineage evolution was measurable as a unit of time. The gene sequences for the fifty-three ribosomal protein genes corresponding to the 579 unique rMLST profiles, from the 1109 genomes) were extracted and concatenated. The analysis was performed at two levels. The first analysis used all of the nucleotide sites and the second analysis stripped all conserved sites from the sequences leaving only the polymorphic sites to reduce the computational complexity.

The results from the analysis showed very broad confidence intervals on the estimation of the root of the tree that rendered the comparison of reasonable hypothesis statistically impossible. Furthermore, although different strains were separated by long and well-supported branches, sequences within them showed little to no variation. Additionally, maximum likelihood trees for individual genes showed

different groupings among the same samples, providing strong evidence for recombination. For these reasons, this dataset was not appropriate for further testing using this method.

4.3. Antigenic characterization

All 1109 sequenced genomes were typed and grouped using the variable loop regions of the antigens PorA and FetA. The three typing fragments produced 247 allelic combinations, of which 33 were most common and found in five or more of the genomes. The 33 antigen combinations described 73.3% of the Lineage 3 genomes. There were 214 antigenic types observed less than five times each in the data, and of those 157 profiles were observed only once. The most common antigen type was P1.7-2,4:F1-5 (40.4%); the next most observed type P1.18-1,3:F1-5 was only found in 3.2% of the genomes. There were 27 antigen types (10.9%) which were found in isolates recovered from both invasive and non-invasive cases. The carriage association was higher than disease association for three antigen types; 65.2% of profile P1.17-1,23:F1-5 (15 of 23 isolates), 63.6 % of profile P1.17,16-3:F1-20 (7 of 11 isolates), and 60.0% of P1.19,15:F1-5 (10 of 14 isolates).

The *Neisseria* sequence definition database contains the Bexsero ® vaccine antigen scheme, BAST (Brehony - manuscript in preparation). The multicomponent vaccine was designed to provide cross-protection against a broad range of meningococcal serogroup B strains [125]. The vaccine contains four antigenic components: factor H binding protein (fHbp), *Neisseria* adhesion A (NadA), *Neisseria* heparin-binding antigen (NHBA), and outer membrane vesicles providing PorA. Of particular interest to this investigation, the PorA component was prepared from a 41/44cc P1.7-2,4 epidemic strain from New Zealand. The gene sequence encoding the peptide for *nadA* was found in only two of the Lineage 3 genomes, and both were carriage associated isolates. Both loci had full *nadA* coding sequences (NEIS1969), and both were peptide 1. To verify the general absence of *nadA* from the lineage the genomes were also searched using tblastx, a search tool embedded in

the BIGSdb platform. This option compared the six-frame translations of the gene sequence query against all of the six-frame translations of the sequence definition database. The additional searches found no matches for *nadA*. There were 72 unique fHbp peptides found in the dataset, and none of the genomes contained the fHbp peptide variant 1.1 of the vaccine. The most common fHbp peptides were 4 (457 isolates), 14 (288 isolates), and 19 (183 isolates), representing 83.7% of the strains. The most common NHBA peptide found was type 1.2, the vaccine type; it was present in 673 of the genomes (60.7%). The PorA VR1 type P1.7-2 was present in 55.6% (617) of the genomes, and the PorA VR2 type 4 was found in 51.4% (570) of the genomes. The PorA VR1, PorA VR2, and NHBA components P1.7-2,4 and 1.2 were found together 35.3% in the strain collection.

The carriage associated isolates (n=232) had 22.0% coverage with both PorA components; the coverage increased to 27.2% for just PorA VR1 and the coverage was 22.8% for the PorA VR2 alone. There was coverage of 46.1% in the carriage strains for the NHBA component, but only 12.9% coverage for both PorA and NHBA together. Fourteen isolate records did not list a disease association. Six of those genomes contained both the PorA P1.7-2,4 and the NHBA 1.2 components, two contained only the PorA P1.7-2,4, and two only the NHBA 1.2 component. The isolates recovered from incidents of disease had higher coverage overall; 58.5% of the isolates had full PorA coverage, and each PorA component individually had increased coverage: PorA VR1 63% coverage and PorA VR2 58.7% coverage. There was a 63.4% coverage of the isolates recovered from incidents of disease by NHBA and a 40.1% coverage with both PorA and NHBA.

4.4. Mobile genetic elements

4.4.1. The meningococcal disease associated island

The previously identified genetic element the Meningococcal Disease Associated (MDA) island is significantly linked with meningococcal strains causing invasive disease in teenagers [126]. The MDA island includes genes encoding a phage replication initiation

factor, four hypothetical proteins, the T and B cell-stimulating protein B (*tspB*), the zonula occludens toxin (*zot*), and in some strains an integral membrane protein. The phage replication initiation factor (NEIS0031) was found to have two different gene encoding lengths of 1215 and 1209 bases; at least one copy of each sequence variant was identified in each of the 1109 genomes. The *zot* gene (NEIS0023) was found to have two lengths, 1188 and 1103 bases. The shorter version of the *zot* gene is defined in the *Neisseria* sequence definition database using the full-length version, and the shortened sequences were flagged as having an internal stop codon. The majority of the genomes (98.6%) possessed at least one copy of *zot,* and 89.3% of those were a single type, allele 228, the shorter isotype. The remaining six MDA island loci, four hypothetical proteins, and a single transposase were also observed at least once in each Lineage 3 genome.

4.4.2. Plasmids and prophage

The analysis found several uncharacterized phage-related proteins in the 1109 genomes. Online BLAST searches of the loci characterized several as belonging to the Mu-like prophage [127]. The regions identified as the Mu-like prophage were fragmented by genes of unknown function. The composition of the Mu-like prophage included genes encoding Mu-like structural proteins: VpN (NEIS2426), VpP (NEIS2426), gp45 (NEIS2426), gp46 (NEIS2426), gp47 (NEIS2426), gp48 (NEIS2426), and VpH (NEIS2426) and the genes encoding Mu-like regulatory functions: MuA (NEIS1046) and MuB (NEIS1048). Within the lineage, 99% of the isolates contained at least one copy of the Mu-like prophage.

The Lineage 3 genomes also contain plasmid inserts. Two previously undefined insertion regions were found, and each contained one of two type II toxin-antitoxin (TA) systems. The first was characterized by the genes encoding the PemK (toxin) and PemI (antitoxin) protein module [128]. The gene sequences of both *pemK* and *pemI* are

highly conserved across the lineage, *p*-distance 0.004 and 0.001 respectively, and the module was found in over 98% of the genomes. The second insertion contained a TA system encoded within the prophage P1 and encoded the genes for the *doc* gene and the *phd* gene [129]. The *doc/phd* TA system was found in over 98% of the Lineage 3 genomes; and while the gene encoding the *doc* toxin is highly conserved, *p*-distance 0.002, the *phd* antitoxin gene is more variable, *p*-distance of 0.504.

4.4.3. Exploration of proposed virulence factors

To investigate the evolution of virulence traits in the 1109 genomes the literature was searched for proposed and/or identified virulence factors and 88 loci were identified within Lineage 3 isolates (Table 3). The genes were grouped into nine factors based on the pathway or type of gene product. The genomes were compared using the BIGSdb Genome Comparator tool and the compiled list of factors. Forty-two of the genes were present in all genomes with a mean *p*-distance of 0 to 0.14. There were four loci estimated to be completely conserved (*p*-distance 0), ie genes encoding NEIS0053 (*cssB*), NEIS0409 (*pilB*), NEIS1648 (*exbB*), and NEIS1649 (*exbD*), and loci were present in 97.5% to 100% of the genomes. There were ten loci that were not present in either or both of the reference genomes but were present in a few of the Lineage 3 isolates. The genes encoded four proteins which contribute to adhesive properties of meningococci NEIS2155 (*lgtD*), NEIS2011 (*lgtG*), NEIS1969 (*nadA*), one capsule locus NEIS0050 (*cssE*), two iron acquisition NEIS1946 (*hpuA*), NEIS1947 (*hpuB*), one protease NEIS0651 (*iga*), and two toxin NEIS0527 (*frpA/C*), NEIS1344 (*frpA/C*) proteins. Three of the six genes, l*gtG, nadA, and frpA*, were found in a small number, 0.2% to 17.7%, of the isolate genomes. With few exceptions, each of the virulence factors was found in at least 95% of the genomes. The highest *p*-distance estimates (>0.099) were found in 13 loci. These loci belonged to one of six categories: adherence factors *lgtB, lgtE*, *lst, pilC1, pilC2* and *pilJ*; iron uptake factors *hpuA, lbpB* and *tbpB*; immune evasion factors *ctrA*

and *cssA*; invasion factors *opaA* and *porB*; immune modulation factor *fHbp*; and a toxin factor *frpA*.

Table 3. Genes known or hypothesized to be involved in virulence

Virulence factors	Demeric gene name	M01-240149 reference genome locus tag identifier	NZ-05/33 reference genome locus tag identifier	pubMLST.org locus	percent of genomes locus was found in	mean *p*-distance
Adherence						
Adhesion and penetration protein	*App*	NMBM01240149_0201	NMBNZ0533_0338	NEIS1959	99.9	0.034
LOS sialylation	*Lst*	NMBM01240149_1165	NMBNZ0533_0973	NEIS0899	100	0.136
LOS synthesis	*lgtA*	NMBM01240149_0261	NMBNZ0533_0398	NEIS1902	99.4	0.048
	lgtB	NMBM01240149_0262	NMBNZ0533_0399	NEIS1901	99.9	0.108
	lgtC	NMBM01240149_0260	NMBNZ0533_0397	NEIS2154	75.8	0.075
	lgtD	-	-	NEIS2155	99.9	0.049
	lgtE	NMBM01240149_0263	NMBNZ0533_0400	NEIS1900	99.9	0.613
	lgtF	NMBM01240149_0482	NMBNZ0533_1680	NEIS1618	100	0.006
	lgtG	-	-	NEIS2011	4.5	0.016
	lgtH	-	-	n/d	n/d	n/d
	rfaK	NMBM01240149_0481	NMBNZ0533_1681	NEIS1619	100	0.005
	rfaF	NMBM01240149_0642	NMBNZ0533_1504	NEIS1456	100	0.016
	kdtA	NMBM01240149_0016	NMBNZ0533_0016	NEIS2152	100	0.012
	rfaC	NMBM01240149_2054	NMBNZ0533_2082	NEIS2134	100	0.003
***Neisseria* adhesion A**	*nadA*	-	-	NEIS1969	0.2	0
Phosphoethanolamine modification	*lptA*	NMBM01240149_0550	NMBNZ0533_1612	NEIS1553	100	0.002
Type IV pili	*pilC1/pilC2*	NMBM01240149_0043, NMBM01240149_0338	NMBNZ0533_0050, NMBNZ0533_0475	NEIS0371 NEIS0033	86.6 87	0.417 0.365
	pilE	NMBM01240149_0020	NMBNZ0533_0019	n/d	n/d	n/d
	pilS	NMBM01240149_0021, NMBM01240149_0022, NMBM01240149_0024	NMBNZ0533_0021, NMBNZ0533_0022, NMBNZ0533_0024, NMBNZ0533_0025, NMBNZ0533_0026, NMBNZ0533_0029	n/d	n/d	n/d
	pilD	NMBM01240149_1752	NMBNZ0533_1913	NEIS1839	100	0.006
	pilF	NMBM01240149_1755	NMBNZ0533_1916	NEIS1844	100	0.053
	pilG	NMBM01240149_1751	NMBNZ0533_1912	NEIS1838	100	0.001
	pilX	NMBM01240149_1198	NMBNZ0533_0940	NEIS0831	99.8	0.067
	pilW	NMBM01240149_0844	NMBNZ0533_1298	NEIS1246	100	0.001
	pilM	NMBM01240149_0381	NMBNZ0533_0519	NEIS0412	100	0.002
	pilN	NMBM01240149_0380	NMBNZ0533_0518	NEIS0411	100	0.002

Virulence factors	Demeric gene name	M01-240149 reference genome locus tag identifier	NZ-05/33 reference genome locus tag identifier	pubMLST.org locus	percent of genomes locus was found in	mean *p*-distance
	pilO	NMBM01240149_0379	NMBNZ0533_0517	NEIS0410	99.9	0.001
	pilP	NMBM01240149_0378	NMBNZ0533_0516	NEIS0409	99.9	0
	pilQ	NMBM01240149_0377	NMBNZ0533_0515	NEIS0408	100	0.058
	pilT1	NMBM01240149_0046	NMBNZ0533_0053	NEIS0036	100	0.003
	pilT2	NMBM01240149_1320	NMBNZ0533_0819	NEIS0721	100	0.001
	pilU	NMBM01240149_0045	NMBNZ0533_0052	NEIS0035	100	0.004
	pilH	NMBM01240149_1202	NMBNZ0533_0936	NEIS0827	99.8	0.054
	pilI	NMBM01240149_1201	NMBNZ0533_0937	NEIS0828	99.8	0.06
	pilJ	NMBM01240149_1200	NMBNZ0533_0938	NEIS0829	99.8	0.215
	pilK	NMBM01240149_1199	NMBNZ0533_0939	NEIS0830	99.8	0.053
	pilV	NMBM01240149_1556	NMBNZ0533_0585	NEIS0487	100	0.051
	pilZ	NMBM01240149_1318	NMBNZ0533_0821	NEIS0723	100	0.01
Efflux pump						
FarAB	*farA*	NMBM01240149_1765	NMBNZ0533_1926	NEIS1852	100	0.009
	farB	NMBM01240149_1764	NMBNZ0533_1925	NEIS1853	99.9	0.044
MtrCDE	*mtrC*	NMBM01240149_0471	NMBNZ0533_1691	NEIS1634	100	0.006
	mtrD	NMBM01240149_0472	NMBNZ0533_1690	NEIS1633	100	0.002
	mtrE	NMBM01240149_0473	NMBNZ0533_1689	NEIS1632	100	0.002
Immune evasion						
Capsule	*ctrA*	NMBM01240149_0070	NMBNZ0533_0077	NEIS0055	97.9	0.517
	ctrB	NMBM01240149_0071	NMBNZ0533_0078	NEIS0056	98	0.004
	ctrC	NMBM01240149_0072	NMBNZ0533_0079	NEIS0057	98	0.015
	ctrD	NMBM01240149_0073	NMBNZ0533_0080	NEIS0058	98	0.008
	ctrE	NMBM01240149_0081	NMBNZ0533_0087	NEIS0066	98.6	0.022
	ctrF	NMBM01240149_0082	NMBNZ0533_0088	NEIS0067	98.6	0.002
	cssA	NMBM01240149_0069	NMBNZ0533_0076	NEIS0054	97.5	0.116
	cssB	NMBM01240149_0068	NMBNZ0533_0075	NEIS0053	97.5	0
	cssC	NMBM01240149_0067	NMBNZ0533_0074	NEIS0052	97.5	0.004
	csc	NMBM01240149_0066	NMBNZ0533_0073	NEIS0051	3	0.002
	cssE	-	-	NEIS0050	3	0
Immune modulator						
Factor H binding protein	*fHbp*	NMBM01240149_0319	NMBNZ0533_0457	NEIS0349	100	0.123
***Neisseria* surface protein A**	*nspA*	NMBM01240149_1429	NMBNZ0533_0710	NEIS0612	100	0.009
Invasion						
Class V OMPs	*opcA*	NMBM01240149_1061	NMBNZ0533_1078	NEIS2198	95.6	0.003
Opacity protein	*opaB,C,D*	NMBM01240149_0552, NMBM01240149_0700, NMBM01240149_1161, NMBM01240149_1646	NMBNZ0533_0977, NMBNZ0533_1444, NMBNZ0533_1610, NMBNZ0533_1807	NEIS1403 NEIS1551 NEIS0903	90.2 0 0	0.717 0 0
PorA	*porA*	NMBM01240149_0736	NMBNZ0533_1408	NEIS1364	99.9	0.12
PorB	*porB*	NMBM01240149_0143	NMBNZ0533_1971	NEIS2020	100	0.101

Virulence factors	Demeric gene name	M01-240149 reference genome locus tag identifier	NZ-05/33 reference genome locus tag identifier	pubMLST.org locus	percent of genomes locus was found in	mean *p*-distance
Iron uptake systems						
ABC transporter	*fbpA*	NMBM01240149_1462	NMBNZ0533_0677	NEIS0578	100	0.002
	fbpB	NMBM01240149_1463	NMBNZ0533_0676	NEIS0577	100	0.001
	fbpC	NMBM01240149_1464	NMBNZ0533_0675	NEIS0576	100	0.002
Ferric enterobactin transport protein A / ferric-repressed protein B	*fetA/frpB*	NMBM01240149_0198	NMBNZ0533_0334	NEIS1963	99.5	0.063
Haemoglobin receptor	*hmbR*	NMBM01240149_0516	NMBNZ0533_1646	hmbR	99.3	0.092
Heme uptake	*hpuA*	-	-	NEIS1946	17.7	0.137
	hpuB	-	-	NEIS1947	17.7	0.025
Lactoferrin-binding protein	*lbpA*	NMBM01240149_0629	NMBNZ0533_1525	NEIS1468	98.6	0.019
	lbpB	NMBM01240149_0628	NMBNZ0533_1526	NEIS1469	98.7	0.379
Ton system	*tonB*	NMBM01240149_0458	NMBNZ0533_1704	NEIS1650	100	0.081
	exbB	NMBM01240149_0459	NMBNZ0533_1703	NEIS1649	100	0
	exbD	NMBM01240149_0460	NMBNZ0533_1702	NEIS1648	100	0
Transferrin-binding protein	*tbpA*	NMBM01240149_1629	NMBNZ0533_1790	NEIS1690	100	0.042
	tbpB	NMBM01240149_1630	NMBNZ0533_1791	NEIS1691	100	0.974
Protease						
IgA protease	*iga*	NMBM01240149_1392	-	NEIS0651	100	0.035
Stress protein						
Catalase	*katA*	NMBM01240149_1867	NMBNZ0533_0222	NEIS0211	100	0.004
Manganese transport system	*mntA*	NMBM01240149_1507	NMBNZ0533_0632	NEIS0530	100	0.026
	mntB	NMBM01240149_1508	NMBNZ0533_0631	NEIS0529	100	0.013
	mntC	NMBM01240149_1509	NMBNZ0533_0630	NEIS0528	100	0.06
Methionine sulphoxide reductase	*msrA/B (pilB)*	NMBM01240149_0041	NMBNZ0533_0048	NEIS0020	100	0.005
Recombinational repair protein	*recN*	NMBM01240149_1348	NMBNZ0533_0791	NEIS0694	100	0.007
Toxin						
Neisseria ADP-ribosylating enzyme	*narE*	NMBM01240149_0812	NMBNZ0533_1330	NEIS2492	96.2	0
RTX toxin	*frpA*	NMBM01240149_0750	-	NEIS0527	1.1	0.271
	frpC	-	-	NEIS1344	86.6	0.663

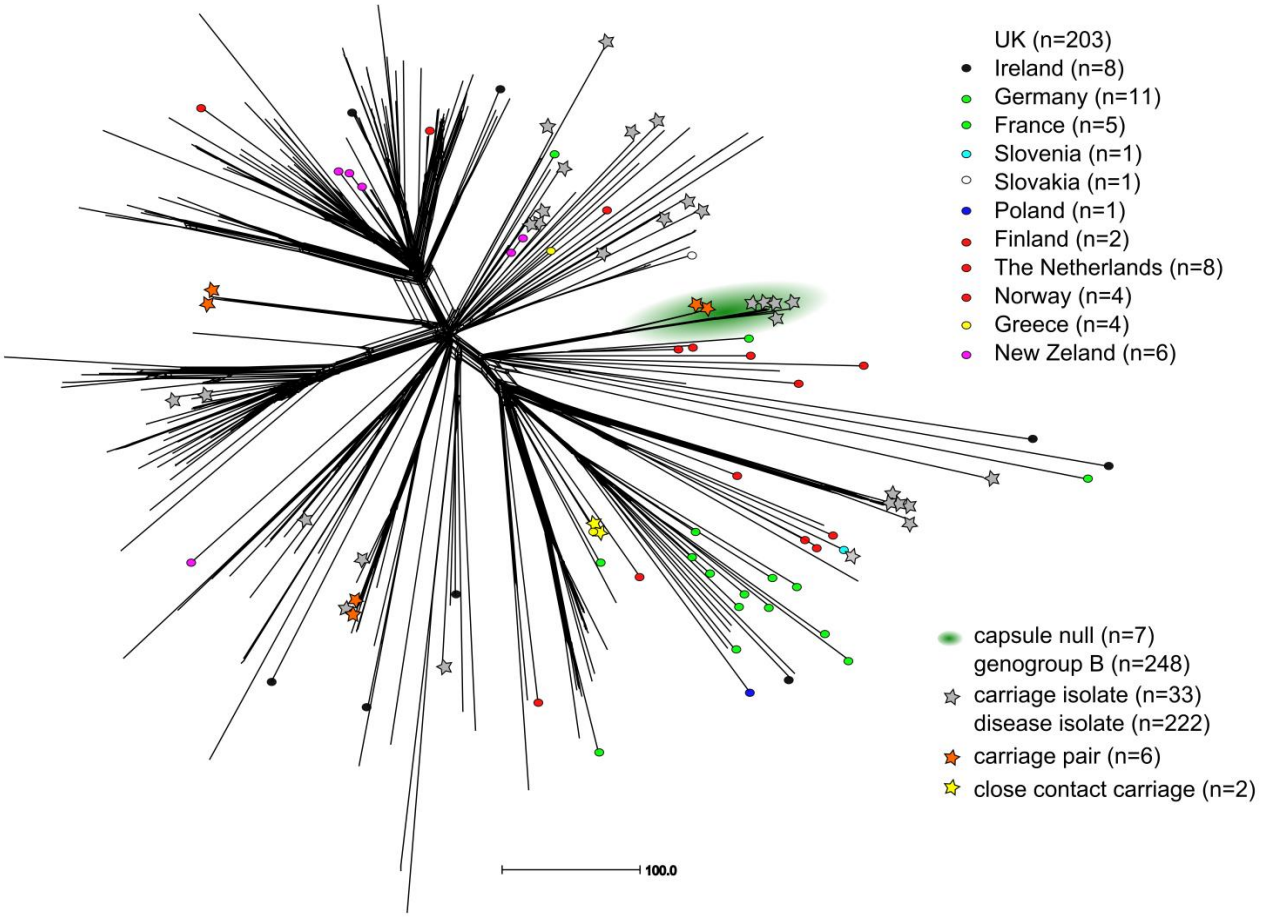
n/d – The unidirectional intergenomic recombination of the *pilE*/*pilS* exchanges generates high levels of genetic and antigenic variation. The locus sequence definition is unstable in the sequence definition database, due to assembly uncertainty, therefore comparison of the locus was not done.

5. Single sequence type characterization, ST-41

The *Neisseria* PubMLST.org database was queried in August 2015 for all 41/44cc isolates typed as ST-41, and that also had genomic data, 255 records were identified. The ST-41 isolates were recorded in the database as 38 carriage isolates, 209 disease isolates and 8 with their disease status undefined. The isolates were collected between 1986 and 2014 from 12 countries. Thirty-five of the carriage isolates were collected as part of a UK carriage study between 1999-2001. Among the strains, there were three sets of temporally paired isolates collected from three single carriers approximately one year apart; and a Greek disease isolate with two associated isolates from close contacts of the patient. A genome comparison was run using the L3 cgMLST scheme and the 255 ST-41 isolates (Figure 4).

The isolates were genotyped using the capsule region A genes and all but seven contained genes encoding the serogroup B phenotype. The seven isolates without capsule region A were confirmed to be *cnl* and all grouped together in the neighbour-net. Two *cnl* strains were isolated during the UK carriage study in 1999/2000 and were a single carrier pair; five isolates were from an independent carriage study during the epidemiological year 2001/2002, also conducted in the UK. The *cnl* genomes are highly conserved; 1455 loci had a *p*-distance estimate of 0, 211 loci had estimates >0 to 0.02 and 43 loci had estimates of >0.02 to 0.03. The locus with the highest sequence variation was the gene encoding the *Neisseria* heparin binding antigen, *nhba* (NEIS2109). In addition to missing the region A capsule genes, all seven genomes were missing NEIS0068, an uncharacterized protein with sequence similarity to the repeat protein signal transduction mechanisms of the *sel1* gene. The seven *cnl* strains also contained 11 genes identified as belonging to the gonnococcal genetic island

Figure 4. L3 cgMLST comparison of ST-41 genomes (n=255) Core genome analysis of the ST-41 isolates found 1574 (90.4%) of the loci in all 255 isolates and included 14 loci with identical gene sequences. 1458 loci had a *p*-distance estimate ≤ 0.04, and 1329 of those were above 0 but ≤0.02; therefore 83.8% of the ST-41 strain genome is highly conserved. The core genome was composed of 27.7% hypothetical proteins and 34 genes known to be phage and/or plasmid-related, nine of which are present in all genomes.



.

Despite being scattered throughout the population network, the carriage genomes were found clustered together in groups or alone within a larger group of isolates collected from disease cases. There were two distinct clusters without carriage associated recovery. Genome comparisons between the seven *cnl* strains and five identifiable groupings within the network

identified a shared gene complement ranging from 409 to 910 of the core loci. The comparisons revealed an average of 9.3% of the allelic variation was due to indels in 88 to 100 loci, and the remaining allelic differences were due to sequence variation.

The core genome analysis of the ST-41 isolates found 1574 (90.4%) of the loci in all 255 isolates and included 14 loci with identical gene sequences: six ribosomal protein genes NEIS0137 (*rplV*), NEIS0141 (*rpsQ*), NEIS0142 (*rplN*), NEIS0143 (*rplX*), NEIS0154 (*rpmJ*), and NEIS0156 (*rpsK*), five hypothetical proteins NEIS0952, NEIS1274, NEIS1663, NEIS2366, and NEIS2384, the two-component ribosomal biogenesis proteins NEIS0153 (*infA*) and NEIS1239 (*infB*), and the metabolism enzyme NEIS0767 (*adk*). The majority of the loci, 1458, had a *p*-distance estimate ≤ 0.04, and 1329 of those were above 0 but ≤0.02; therefore 83.8% of the ST-41 strain genome is highly conserved. The core genome was composed of 27.7% hypothetical proteins and 34 genes known to be phage and/or plasmid-related, nine of wich are present in all 255 genomes. The highest variation was found in the genes encoding NEIS1691 the transferrin binding protein (tbpB) and NEIS2109 (nhba), with *p*-distance estimates of 0.93 and 0.822 respectively. There were 115 rST profiles among the genomes and 45 antigen types.

The strains all contained annotated phage/plasmid genes identified as belonging to the MDA island, the Mu-like prophage, and a yet as undefined phage or plasmid insert. There could be several more phage or plasmid-related loci, however. There are loci integrated into the regions annotated as phage/plasmid-like that are annotated only as hypothetical loci and online searches provided no further information. With rare exception each of the MDA and Mu-like prophage genes, as described in section 4.4.2., were identified in as present in 96.9% and 96.5% of the ST-41 strains.

6. Discussion

Few whole-genome population level comparisons of the scale and variation investigated here exist for bacteria of a single species [130-133]. Meningococcal Lineage 3 is characterized by the epidemiological diversity observed historically and globally for the meningococcal species. This makes the Lineage ideal for population genetic studies. The Lineage is phenotypically diverse and has one of the highest levels of prevalence for a serogroup B strain in most geographical regions [134-136]. The Lineage is known to be hyperinvasive, resulting in low-level outbreaks to epidemics [105, 111, 136]. The high and low invasive potential of the Lineage has shifted over time, and in different global regions different STs have been associated with invasive disease [108, 136]. A proportion of this diversity has been captured in the *Neisseria* PubMLST.org database and is typed as the ST-41/44 clonal complex. In August 2015, a survey of the database found 17.1% (1933) of defined STs belonged to the 41/44cc; which is more than any other clonal complex in the database and three times as many STs as the next largest clonal complex ST-32 with 621 STs. This amount of variation is due, in part, to how an ST is defined for the 41/44cc. Instead of a single central ST the complex is defined by two STs the ST-41 and ST-44; and based on MLST, the two STs are genetically two-locus variants of each other. The designation of two central STs is rare for a meningococcal clonal complex and was chosen based on the dual epidemiological nature. While seven loci are used for the typing, there are nine possible allele matches, allowing the 41/44cc a broader inclusion criteria.

Only a small proportion of the 41/44cc STs in the database, 272 (14%), were represented by one or more whole genome sequences. However, the comparison of a single sequence type (ST-41) showed that the designation of a 'single type' masks a large amount of genetic and phenotypic variation (Figure 4). This was consistent with previous studies observing the limited power of MLST to detect evolutionary relationships or provide a high level of resolution required for some applications [137, 138]. The ability of MLST to robustly categorize related strains into a group, which can be further explored with population biology and evolutionary methods if whole genome sequences are available, is, however, evident. The results presented here emphasize that referring

to a group of genomically defined strains that share the same ST as 'the same' is somewhat misleading. The high level of resolution provided by whole genome sequence analysis underscores the level of diversity within a bacterial lineage.

The analysis found that disease-associated and asymptomatic carried genomes did not form distinct phylogenetic groups and supports the findings of Schoen *et al*. who performed the first whole genome comparison of carried and disease associated meningococcal isolates [139]. When rSTs were mapped to the Neighbour-Net, with two exceptions, those rSTs with a higher frequency in the dataset grouped in one of two distinct sublineages (Figure 3). This finding indicates two groups of related strains each with a more recent common ancestry, conceivably based on a division based on higher fitness potential.

High rates of gene flow combined with strong selection and abundant, diverse DNA repeat sequences can result in the formation of mosaic genomes [98, 140]. PROKKA identified new hypothetical proteins in the Lineage 3 genomes that were conserved throughout the lineage and were associated with regions where mobile genetic elements were found. In addition to confirming the high levels of recombination, the BEAST analysis also found evidence of clonal expansion. The implication of these two findings being that the phage and plasmid inserts were incorporated into the genome at the point, or just prior to, the proliferation of the lineage; and were major forces shaping the population structure. These results were also congruent with the observed deep branching phylogeny population structure that distinctly separates Lineage 3 from the other meningococcal lineages; in fact many meningococcal lineages share this feature [96]. Unfortunately dating these expansions is difficult, as there was not enough evolutionary signal present. This is a common issue when BEAST is applied to bacterial genomes [141]. However, it may also indicate that not enough time for substitutions had occurred that would allow for an estimate of temporal history; indicating that the addition of more, older samples may be beneficial.

The concept of a species pan-genome was first described in 2005 as a collection of all genetic material within the same species composed of a core genome, an accessory genome and unique or strain specific loci [58]. Previous comparative genome studies failed to find any obvious genomic

differences between the disease-associated and carried genomes based on the presence or absence of any particular genes. The Lineage 3 core genome contained 1740 loci and included genes involved in core metabolic pathways, specifically genes associated with metabolism and biosynthesis, all of which are essential to the lifecycle and fitness of the meningococcus. There were no identified genes in the accessory genome that were a part of these critical pathways. Comparison of the rMLST loci clustered isolates in two distinct sublineages (Figure 3). This was consistent with persistent genome-wide allelic structure for a meningococcal lineage and indicates a conservation of core genes [35]. The virulence factors identified from the literature were present in >86% of the lineage 3 genomes, and all but thirteen genes are part of the Lineage 3 cgMLST and the distance estimates indicated most were conserved, *p*-distance <0.080. This was in agreement with previous findings that demonstrated no virulence gene pool exists exclusively in the hyperinvasive meningococcal strains [98, 139]. Some virulence factors are known to be phase variable indicating a potential for phenotypic modification mediated by gene switching [142]. The repetitive nature of these genes causes problems with short-read assembly methods and in some genomes an absent virulence associated locus may be present but did not assemble. In time, sequencing technologies and bioinformatics should improve, and incomplete genome assemblies decrease.

The persistence of the hyperinvasive 41/44cc strain types containing PorA VR1, PorA VR2:FetA association have been described predominantly as antigenic type P1.7-2,4:F1-5 as was found in this collection of isolates [143]. The diversity of antigenic types in this collection of over 1000 isolates was high, 406 types; but not congruent with the phylogenetic analysis. Data presented here found up to 60% of the 41/44cc isolates would have been covered by one or more of the Bexsero® vaccine components, a recently approved meningococcal serogroup B vaccine being used in the EU, North America and Australia [144].

Two insertion regions were identified in the lineage genomes, and each contained one of two type II toxin-antitoxin (TA) systems [145]. The first region contained the gene encoding *pemK* that has been shown to be a ribonuclease that inhibits replication by affecting DNA synthesis at the initiation phase [146]. The overexpression of *pemK* is toxic to the host cell and inhibits protein

synthesis without affecting DNA replication [128]. This TA system was found in 99% of the genomes. The second TA system (*phd/doc*) was a pair of genes known to contribute to plasmid stabilization by the introduction of a lethal response to plasmid loss [129]. The target of the system is not clear; however indirect evidence indicates that it inhibits translation [147]. The prevent-host-death (*phd*) antitoxin gene was highly variable and had a *p*-distance of 0.504 while the death-on-curing (*doc*) gene had a *p*-distance of 0.001 and, was, therefore, highly conserved. The system is part of a post-segregational killing system that is translationally coupled, and the variation suggests that the anti-toxin may actively block a host of different type II toxins. A comparative study of the Lineage 3 TA systems and comparison to other lineages could identify another clade specific mechanism, much like the restriction-modification systems found by Budroni *et al*. [44]. This finding is interesting because the Lineage 3 genomes contain at least two identified TA systems and possibly more.

The Lineage 3 genomes contained one or more regions with high sequence similarity to the phage Mu, which was interspersed with segments that are seemingly unrelated [127]. Further inquiry using online BLAST searches of the loci characterized several as belonging to the Mu-like prophage, which was previously identified in the meningococcus serogroup B isolate MC58 [127].

Taken together these findings provide a gene-be-gene account of the genomic data that structures the Lineage 3 population. The identification of the core and accessory genes using the gene-by-gene approach enabled the identification of elements forming the population structure of the Lineage and the biological significance associated with allelic variation. Analysis of a defined population of isolates that are known to have diverse phenotypes provides an invaluable tool for exploring the evolution of lineages within a species. Also, the identification of multiple phage and plasmids contributed substantially to the genetic diversity and genome plasticity and may increase the fitness of this Lineage. The identification of multiple elements, such as those importing TA systems and genes of unknown function, would offer new additional prospects in understanding the diversification of this and other lineages over time.

Chapter 5 Genomic Similarity of *Neisseria meningitidis* carriage isolates of the same lineage found in individual carriers

1. Introduction

Neisseria meningitidis is an obligate commensal bacterium that asymptomatically colonizes the human upper respiratory mucosa, rarely causing disease. Recognized as an epidemiological state by 1890, the phenomenon known as carriage has been shown to have an immunizing effect in both children and adults and thus may be important for the development of natural immunity [148, 149]. Many studies from the from the last 40 years found that 10% of a healthy population could be carriers of the meningococcus in the nasopharynx in non-epidemic settings; whereas overall incidence of meningococcal disease is around 1-3/100,000 in most of Europe and North America [94, 150-152]. Not only is asymptomatic carriage more prevalent than invasive disease, but the patterns of carriage also differ. It has been estimated that a typical individual will asymptomatically carry meningococci up to ten times in the first 30 years of life and at least 10% of the general population will harbour the organism during periods of endemic disease [153, 154]. Carriage prevalence also varies with age, with the lowest rates found in young children and a peak in teenagers and young adults that can be as high as 30% [91, 153]. Asymptomatic carriers are thought to be the transmission source of invasive meningococcal disease to susceptible hosts because most patients have not been in contact with other cases. Transmissibility and long-term colonization, rather than invasive potential, are considered essential attributes in the meningococcal lifecycle, and the state of carriage is suspected to be a factor affecting disease incidence.

Many questions remain in understanding the meningococcal non-invasive carriage state and its population structure. Previously serological and molecular typing methods have shown that carriage populations are genotypically more variable than collections of disease isolates [91, 155-157]; however, how similar persistently carried strains are at a genomic level or what role reinfection or co-colonization play in long-term carriage is unknown, as are the factors that determine the actual length of each carriage episode. Correspondingly, an explanation for why a

normally commensal pharyngeal bacterium occasionally crosses the epithelial and or blood/brain barrier to cause sepsis and invasive meningococcal disease also, to a certain degree, elusive. *N. meningitidis* is known to evade complement-mediated killing, however a study by Davila *et al* in 2010 also suggests that host genetic variation in regulators of complement activation contribute to the occurrence of invasive disease versus asymptomatic colonization [158, 159]. And so, is the propensity to cause disease a stochastic host-bacteria interaction - in other words is it chance alone, *i.e.* host susceptibility, or adaptive within-host evolution of the strain? Further molecular and genomic studies on episodes of meningococcal carriage are needed to provide information on the relationship between non-invasive strains, hyperinvasive strains, and the mechanisms generating the genetic diversity of this organism. This chapter will begin with a background section on the large dataset from which a subsequent smaller collection was taken for further analysis. The large study provided a baseline comparison of UK carriage, as observed from 1999-2001, for this analysis. In the subsequent sections, the focus is on a subset of isolates from a single region, which were used to investigate the genomic level of similarity between meningococcal strains of the same ST recovered from individual carriers approximately one year apart. The aim of this study was to investigate within-host evolution using subsets of *N. meningitidis* isolates with the same MLST profile, that were recovered from individual carriers and thought to be instances of a long-term asymptomatic carriage state in an attempt to identify the level of genetic similarity between the related isolate pairs and the mechanisms of genetic diversity present.

2. Background of the study

From 1999 to 2001 a large cross-sectional survey of carried meningococci in young adults aged 15-19 was conducted in seven regions of England, Scotland and Wales (UKMenCar) [160]. The coordinating centres were located in Glasgow, Stockport, Nottingham, Oxford, Bangor, Cardiff, London and Plymouth; with samples taken from across the area of each centre. The study was undertaken during a period of increased disease across the United Kingdom [161]. The aim of

the UK-wide study was to determine the effect of the meningococcal serogroup C conjugate (MCC) vaccine on meningococcal populations; specifically carriage of serogroup C meningococci to determine the effects of the vaccine on herd immunity [162]. Oropharyngeal swabs were taken annually over a period of three years, and a questionnaire was completed for demographic purposes. The isolate collection also offered a unique chance to investigate within-host evolution and potential diversity of longitudinal carriage in a high transmission environment during a period of high disease incidence using a regional subset from the study. Recovered isolates were typed using Gram's stain, oxidase testing, serogroup agglutination and MLST [26].

2.1. Baseline comparison of regional carriage data in the United Kingdom

To contextualize and understand the possible impact of the regional study results, a baseline comparison was needed. The following section is an overview of the data from all regional centres in the UK study. The sample database was queried (August 2013) for all records where recovered isolates were identified as *N. meningitidis* within each region.

Table 1: Regional sampling (swabs) and frequency of *N. meningitidis (Nm).* Isolates are grouped by year of recovery and compared by centre. Rates of recovery varied by centre and increased over the three year period. The rate of carriage increased 2% over the three years with an overall prevalence of 17.7%.

	1999			2000			2001			overall total		
Centre	**swabs**	***Nm***	**%**	**swabs**	***Nm***	**%**	**swabs**	***Nm***	**%**	**swabs**	**Nm**	**%**
Bangor	972	183	*18.8*	1087	198	*18.2*	1025	224	*21.9*	3084	605	*19.6*
Cardiff	1718	236	*13.7*	1723	263	*15.3*	1655	257	*15.5*	5096	756	*14.8*
Glasgow	2896	300	*10.4*	2995	461	*15.4*	2996	495	*16.5*	8887	1256	*14.1*
Nottingham	1669	289	*17.3*	2342	439	*18.7*	2586	391	*15.1*	6597	1119	*17.0*
Oxford	2398	524	*21.9*	2147	410	*19.1*	2471	460	*18.6*	7016	1394	*19.9*
Plymouth	1393	107	*7.7**	2697	395	*14.6*	3475	686	*19.7*	7565	1188	*15.7*
Stockport	3011	696	*23.1*	3491	748	*21.4*	3562	798	*22.4*	10064	2242	*22.3*
total	**14057**	**2335**	**16.6**	**16482**	**2914**	**17.7**	**17770**	**3311**	**18.6**	**48309**	**8560**	**17.7**

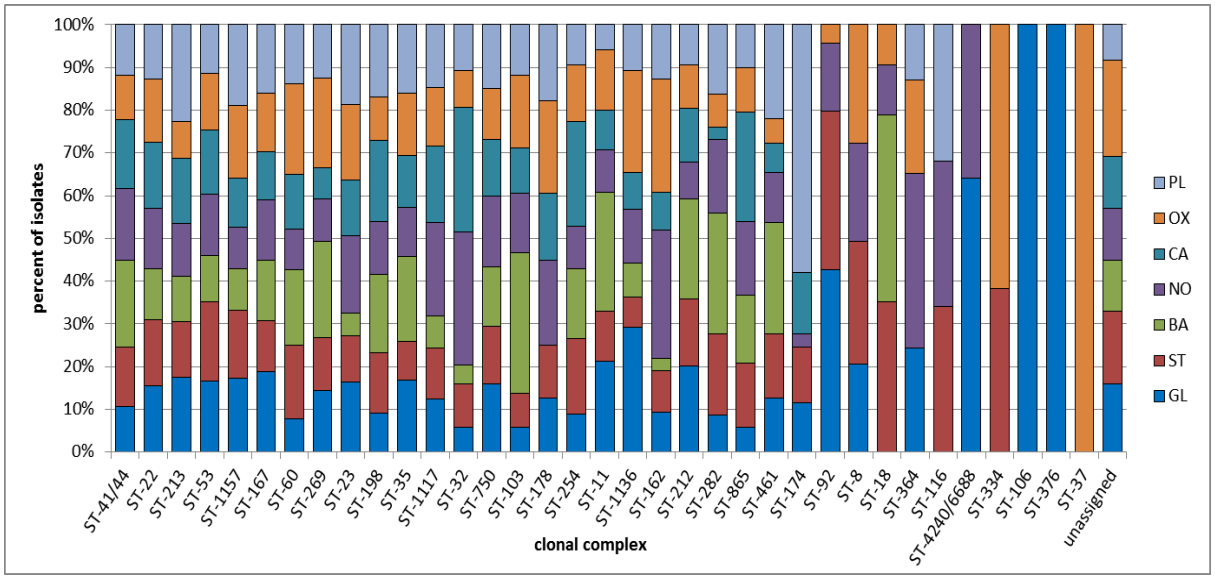
* The initial low recovery rate in 1999 for Plymouth was caused by the plating method used. Their method was subsequently changed in 2000, and the recovery rates were more in line with other regions. ***NOTE:*** The London site lost their samples due to a freezer failure and are not included.

A total of 48,309 swabs were taken over the three-year study period, and 8560 (17.7%) were identified as *N. meningitidis* isolates (Table 1). The overall carriage rate among centres increased slightly from 16.6% to 18.6% over the three years and rates of carriage varied by individual centres. The clonal complex and ST were determined for each isolate, and the prevalence of both was compared between the centres (Figure 1A). Over 93% of the isolates (7985) could be grouped into one of 35 defined clonal complexes, which contained 1312 (81.4%) of the STs found in the study. The remaining 300 STs (18.6%) were not assigned to a clonal complex and represented 575 isolates (6.7%). There were 23 clonal complexes (65.7%) present in all seven regions.

The most prevalent clonal complexes were ST-41/44, ST-22, with 1182 and 1118 isolates respectively, and complexes ST-213 and ST-53 with 816 and 805 isolates (Figure 1B). Together these four complexes represented over 45% of the total dataset. Three clonal complexes had only one isolate, complexes ST-106, ST-376 and ST-37, and three clonal complexes were represented by two, three and four isolates, ST-334, ST-4240/6688 and ST-116 respectively. The remaining complexes are represented by at least eight isolates each. Due to the design of the study, some participants were sampled in successive years. The database was queried a second time (August 2013) for individual carriers with an isolate record and a positive meningococcal culture in two successive years. This was to determine if these isolates represented two separate carriage episodes from unrelated meningococci, or whether they were related and represented long-term carriage.

Figure 1. Typing of UK meningococcal carriage study isolates. (A) Comparison of all carried meningococcal clonal complexes by region, n=8560; and (B) regional distribution of carried meningococcal clonal complexes. Clonal complexes are arranged in descending order of prevalence. Clonal complexes ST-37, ST-376 and ST-106, have one isolate each, ST-334 represents two isolates, ST4240/668 and ST-116 contain three and four isolates respectively. There were 575 STs unassigned to a clonal complex. Coordinating study centre: (PL) Plymouth n=1188, (OX) Oxford n=1394, (CA) Cardiff n=756, Wales n=, (NO) Nottingham n=1119, (BA) Bangor, Wales n=605, (ST) Stockport n=2242, (GL) Glasgow, Scotland n=1256.

A.



B.

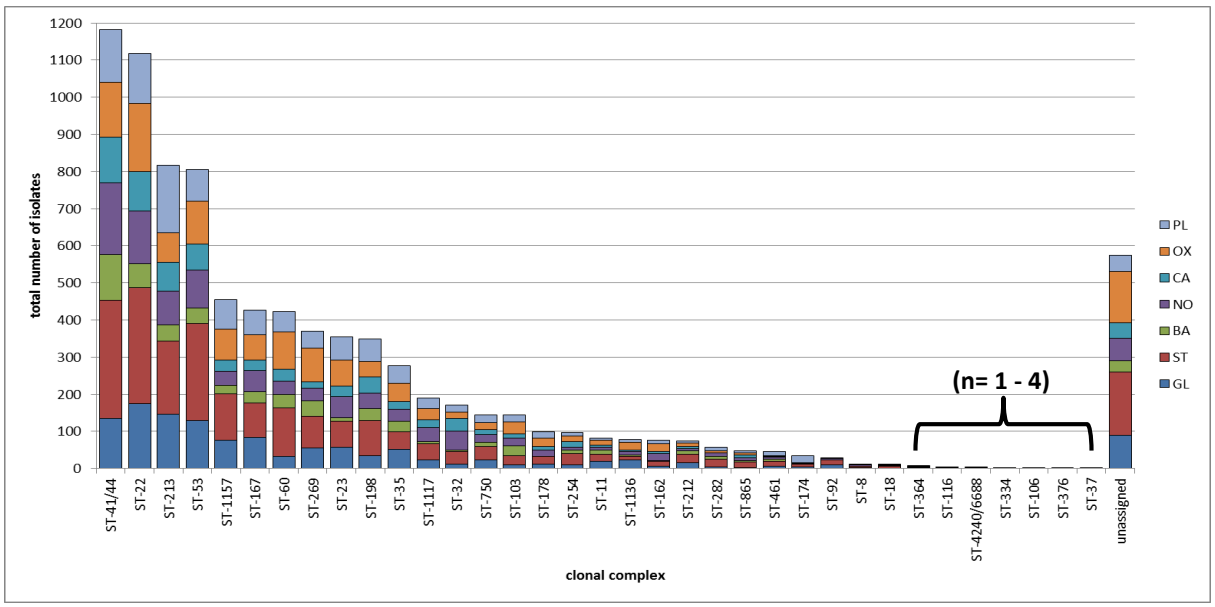


Table 2. Meningococcal isolate pairs by year of recovery. Overall there was a total of 924 meningococci successively recovered from individual carriers on more than one occasion. The number of pairs per consecutive year and the percentage for each centre by pair year based on the total number of pairs. There was an increase in the number of long-term individual carriers between the 1999/2000 and 2000/2001 epidemiological years.

Centre	number of pairs								Total	
	1999/2000		2000/2001		1999/2001		1999-2001*			
Bangor	14	3.0%	21	4.5%	2	0.4%	0	0.0%	37	8.0%
Cardiff	12	2.6%	15	3.2%	0	0.0%	0	0.0%	27	5.8%
Glasgow	33	7.1%	38	8.2%	0	0.0%	0	0.0%	71	15.3%
Nottingham	15	3.2%	20	4.3%	2	0.4%	0	0.0%	37	8.0%
Oxford	36	7.8%	42	9.1%	3	0.6%	2	0.4%	83	17.9%
Plymouth	7	1.5%	46	10.0%	0	0.0%	0	0.0%	53	11.5%
Stockport	71	15.4%	80	17.3%	2	0.4%	2	0.4%	155	33.5%
Total	188	40.7%	262	56.7%	9	1.9%	4	0.9%	463	100%

* The two individuals that tested positive all three years, 1999-2001, each contribute two pairs, a 1999/2000 pair and a 2000/2001 pair.

Overall, 188 participants with a *N. meningitidis* positive swab in 1999 were also carrying a meningococcus in 2000; 262 participants in 2000 were carrying again in 2001; nine individuals participated in all three years but only had positive cultures in 1999 and 2001; and two people who also participated all three years were carrying meningococci in all three years (Table 2). In total 924 temporally associated isolates were found. For each individual carrying the meningococcus for more than one year, the two recovered isolates will be referred to as a pair. There were 463 pairs found across the UK study regions. The pairs were grouped by centre and categorized by clonal complex and the number of MLST allelic differences between isolates of each pair, termed a locus variant, to determine the typing similarity of each pair. Briefly, the ST is a genetic profile defined using the DNA sequence of seven housekeeping gene fragments of approximatly 400 to 500 base pairs each. The MLST profiles are grouped into clonal complexes by their similarity to a central or most commonly found genetic profile (Table 3). Association with a central ST is defined as including any MLST profile that matches the central ST at four or more of the seven loci. The three-locus variant (3LV) is the cut-off point for inclusion in a clonal complex because four-locus variants (4LV) and higher (5LV to 7LV) are increasingly associated with unrelated

central STs. The majority of pairs found in the dataset, 221 (47.7%), were zero-locus variants (0LV), and the second largest group was the seven-locus variant (7LV), 130 (28.1%). The number of pairs found in the UK study and those pairs considered to be related, *i.e.* both isolates in the same clonal complex, was 58.7% of the pairs.

Table 3: Multilocus sequence type variant of pairs by centre. Overall 463 pairs were found. Each column represents a locus variant (LV), and the total is the number of pairs per variant, and the percent is based on the overall study.

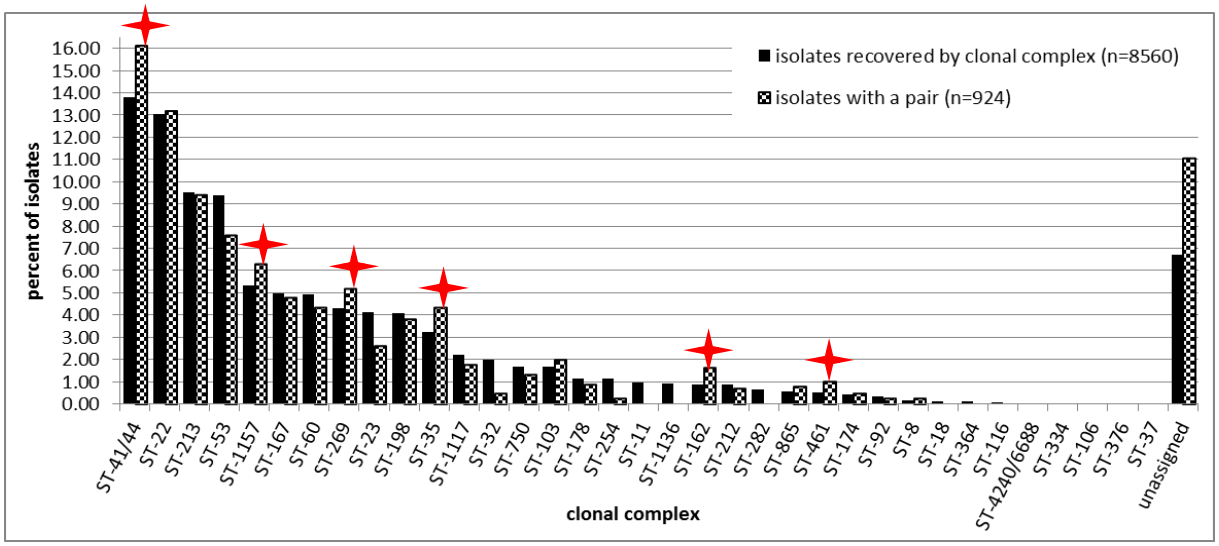
centre	related sequence types				unrelated sequence types			
	0LV	1LV	2LV	3LV	4LV	5LV	6LV	7LV
Bangor	14	4	1		1		6	11
Cardiff	13						3	11
Glasgow	40	3	2		1	3	6	16
Nottingham	24			1		1	4	7
Oxford	39	8	2		2	1	11	20
Plymouth	23	4					8	18
Stockport	68	15	8	3		2	12	47
total	221	34	13	4	4	7	50	130
	47.7%	7.3%	2.8%	0.9%	0.9%	1.5%	10.8%	28.1%

Regional datasets were investigated and compared to determine; (i) what proportion of individuals sampled on more than one occasion had a meningococcus recovered on more than one occasion; (ii) what proportion of pairs were related to each other; and, (iii) what clonal complexes the related pairs were associated with (Figure 2A-H). Clonal complex presence varied among regions and ranged from 24 clonal complexes in Bangor to 31 in Glasgow and Nottingham. There were 23 clonal complexes present in all regions. Seven clonal complexes had pairs in all regions, 14 complexes have pairs in two to six regions, and 11 complexes were not represented by any individual pairs. The latter 11 complexes represented only a single isolate each. The distribution of the clonal complexes containing pairs reflected, in part, the rank order of the clonal complex prevalence for the whole of the UK study dataset. Three regions all had a higher paired prevalence in clonal complex ST-461, two regions had higher pair prevalence in clonal complex ST-35 and two additional

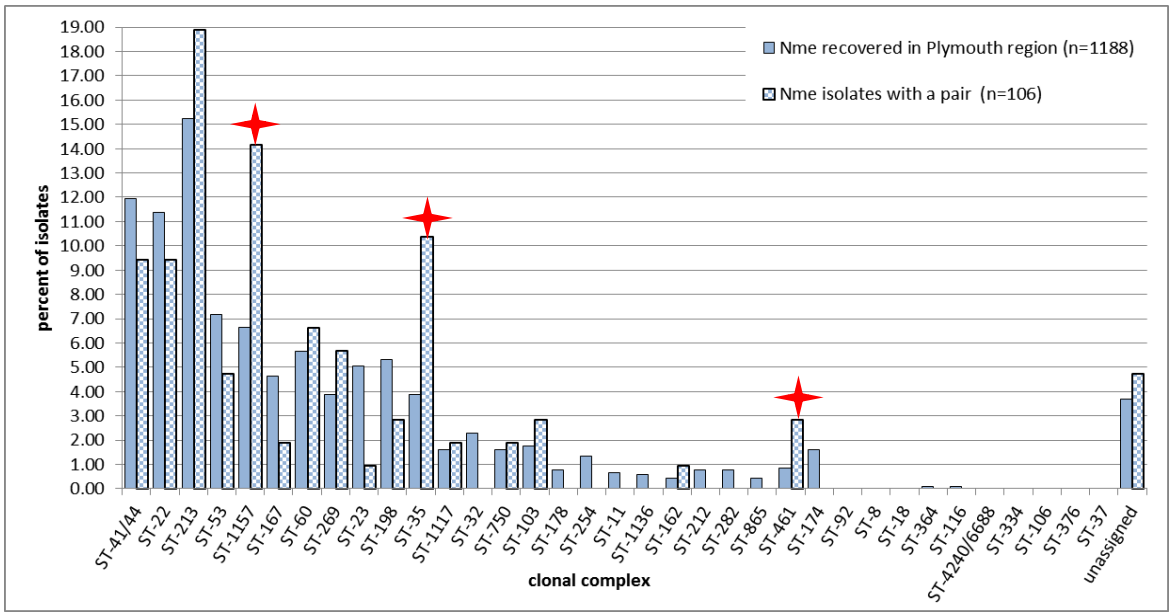
regions had a higher paired prevalence for clonal complex ST-41/44. Cumulatively complex ST-269 had a significantly higher paired prevalence, *p*-value 0.036, calculated using chi-square 2-tailed test. However, no individual region had a high prevalence of ST-269 pairs. Six clonal complexes, ST-41/44, ST-1157, ST-269, ST-35, ST-162 and ST461, had a higher frequency of temporally linked pairs, *p*-value 0.009 to 0.304, calculated using chi-square 2-tailed test (Figure 2A). Clonal complexes ST-22, ST-103 and ST-856 also appeared to have elevated association; however, they were not statistically significant (*p*-value 0.21 to 0.26), calculated using chi-square 2-tailed test.

Figure 2 (A-H). Frequency of paired *N. meningitidis* isolates recovered by clonal complex by region, from 1999-2001. Graph (A) shows the frequency of each clonal complex in the UK data set as compared to the frequency of paired samples in each regions. Six clonal complexes had a statistically higher, *p*-value <0.04 calculated using chi-square 2-tailed test, number of temporally linked pairs: ST-41/44, ST-1157, ST-269, ST-35, ST-162 and ST461 (indicated by a red star). Graphs (B) to (H) are single study regions with statistically higher paired frequencies indicated by a red star. ST-269 was the only complex over all that did not have a significantly higher paired assosciation in any single year. Colors between Figure 1 and Figure 2 panels B-H are coordinated.

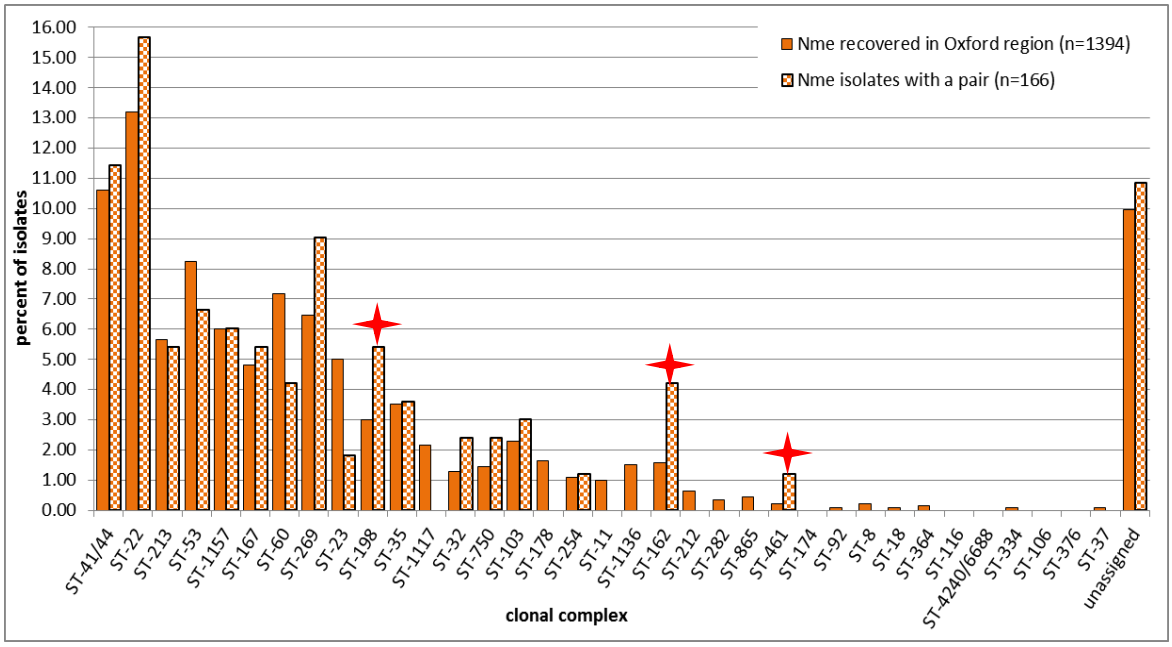
A. Comparison of all seven regions by clonal complex



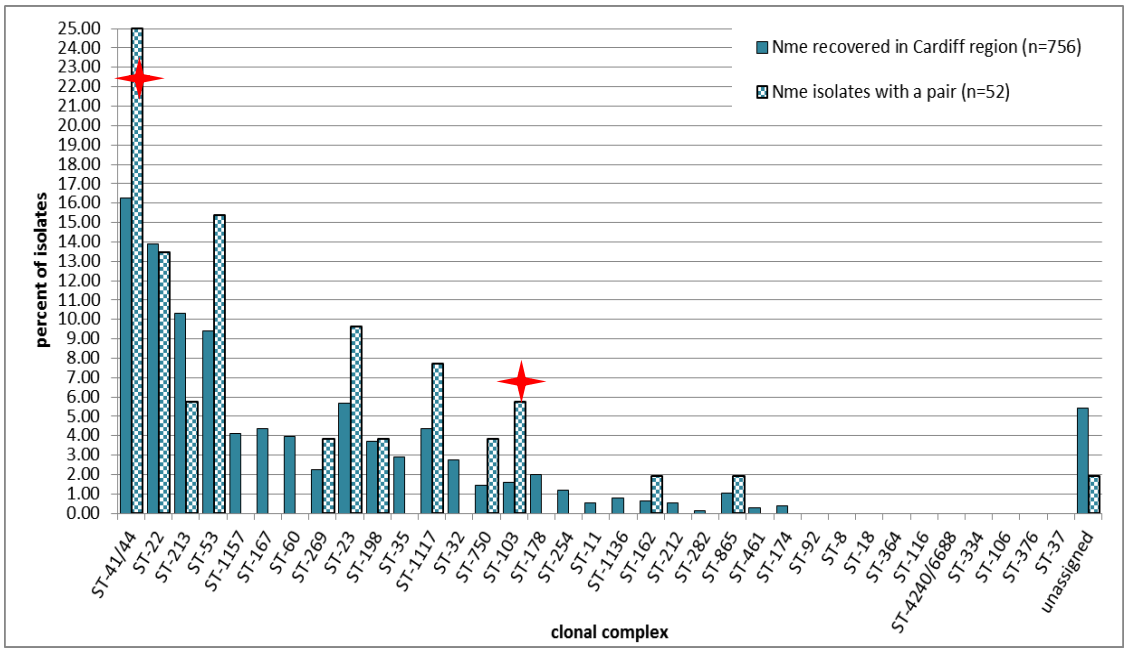
B. Plymouth region



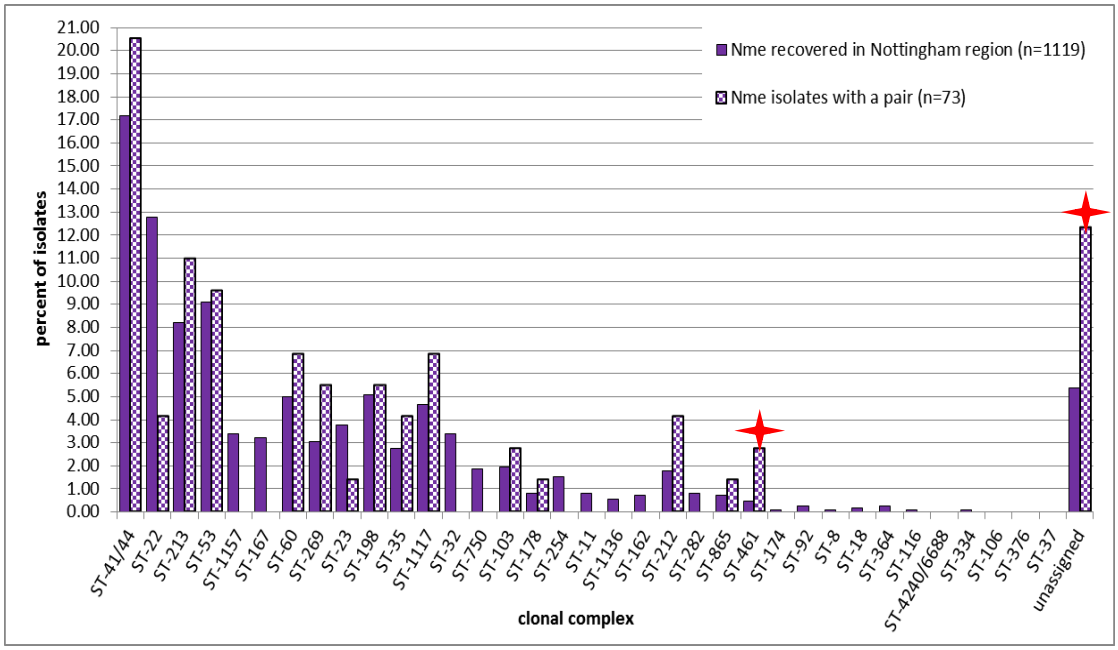
C. Oxford region



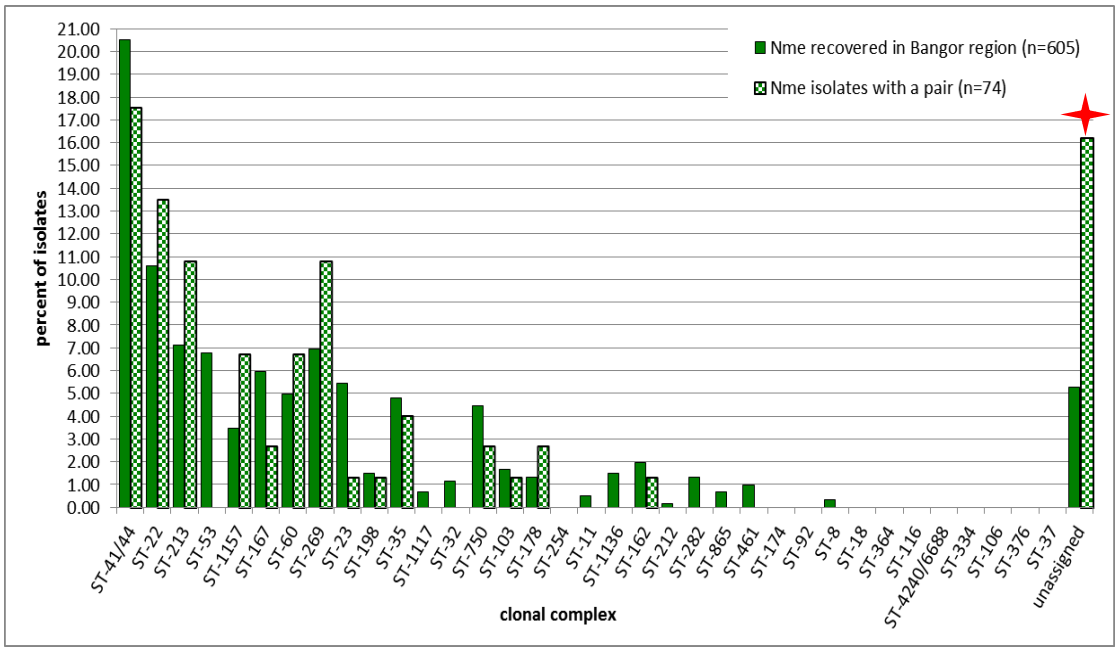
D. Cardiff region



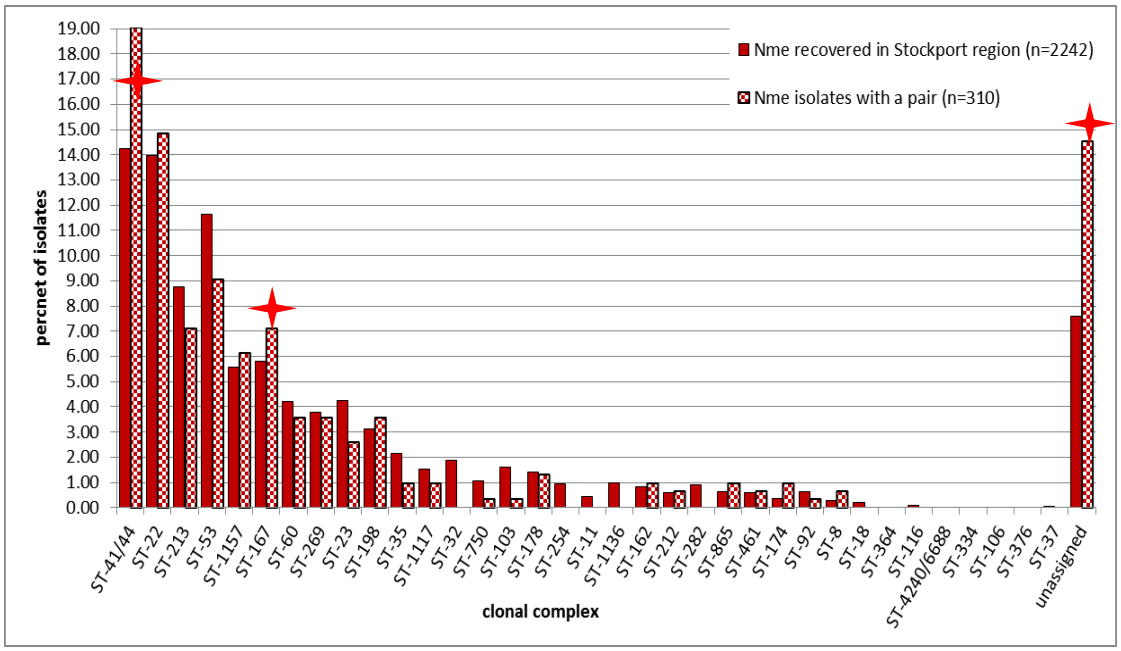
E. Nottingham region



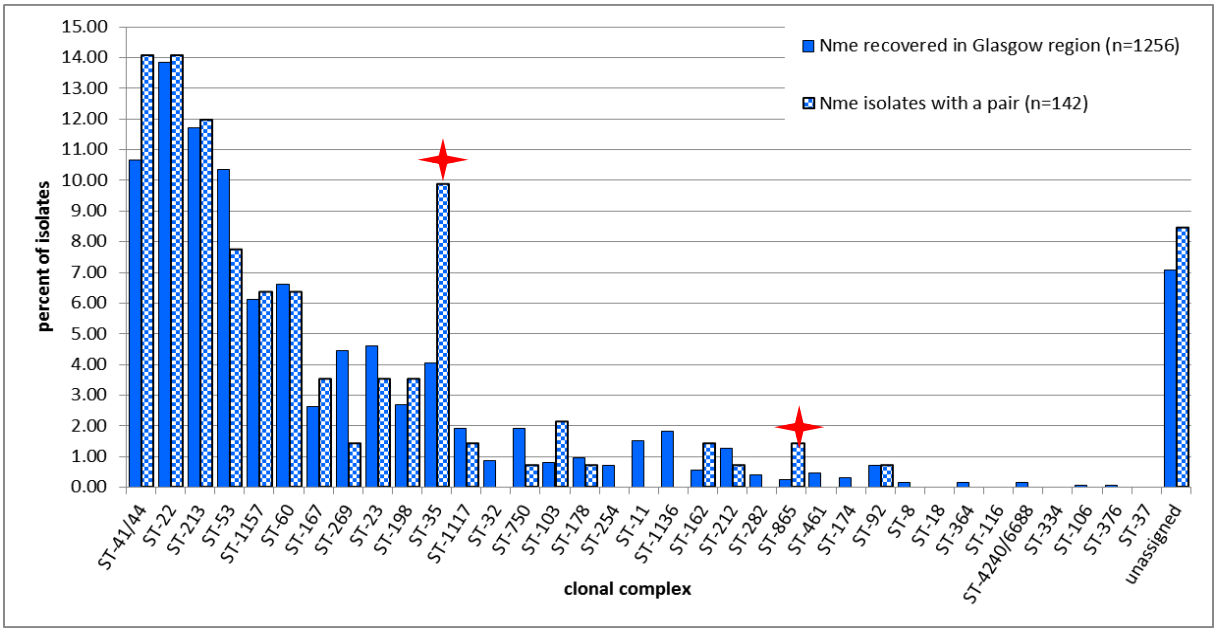
F. Bangor region



G. Stockport region



H. Glasgow region



2.2. Oxford regional carriage isolate collection

The meningococcal isolates recovered from Oxfordshire are held in the Department of Zoology at the University of Oxford, UK. The dataset is the second largest in the UK study, contained 16.3% of the study isolates, and was representative of the clonal complex prevalence of the large study. For these reasons, these isolates were chosen as the subset to study more in-depth.

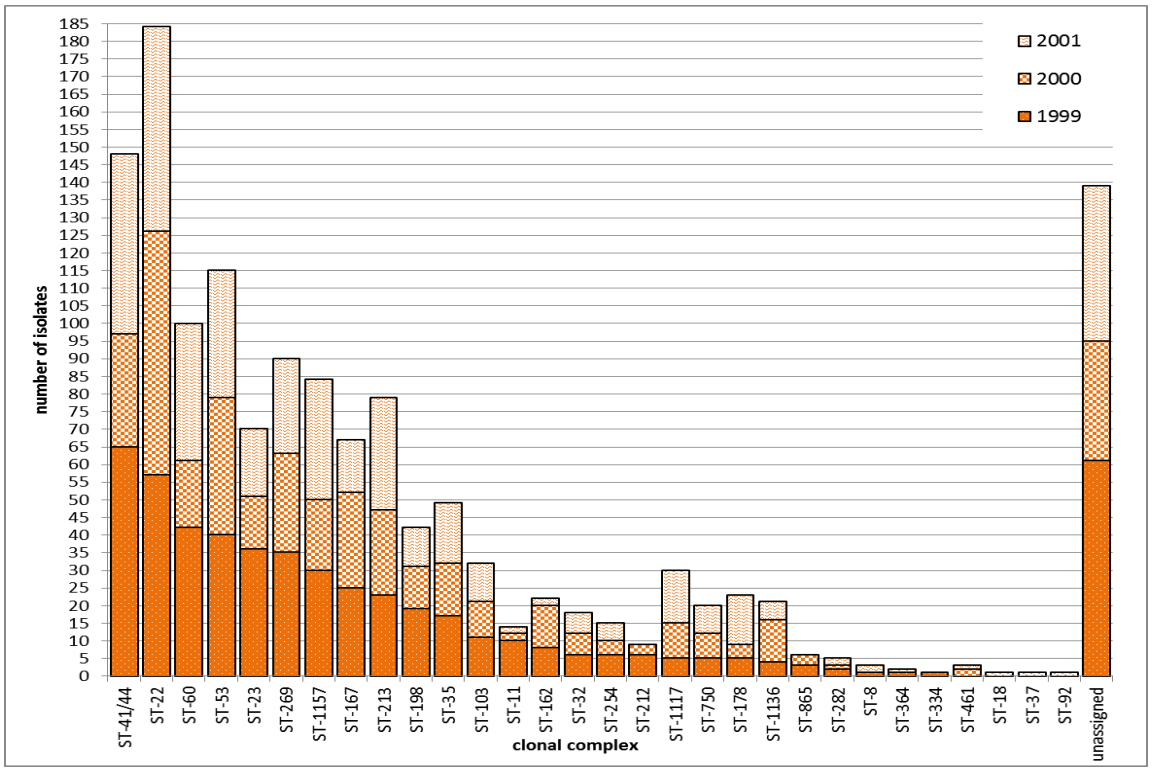
N. meningitidis was recovered from 1394 participants in Oxfordshire at least once over the three-year period. The most prevalent clonal complexes over the three years were ST-22 (184 isolates) and ST-41/44 (148 isolates) followed by the ST-53 (115 isolates) and ST-60 (100 isolates) (Figure 2C). The proportion of each clonal complex varied over the three years and contained a diverse range of genotypes (Figure 3). Thirty of the 35 clonal complexes found in the UK-wide study were present the Oxfordshire subset. There were 139 STs unassigned to a clonal complex and of those 76 (85.4%) were single occurrences, twelve STs were present two to six times and 20 isolates shared one profile, ST-963. This ST was not isolated in a successive year in any of the original carriers in which it was found but was found in all three years of the study in one school. Further investigation of the

database records for this ST revealed that 70% of the isolates were recovered from one school and 20% from a second school over the three years. All but two participants (90%) lived in the same postcode area, no further information was available in the database. This grouping of STs was the only phylogeographic correlation found in this dataset.

Based on the characteristics of clonal complexes defined in the *Neisseria* pubMLST.org database, 14 of the complexes (46.7%) were more often found associated with carriage, seven (23.3%) were more associated with isolates recovered from disease, and nine (30%) were commonly associated with both carriage and disease. However, the clonal complexes found in the Oxford region (and the UK-wide study) had some level of invasive meningococcal disease association (Appendix D, page 135).

Figure 3. Clonal complex prevalence in the Oxfordshire isolates, n=1394.

Clonal complexes are arranged left to right based on prevalence in 1999. Prevalence of each clonal complex varied year by year. Twenty complexes were found all three years. Complex ST-212 was not found in 2001, and complexes ST-461, ST-18, ST-37, and ST-92 were only found in 2000/2001



2.2.1. Oxford paired isolate collection

The isolates collected from Oxfordshire contained 83 individual pairs that were related by MLST typing. Of a total of 165 isolates, 81 individuals had *N. meningitidis* isolated in two consecutive years, and one individual in three successive years. The frequency of related pairs found among the clonal complexes was compared to the overall frequency of each clonal complex to determine if any particular clonal complex was associated with long-term carriage. Nine clonal complexes have a higher frequency of pairs; however, only three were statistically significant, *p*-value of 0.001 to 0.031 (Figure 2C) calculated using chi-square 2-tailed test. These were ST- 198, ST-162, and ST-461. Clonal complexes ST-41/44, ST-22, ST-269, ST-32, ST750 and ST-103 while elevated were not significant, *p*-value 0.061 to 0.378, calculated using chi-square 2-tailed test. Complexes ST-162 and ST-461 are associated with both carriage and disease while ST-198 is typically found in carriage isolates (Appendix D, page 135). The clonal complexes ST-41/44, ST-167, ST-32, ST-750, and ST-103 also showed a marginal increase in prevalence for paired isolates from an individual. While the ST-32 complex is commonly associated with disease, the clonal complexes ST-41/44, ST-167, ST-750 and ST-103 were associated with both carriage and disease states. There were also 34 pairs (41.0%) that were four to seven-locus variants and, therefore by definition, were unrelated to the clonal complex and represented two separate carriage episodes and were therefore not included in this analysis.

The original isolate typing data was extracted from the database for the 49 related pairs, those that were 0LV to 3LV. The list included 23 isolate pairs from 1999/2000, 25 pairs from 2000/2001 and one pair from 1999/2001. The samples were given identification codes by pair to simplify analysis (Table 4). The dataset contained 16 0LV pairs from 1999/2000 and 23 0LV pairs from 2000/2001, five 1LV from 1999/2000 and two 1LV from 2000/2001, two 2LV from 1999/2000, and a 1LV from a 1999/2001 pair. No triple-locus variants (3LV) were found in the dataset.

Table 4. Typing and ribosomal protein gene profiles. Pairs were coded alpha-numerically; pair id A-S represent Lineages 22, 2, 3, 13, 12, 15, 25, 19, 35, 6, 20, 26, 5, 31, 39, and 43 respectively and each iolate each lineage are numbered 1 to 18. Paired isolates were considered related if their MLST profiles were a zero-, one-, two- or three-locus variant. The Oxfordshire dataset did not contain any three-locus variants.

pair ID		isolate	strain designation					MLST variant	rST
			sero-group	PorA VR1,VR2	FetA VR	sequence type	clonal complex		
A	**1**	OX9930886	W	P1.18-1,3	F4-1	ST-1426	22	0LV	2415
	2	OX0050092							
	3	OX0052051	W	P1.18-1,3	F4-1	ST-184	22	0LV	2415
	4	OX01060508							
	5	OX0050781	W/Y	P1.18-1,3	F4-1	ST-2638	22	0LV	*I*
	6	OX01061392							2415
	7	OX9930103	W	P1.18-1,3	F1-62	ST-1221	22	0LV[a]	3069
	8	OX0050377							
	9	OX0050574	W	P1.18-1,3	F4-1	ST-1673	22	0LV[c]	*I*
	10	OX9931754							*I*
	11	OX9930617	W/Y	P1.18-1,3	F4-1	ST-22	22	0LV	2982
	12	OX0050946							
	13	OX9930387	W/Y	P1.18-1,3	F4-1	ST-22	22	0LV	2982
	14	OX0050965							
	15	OX9930530	Y	P1.18-1,3	F4-1	ST-22	22	0LV	3120
	16	OX0051415							
	17	OX0050776	W	P1.18-1,3	F1-5	ST-22	22	0LV	3194
	18	OX01061379							*
B	**1**	OX9932169	B	P1.19-1,15-11	F5-1	ST-269	269	0LV[c]	2423
	2	OX0050607							*I*
	3	OX9931800	B	P1.19-1,3-5	F5-1	ST-269	269	0LV	*I*
	4	OX0052881							2986
	5	OX0051091	B	P1.19-1,30-1	F5-1	ST-2091	269	0LV	3004
	6	OX01060731							
	7	OX0050049	B	P1.22,9	F5-12	ST-275	269	0LV	3011
	8	OX01061080							
	9	OX0050393	B	P1.19-1,3-5	F5-1	ST-269	269	0LV	2986
	10	OX01061948							6973
C	**1**	OX9932316	B	P1.7-2,4 P1.7-2,*deleted*	F1-5	ST-41	41/44	0LV[c]	3114
	2	OX0050570							
	3	OX9931715	*cnl*	P1.7-2,4-13	F1-5	ST-41	41/44	0LV	2991
	4	OX0050209							
	5	OX0050898	B	P1.19-1,26	F1-7	ST-44	41/44	0LV	2978
	6	OX01061840							
	7	OX0050873	B	P1.7-2,4	F1-5	ST-41	41/44	0LV	2554
	8	OX01062252							
D	**1**	OX9931776	B	P1.22,14	F5-5	ST-1644	213	0LV[b]	3010
	2	OX0051114							
	3	OX0050902	B	P1.22,14	F5-5	ST-213	213	0LV	2993
	4	OX01060373							
	5	OX0050325	B	P1.22,14-32	F3-9	ST-213	213	0LV	2979
	6	OX01060426							
	7	OX0051099	B	P1.22,14-32	F3-9	ST-213	213	0LV	2979
	8	OX01060686							

pair ID		isolate	strain designation					MLST variant	rST
			sero-group	PorA VR1,VR2	FetA VR	sequence type	clonal complex		
E	1	OX9930244	*cnl*	P1.7,30	F1-2	ST-53	53	0LV[a]	2984
	2	OX0051080							
	3	OX0052813	*cnl*	P1.7,30	F1-2	ST-53	53	0LV	2984
	4	OX01061911							3104
	5	OX9931971	*cnl*	P1.7,30	F1-2	ST-53	53	0LV[a]	2975
	6	OX0051106							
	7	OX9931590	*cnl*	P1.7,30-9 P1.7,30-4	F1-2	ST-53	53	0LV	2975
	8	OX0053048							
F	1	OX9932322	E	P1.21-7,16	F5-36	ST-1421	1157	0LV[c]	2933
	2	OX0050468							
	3	OX9931193	E	P1.21-7,16	F5-36	ST-1421	1157	0LV[a]	2933
	4	OX0052158							
	5	OX9931682	E	P1.21-7,16	F5-36	ST-1419	1157	0LV[a]	2933
	6	OX0053039							
	7	OX0050858	E	P1.21-7,16	F5-36	ST-1157	1157	0LV	2933
	8	OX01060268							
G	1	OX9931942	B	P1.22,14	F5-9	ST-162	162	0LV	2429
	2	OX0051290							
	3	OX9931975	B	P1.22,14	F5-9	ST-755	162	0LV[a,c]	2429
	4	OX0051317							
	5	OX0050158	B	P1.22,14	F5-9	ST-162	162	0LV	3030
	6	OX01061128							
H	1	OX9932212	*cnl*	P1.18,25	F5-5	ST-1956	198	0LV	*i*
	2	OX0050673							3008
	3	OX9930296	*cnl*	P1.18-9,25-9	F5-5	ST-198	198	0LV	3017
	4	OX0051153							
	5	OX0051143	*cnl*	P1.17-4,9	F5-5	ST-823	198	0LV[a]	2996
	6	OX01060852							
I	1	OX9931983	B	P1.19-3,15	F5-5	ST-3073	35	0LV[a]	3235
	2	OX01061987							*
	3	OX0050256	B	P1.22-1,14-3	F1-20	ST-35	35	0LV	2548
	4	OX01061099							
J	1	OX0051410	E	P1.19,15-1	F1-7	ST-60	60	0LV	3089
	2	OX01060129							3044
	3	OX0050852	E	P1.5,2	F1-7	ST-60	60	0LV	3179
	4	OX01060273							
K	1	OX0050185	Y	P1.18-1,3-16	F5-7	ST-862	103	0LV	3111
	2	OX01061141							
L	1	OX0052328	Y	P1.5-1,10-8 P1.5-1,10-1	F1-3	ST-767	167	0LV	3037
	2	OX01062316							
M	1	OX0050360	B	P1.7,16-2	F1-5	ST-74	32	0LV	3092
	2	OX01060380							
N	1	OX9931422	E	P1.5-1,2-2	F3-6	ST-254	254	0LV	*i*
	2	OX0050268							3141
O	1	OX0051189	B	P1.19-10,13-2	F5-5	ST-461	461	0LV	3001
	2	OX01060836							
P	1	OX0050081	*cnl*	P1.18-4,25 P1.18-4,25-49	F4-1	ST-1136	1136	0LV	3014
	2	OX01061117							
Q	1	OX9930407	*cnl*	P1.22-11, 13-22	F3-9	ST-2121	UA1	0LV[b]	3250
	2	OX0050971							
R^	1	OX0051148	ND	ND	ND	ST-1653	UA2	0LV	ND
	2	OX01060795							

pair ID		isolate	strain designation					MLST variant	rST
			sero-group	PorA VR1,VR2	FetA VR	sequence type	clonal complex		
S	1	OX0050901	*cnl*	P1.21,ND	F1-5	ST-3104	UA3	1LV[a]	3249
	2	OX01060284		P1.21,4		ST-2196			

ND - not done; NG - non-groupable; ST - sequence type; rST - ribosomal sequence type; *i* -1/53 loci found at the end of a contig (30-50 nucleotides missing); * Bowtie assembly was used to assemble the locus; [a] original typing recorded as a one-locus variant; [b] original typing recorded as a two-locus variant; [c] one locus from the genome assembly was incompletely assembled; 0LV - zero locus variant, both isolates have the same ST; 1LV single locus variant, isolates are different at 1 of the 7 MLST alleles; 2LV - double locus variant, isolates are different at 2 of the 7 MLST alleles; UA - unassigned ; ^ pair was removed from analysis, R1 could not be grown; cnl – capsule nul locus; *deleted* – no porA loop regions.

2.2.2. DNA preparation and whole genome sequencing of related pairs

The isolates had been archived as pure cultures in Brain Heart Infusion plus 10% glycerol and frozen at -80^{O}C for over ten years. The 49 related pairs were revived, cultured and their DNA was extracted as described in the Methods section 2-4, page 14-15. The majority of the isolates required an additional 24 hours of incubation (48 hours in total) at 56^{O}C in 5% CO_2 before subculturing for DNA extraction could be completed. One pair (R1/R2) was removed from the dataset because the R1 isolate could not be revived from the archived culture. Once extracted, the DNA was sent for Illumina® paired-end sequencing using 54 nucleotide base reads (Methods Section 4-5, page 15-16). The genomes were sequenced using the Illumina® GAII platform.

2.2.3. Genome assembly and assessment of coverage

The short-read sequences were assembled to a draft level using VELVET and the VELVETOptimiser script with scaffolding options switched off, and all other default settings unaltered [32, 33]. The draft genomes contained an average of 342 contigs each with a mean length of 2,112,790 nucleotides per assembly (Table 5 – assembly statistics). The L50 value, that represents the length at which contigs of equal or

longer length contain at least 50% of the assembled sequence, ranged from 5661 bp to 47,907 bp; with a mean length of 24,980 bp. The N50 mean was 30 contigs, however, it varied from 15 to 112 contigs per genome. The N50 statistic refers to the number of contiguous sequence assemblies with the L50 value. However, the L50 and N50 are not a measure of genome assembly quality. Overall, a higher k-mer setting for the assembly was associated with the higher N50 values and, within the bounds of the read length, assembled each of the genomes to the standard required for gene-by-gene analysis.

Genomes were uploaded to the *Neisseria* PubMLST.org sequence database and linked with their isolate record. The genomes were scanned against the sequence definition database for all defined loci, no predefined gene scheme was used. There was an average of 1794 NEIS loci identified per genome, and 97.5% of those loci were assigned a known allele number (Table 5 – annotation statistics). However because of the draft nature of the genome assemblies, and the short-read length (79 base reads) used for this dataset, annotation of the genomes fell as low as 1683 and only 82.1% (1411) alleles assigned. The assemblies, therefore, encoded an average of 97.5% of the 2000 coding sequences expected for a meningococcal genome.

Table 5: Sequence assembly statistics. Genomes were assessed using 2281 NEIS loci. Frequencies of tagged and designated are based on 2281 possible loci in the sequence definition database. All genome assemblies were within the established ranges for total sequence length and expected number of loci for a draft genome. The L50 represents the length at which contigs of equal or longer length contain at least 50% of the assembled sequence across the genome. The mean N50, or the number of contigs, per genome, was 30 but ranged from 15 to 112.

	assembly statistics							annotation statistics		
	contig count	Total length	Min	Max	Mean	Std Dev	L50	total loci tagged (%)	total alleles designated (%)	tagged loci with alleles designated
Mean	342	2112790	114	94045	6765	12447	24980	1794 (78.7)	1750 (76.7)	97.5%
Max	681	3895243	265	207453	14659	23277	47907	1882 (82.5)	1837 (80.5)	100.1%
Min	143	2004868	100	21054	3017	3368	5661	1683 (73.8)	1441 (61.9)	82.1%

Std Dev - standard deviation

2.3. Genomic typing of related pairs

2.3.1. Multi-locus sequence typing (MLST)

Analysis of the 96 genome sequences revealed that seven genomes from seven 1LV pairs (A7/A8, E1/E2, E5/E6, F5/F6, G3/G4, H5/H6, I1/I2) and one genome from each of the 2LV pairs (D1/D2, Q1/Q2) had MLST allele discrepancies (Table 4). The typing loci with discrepancies included five of the seven loci; *fumC* and *pgm* were not affected in this set of genomes. Analysis also found four additional MLST loci at the ends of contigs which failed to assemble fully. The incomplete loci that did assemble were missing 30% or less of their nucleotide sequence. To check the genomic sequence validity, the seven loci and the four incomplete loci were manually typed at the conflicting locus using a reserved lot of DNA archived from the genome sequencing stocks. The 2LV pairs were also genome sequenced a second time. The typing of each discrepant locus and the resequenced genomes had the same alleles as the typing derived from the first genome assembly for each of the discrepant allele sequences. Therefore, the sequence discrepancies could be attributed to

either editing or labeling problems in the original Sanger sequencing experiments. The pairs used for the genome comparisons therefore composed of forty-seven 0LV pairs and one 1LV pair, S1/S2.

2.3.2. Capsule type and antigen profiles

The serogroup had not previously been determined for several of the isolates; however, the capsular operon region A genes, responsible for capsule synthesis, were identified and used to predict the serogroup (Table 4). The capsule region A genes were identified in 74 genomes (77.1%) and included 19 pairs of serogroup B, seven pairs of serogroup E, five pairs of serogroup W and three pairs of serogroup Y; no pair had more than one serogroup. The remaining three isolates had an incompletely assembled sialic acid capsule biosynthesis gene, sialyltransferase, and were either serogroup W or Y. The differentiation of the locus that codes for both the W and Y capsule is a single amino acid at residue 310 in the N-terminal glycosyltransferase domain of the capsule polymerases, which is responsible for substrate specificity toward UDP-galactose or UDP-glucose [163]. The W/Y isolates were also identified as not being serogroup B or C types based on the orientation of the O-acetyltransferase gene (*csw/csy*) NEIS2164. The gene encoding UTP-glucose-1-phosphate uridyltransferase (*galU*) NEIS0580 was also present in all three of the W/Y genomes. Twenty-two genomes (11 pairs) did not contain the capsular region A genes and were annotated as *cnl*, capsule null.

Strain type including PorA VR1, PorA VR2, and FetA VR were extracted from each genome. Each pair shared identical PorA VR1 and FetA VR types, and 44 pairs each had identical PorA VR2 types. The remaining four pairs (C1/C2, E7/E8, L1/L2, and P1/P2) had differing PorA VR2 types between their genomes. In three pairs, E7/E8, L1/L2, and P1/P2, the PorA VR2 antigen family was the same but the variant differed between the genomes in each pair. Genome E7 has a homopolymeric twelve amino

acid repeat in the PorA VR2 region, half of which were absent in genome E8. Likewise, genome L1 has a six amino acid inverted repeat, half of which was not present in genome L2. The pair P1/P2 PorA VR2 was different between two consecutive codons, both non-synonymous nucleotide changes, in the PorA VR2 region. In the fourth pair, while genome C1 contained a complete PorA VR2 loop, C2 had a deletion of the PorA VR2 loop.

2.3.3. Ribosomal protein gene sequence type (rST)

The paralogous loci for each pair were identified and their alleles, when compared, were found to be identical between the genomes of each pair. Overall there were 19 allele combinations for *rpmE1/rpmE2* and ten combinations of *rpmJ1/rpmJ2* alleles found among the paired genomes. The allele combination 1/4899 was most prevalent (18 occurrences, 18.4%) for *rpmE1/rpmE2* and 1/1010 was most common (26 occurrences, 27.1%) in *rpmJ1/rpmJ2*. The combinations are somewhat congruent with lineage association, but the dataset is too small to draw any conclusions. For example, the *rpmE1/rpmE2* allele combination 1/4899 is observed in all Lineage 22 pairs, and the *rpmJ1/rpmJ2* combination 1/1010 is present in all but one pair (which is 1/3583), but the latter is shared with Lineages 19, 20, 43, and is also present in the unassigned S1/S2 pair.

There were nine genomes with unassigned rSTs; this was due to one allele in each profile being found at the end of a contig and therefore incompletely assembled. In two of the nine instances, A18 and I2, the incomplete locus (*rpsA*) was present but split between two contigs. Bowtie was used to reassemble the short sequence reads. Bowtie is a computationally efficient short-read sequence alignment program that indexes a reference sequence for assembly by sequence mapping. For each of the two genomes a complete locus was used as a reference, the sequence was identified by the BIGSdb software as having the nearest sequence identity match to

one or both of the locus fragments. The resulting assembled sequences were uploaded to their respective sequence bins and assigned allele numbers. There was no obvious sequence complexity in the locus sequence, and the incompletely assembled loci could be attributed to the short read length. Three pairs did not share the same rST, pair B9/B10, E4/E5 and J1/J2. Both pair B9/B10 and E4/E5 had a single nucleotide transition in the *rplW* and *rplU* protein genes respectively while pair J1/J2 had a single nucleotide transversion in the *rplJ* protein gene. All remaining pairs had the same rST by pair. All 96 genomes shared the same allele for *rpsS* and r*plX*. The number of rST found by lineage varied. All four pairs of the Lineage 15 genomes had the same rST type while Lineage 22 had nine rSTs among the nine pairs. The remaining lineages where rST profiles were complete and had more than one pair, exhibited at least two or more rST profiles within the lineage.

2.3.4. Whole genome comparisons

Based on the diversity of the isolates no single genome could be used as a reference for whole genome analysis. Therefore, all NEIS loci defined in the *Neisseria* pubMLST.org sequence definition database, 2281 loci at the time of analysis, were used to compare the 48 genome pairs. This comparison was chosen instead of using the predefined *N. meningitidis* core genome because the current scheme definition is almost exclusively based on disease genomes and may not represent the core genes of carried meningococci. Loci that were missing due to the absence in a particular genome or had an incomplete assembly were ignored in pair-wise comparisons. There were 120 NEIS loci not identified in the genomes and the majority (73.3%) were known mobile elements, such as the conjugative plasmid, gonococcal genetic island, beta-lactamase plasmid, the VirB T4SS plasmid, and 14.2% included the capsule genes specific for serogroups not found in this isolate collection, and the genes encoding the Opacity family proteins. The average number of loci tagged

across all genomes was 1794, which is 89.7% of the meningococcal genome complement of coding loci based on the published meningococcal genomes.

The genomes collectively contained a core set of genes totaling 1546 and are 71.5% of the total number of coding loci found within the genomes. The majority of the core loci (68.9%) had a *p*-distance estimate between 0.00 and 0.04, and of these 959 loci had *p*-distance estimates of <0.02. There were loci that shared identical alleles within each lineage. As expected, the number of loci with identical alleles in each lineage decreased as the number of pairs increased (Table 6). For example, lineages represented by a single pair had over 1700 identical loci except pair N1/N2, which had 1671; and the count fell as low as 540 identical loci in pairs B1 to B10 (Lineage 2) which has five pairs. There were six identical loci across all 96 genomes: three ribosomal protein genes (NEIS0136 *rpsS*, NEIS0140 *rpmC*, and NEIS0143 *rplX*), two hypothetical proteins (NEIS0441 and NEIS2515) and the ATP synthetase subunit C protein gene (NEIS1911). Each pair of genomes contained different counts of identical loci. The alleles with 100% identity between pairs ranged from 1412 to 1841 with an average of 1739 per pair, approximately 87% of the expected 2000 meningococcal genes. The core genes with higher *p*-distance estimates, >0.1 (73 in total), were mainly categorized as hypothetical proteins (43.8%), genes encoding membrane-associated proteins (27.4%), and those involved in genetic information processing (9.6%).

Table 6: Annotation by lineage and sequence variation. Genome assembly coverage was assessed using 2281 NEIS loci. The *p*-distances were computed using the Kimuras 2-parameter model (1980) with Bootstrap, which corrects for multiple hits. The model also takes into account transitional and transversional substitution rates, while assuming that the four nucleotide frequencies are the same and that rates of substitution do not vary among sites.

whole genome nomenclature	clonal complex	genome count	same allele*	variable loci	missing in all	incomplete loci	100% core	*p*-distance	S.E.
Lineage 22	ST-22	18	939	662	355	325	1885	0.003	0.000
Lineage 2	ST-269	10	545	943	323	470	1856	0.007	0.000
Lineage 3	ST-41/44	8	621	1194	360	106	1834	0.009	0.000
Lineage 13	ST-213	8	1179	524	381	197	1834	0.003	0.000
Lineage 12	ST-53	8	1433	210	486	154	1787	0.001	0.000
Lineage 15	ST-1157	8	1321	276	444	240	1803	0.002	0.000
Lineage 25	ST-162	6	1496	180	415	190	1854	0.001	0.000
Lineage 19	ST-198	6	871	847	416	147	1840	0.006	0.000
Lineage 35	ST-35	4	770	1003	321	187	1834	0.010	0.000
Lineage 6	ST-60	4	1253	184	459	385	1787	0.002	0.000
Lineage 5	ST-32	2	1796	48	354	83	1915	0.000	0.000
Lineage 20	ST-103	2	1735	45	427	74	1853	0.000	0.000
Lineage 26	ST-167	2	1730	57	408	86	1860	0.000	0.000
Lineage 31	ST-254	2	1671	32	478	100	1800	0.000	0.000
Lineage 39	ST-461	2	1718	34	426	103	1839	0.000	0.000
Lineage 43	ST-1136	2	1718	44	425	94	1855	0.001	0.000
NA	UA1 [a]	2	1706	19	493	63	1777	0.014	0.000
NA	UA3 [a]	2	1766	32	397	86	1879	0.013	0.000

NA – not assigned; [a] the ST is not part of a defined clonal complex; * the total number of loci that share the same allele in all genomes; S.E. – standard error

All 18 of the Lineage 22 and the Lineage 20 genome pair contained remnants of a type IV (T4SS) gonococcal genetic island (GGI). The GGI was homologous to the one described in the Lineage 5 *N. meningitidis* [35]. The island contains 73 genes (NEIS2250 to NEIS2322), and the genomes were all missing seven of the loci NEIS2274-2278, 2295, and 2315. The remaining 66 loci were present in all the Lineage 22 genomes. However, the Lineage 20 genomes were also missing loci NEIS2257 to 2279 and NEIS2296 to NEIS2314. No other lineages contained the gonococcal genetic island or its defined genes.

Several regions of each genome were interrupted by incomplete sequence assembly and the presence or absence of loci in these regions could not be properly assessed. The regions of incomplete assembly contained paralogous loci or loci with repeat sequence tracts longer than the sequence read lengths; such as the genes encoding the Maf proteins, FrpC proteins, and DNA transport competence proteins, as well as transposases and hypothetical proteins.

The MDA island, a member of the filamentous prophage family and found to be associated with hyperinvasive lineage, was found in 40 of the genomes. The genomes containing the MDA island were assigned to lineages associated with disease: the ten Lineage 2, eight Lineage 3 and six Lineage 25 genomes, the Lineage 5 pair and the Lineage 13 pair. The MDA island genes, but not the complete element, was found in the eight Lineage 15 genomes, a lineage associated with carriage; and one of the two Lineage 35 pairs, a lineage linked to both disease and carriage.

A set of three genes were found in only 20 genomes, those isolates belonging to Lineages 12, 13 and 26. The genes included the DNA methylase (NEIS0327), the modification methylase DpnIIB (NEIS0328) and a locus described in the database as a hypothetical protein (NEIS0329). Both of the methylase genes are adenine-specific and involved in DNA mismatch repair and type I and type II site-specific restriction-modification systems. A Pfam search using the hypothetical protein peptide sequence of NEIS0329 found a high level of similarity to the plasmid conjugative transfer protein PilI. These genes were not found in any other Lineage. The Lineages 12 and 13 are carriage and disease associated respectively, and Lineage 26 is equally associated with invasive disease and carriage. A second set of lineage linked restriction-modification related genes was also found. The NEIS2390 and NEIS2391 loci also encode type I restriction enzymes and were found in all genomes except those belonging to Lineage 19 and Lineage 43, with six and two genomes respectively, both carriage associated lineages.

2.4. Phylogenetic analysis

A genomic analysis was carried out using the gene-by-gene approach [87, 96]. The BIGSdb Genome Comparator tool was used to generate distance matrices that were used to construct phylogenetic networks. Networks were then visualized using SplitsTree and annotated using InkScape [45, 164, 165].

2.4.1. Genome-level MLST

The phylogenetic network based on the core genes from the 2281 NEIS loci provided a high-resolution population structure of lineages and multi-lineage clades. Each pair formed a cluster. Those pairs from the same lineage were more closely related and formed distinct monophyletic groups (Figure 4). Isolates from Lineages 2, 5, 35 grouped together to form a larger clade as did Lineages 19 and 43; and each clade showed a clear common ancestry. Branch lengths and the internal nodes of the two multi-lineage clades indicated lineages were older than transmission events. Individual lineages were also compared using the 2281 NEIS loci. There were several loci in each pair that were incompletely assembled; with two exceptions the number of incomplete loci was less than 10% of the total number of loci present each genome. Each pair of genomes had, on average, 1878 loci to compare. Two of the pairs had higher levels of incomplete loci, B1/B2 and J3/J4 with 14% and 20% incomplete loci respectively. The average number of loci different between the genomes of a pair was 40; though in pair A5/6 there were 102 locus sequence differences.

Distance estimations for each pair of sequences were calculated in MEGA6 using the nucleotide sequence and the translated amino acid sequence. The Kimura 2-parameter with uniform rate among sites model was used for both tests with 500 bootstrap replications per run (Table 7). The average *p*-distance among all pairs was 0.008 for the nucleotide comparison and 0.013 for the translated amino acid comparison. As expected the *p*-distance estimate for the amino acid translation

comparisons increased due to the sequence variation found between the paired genomes. The within lineage mean *p*-distances was from 0.000 to 0.014, and an average *p*-distance of 0.004 among the lineages. Three pairs have higher sequence variation, A5/A6 and B1/B2 have 102 and 91 allele differences, and pair C3/C4, while only 23 alleles differed, and the estimated *p*-distance was 0.046. The remaining pairs fell with in a range of 13 to 64 variable alleles and p-distance estimates of 0.000 to 0.025 for nucleotide sequence and 0.000 to 0.037 for amino acid.

Individual lineages were also compared using Genome Comparator and the 2281 NEIS loci. As a representative, only the Lineage 22 tree has been visualized using the SplitsTree and annotated using InkScape (Figure 5). The lineage is representative of many of the phylogenetic characteristics observed in the lineages represented by four or more pairs. Lineage 22 had nine 0LV pairs; and includes six STs, six rSTs, and three antigen profiles. All genomes shared the same PorA VR1, PorA VR2 type (P1.18-1,3) while eight pairs shared the same FetA VR (F4-1), and four shared the same sequence type, ST-22. The STs were either a 1LV or a 2LV of the central sequence type ST-22. The list of variable loci was filtered for paralogous loci and those loci that did not assemble well. Each pair of genomes varied at an average of 48 loci.

2.4.2. Ribosomal protein gene (rST) phylogeny

The 53 rMLST loci were extracted from the database, concatenated, aligned and used to generate the rST number for each isolate [22]. All isolate pairs clustered together, and 39 rST profiles grouped all corresponding pairs into well-defined lineages that were the same as the whole genome phylogenetic clustering. There were no rST profiles shared between lineages. Visual comparison of the whole genome phylogeny to the rMLST phylogeny showed congruence of lineages and lineage clusters.

Figure 4. Whole genome, Neighbor-Net. Genomes assigned to the same clonal complex were compared at 2281 loci. To reflect the increased resolution of whole genome typing, a lineage nomenclature was used to distinguish groupings. Numbers in parenthesis are the total number of isolates; those without a number were a single pair. Isolates labeled UA1 and UA3 are currently not associated with an identified clonal complex and are the only representatives of their lineage in this dataset.

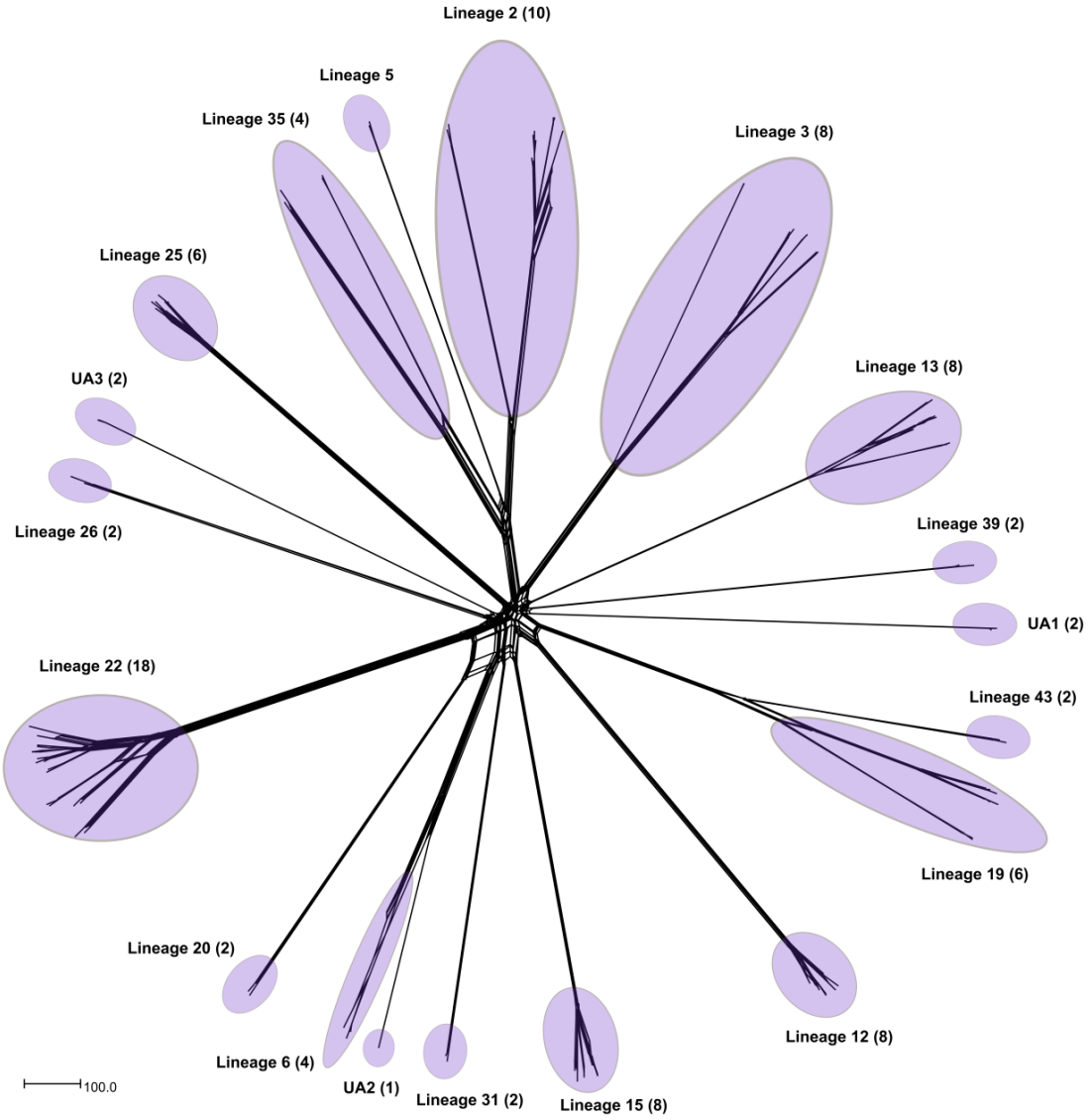


Table 7: Estimates of average evolutionary divergence of each pair. The *p*-distance was computed using the Kimuras 2-parameter model with bootstrap (500 replicates), which corrects for multiple hits and takes into account transitional and transversional substitution rates, while assuming that the four nucleotide frequencies are the same and that rates of substitution do not vary among sites. Standard error (S.E) estimates are shown. The coding data was translated assuming a standard genetic code table. All ambiguous positions were removed for each sequence pair.

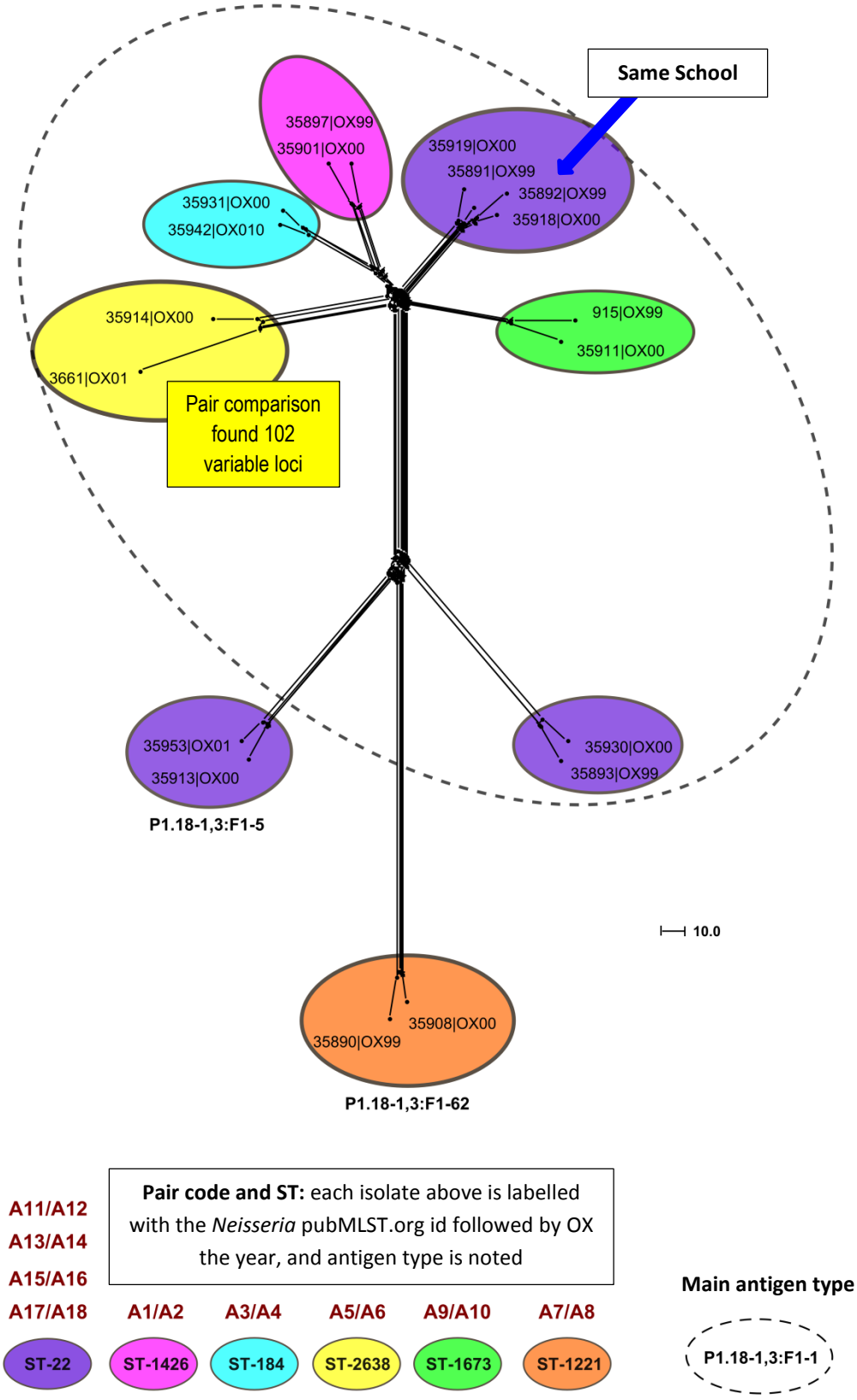
pair	Loci				nucleotide		animo acid	
	variable	missing	identical	incomplete	*p*-distance	S.E.	*p*-distance	S.E.
A1|A2	57	365	1793	73	0.004	0.000	0.006	0.001
A3|A4	27	369	1841	51	0.008	0.001	0.013	0.001
A5|A6	102	365	1761	60	0.002	0.000	0.004	0.000
A7|A8	39	361	1801	88	0.001	0.000	0.002	0.000
A9|A10	64	364	1761	99	0.005	0.000	0.009	0.001
A11|A12	37	371	1798	82	0.010	0.000	0.015	0.001
A13|A14	33	369	1804	82	0.005	0.000	0.008	0.001
A15|A16	44	371	1819	54	0.002	0.000	0.002	0.000
A17|A18	36	361	1739	152	0.009	0.001	0.014	0.001
B1|B2	91	347	1578	272	0.016	0.000	0.022	0.001
B3|B4	44	348	1744	152	0.025	0.001	0.041	0.001
B5|B6	25	360	1827	76	0.011	0.001	0.014	0.001
B7|B8	39	356	1780	113	0.006	0.000	0.009	0.001
B9|B10	59	354	1693	182	0.015	0.000	0.024	0.001
C1|C2	40	389	1786	73	0.003	0.000	0.005	0.001
C3|C4	23	393	1830	42	0.046	0.001	0.043	0.002
C5|C6	16	409	1831	32	0.020	0.001	0.019	0.002
C7|C8	13	379	1841	55	0.016	0.001	0.025	0.002
D1|D2	32	397	1754	105	0.018	0.000	0.037	0.002
D3|D4	18	416	1776	78	0.001	0.000	0.003	0.001
D5|D6	41	413	1682	152	0.025	0.001	0.037	0.001
D7|D8	28	414	1749	97	0.005	0.000	0.007	0.001
E1|E2	30	489	1723	46	0.004	0.000	0.006	0.001
E3|E4	37	490	1683	78	0.009	0.001	0.017	0.001
E5|E6	43	490	1659	96	0.002	0.000	0.004	0.001
E7|E8	47	486	1701	54	0.014	0.000	0.022	0.001
F1|F2	43	460	1627	158	0.006	0.000	0.010	0.001
F3|F4	33	461	1712	82	0.005	0.000	0.008	0.001
F5|F6	49	454	1690	95	0.007	0.000	0.008	0.001
F7|F8	22	452	1735	79	0.004	0.000	0.006	0.001
G1|G2	33	416	1754	85	0.006	0.000	0.007	0.001
G3|G4	35	418	1738	97	0.003	0.000	0.007	0.001
G5|G6	43	415	1699	131	0.008	0.000	0.012	0.001
H1|H2	29	420	1756	83	0.002	0.000	0.003	0.001
H3|H4	24	425	1771	68	0.006	0.000	0.009	0.001
H5|H6	42	433	1740	73	0.002	0.000	0.004	0.000
I1|I2	50	358	1749	131	0.019	0.000	0.023	0.001
I3|I4	50	391	1747	100	0.004	0.000	0.007	0.000

pair	Loci				nucleotide		animo acid	
	variable	missing	identical	incomplete	*p*-distance	S.E.	*p*-distance	S.E.
J1|J2	32	468	1720	68	0.014	0.000	0.020	0.002
J3|J4	44	465	1412	376	0.007	0.001	0.013	0.001
K1|K2	47	427	1739	75	0.000	0.000	0.007	0.001
L1|L2	57	408	1733	90	0.000	0.000	0.008	0.000
M1|M2	48	355	1802	83	0.000	0.000	0.010	0.001
N1|N2	35	478	1673	102	0.000	0.000	0.004	0.000
O1|O2	36	426	1722	104	0.000	0.000	0.003	0.000
P1|P2	46	425	1722	95	0.001	0.000	0.024	0.001
Q1|Q2	20	492	1713	63	0.014	0.000	0.021	0.002
S1|S2	32	397	1772	87	0.013	0.000	0.020	0.001

[a] MLST sequence type is currently not associated with the identified clonal complexes

S.E. standard error

Figure 5. Whole genome comparison of Lineage 22 pairs. The lineage was observed most often in the Oxford carriage pair isolate collection. Each pair was highly conserved with *p*-distance estimates of 0.001 to 0.010 and varied at 27 to 102 loci.



3. Discussion

While the UKMenCar study was not designed to investigate long-term carriage or the analysis of sequential isolates in persistent carriage, the repeated swabbing of the same population successively captured the same participants durring each annual swabbing for each of the three years of the study (Table 2). However, it provided the opportunity to do so in a limited way due to the sampling frame and the repeated annual visits to the same schools.

The persistence and duration of meningococcal carriage varies; it may be transient, with a duration of just a few days or weeks, or chronic lasting for several months, depending on the host and strain characteristics [91, 167]. The findings presented here, are consistent with reports of colonization persisting for at least six months. This supports the results from a modelling study which estimated that a stable carriage relationship may last for up to 29 months [168, 169]. The interval, of approximately one year, between the recovery of related paired isolates in this dataset is consistent with this and indicates that chronic carriage could last much longer in some individuals. The true persistence and duration of carriage, however, remains uncertain.

Certain lineages remained highly stable over time, and some strains appeared to be adapted to the commensal role. In this analysis Lineage 22 was most frequently found, 18 pairs representing all three swabbing time points and multiple schools (Figure 5). This is interesting because within the *Neisseria* pubMLST.org database, about 55% of isolates are from cases of invasive disease. Therefore this lineage is demonstrating attributes of both invasive potential and the ability to be carried for long periods of time. The low level of variation found between each pair, of all lineages, is similar to the variation found in patient contact transmission studies [114]. This indicates that these pairs were in fact carried by a single individual for approximately one year. In contrast the unrelated pairs, those identified as 7LV, indicate reinfection and a non-stable carriage relationship. This finding is similar to the non-over lapping antigen models [170]. Comparing all of the Lineage 22 genome pairs from the remaining regions, as well as the genomes from any single individuals carrying a Lineage 22 one year and not the next, would offer a chance to investigate if the genomes of cleared isolates differ from chronically carried isolates.

Virulence shortens the transmission window, and invasive behaviour is an evolutionary dead-end for the meningococcus. It has been suggested that the ability to transmit is more important to the meningococcal lifecycle than virulence [143]. The mechanism explaining this is 'transmission fitness' and proposed that key metabolism and biosynthesis genes affected by selection influence fitness; and the selection results in a lineage associated allele set [171]. This is consistent with the clustering of lineage-specific and pair specific genomes found (Figure 5). Comparison of 1875 loci in Lineage 22 found a highly conserved core, *p*-distance of 0.003; of those loci 1363 had identical alleles in all 18 genomes. This finding was similar for all lineages within this dataset and supports a genome-wide allelic lineage structure that is stable and persists over long periods of time.

Each of the forty-nine pairs was compared at the genome level and sequence concordance between the genomes of each pair was conserved. There was a suggestion that within-host evolution might be occurring, pairs A5/A6, B1/B2, and C3/C4 (Table 6). However the differences between these pairs could be the result of natural variation in the population colonizing a single individual. Without comparing multiple picks from multiple sequential isolates, taken temporally from a single chronic carrier, no conclusions can be drawn.

The level of diversity found in the current study is concordant with other reports that found carriage associated strains were more diverse than those associated with invasive disease [91]. Lineage replacement was observed in 40% of individuals found to be carrying in a successive year. This was in contrast to previous cohort studies of carriage in teenagers and military recruits that found 0.7 to 13% lineage replacement over a seven to nine month period [172, 173]. While simultaneous colonization by multiple lineages could have occurred, this was not possible to assess as only one isolate was stored for each individual. The analysis presented here also highlights that strains characterized by MLST as being one- or two-locus variants should not be over-interpreted as different and may, in fact, be more closely related than zero-locus variants (Figure 5).

Many of the carriers that did not present with a positive meningococcal swab in a successive year may still have persistently carried at some point during the intervening year, but cleared colonization before the next swabbing. Participant demographics were taken at the time of the

carriage study and included antibiotic use, smoking habits, and attendance at pubs and clubs. While this information was not accessed for this study, these are predisposing host factors shown to affect meningococcal carriage [174, 175]. This is also consistent with other studies demonstrating a link between cigarette smoke and increased adherence in other respiratory pathogens including *Neisseria* speices, *Streptococcus pneumoniae, Bordetella pertussis, Haemophilus influenzae, Moraxella catarrhalis,* and *Staphylococcus aureus* [176, 177].

The aim of the analysis was to provide information on the relationship between paired isolate genomes from individual carriers at two time points approximately one year apart and to identify the level of genetic similarity between the related isolate pairs and the mechanisms of genetic diversity present. The data presented here showed that certain lineages remained highly stable over time and were unique to the individual carrying them. However, while the genomic diversity was low and specific to the individual pairs, the genomic diversity between individuals with the same lineage was high. The genealogies reconstructed with the Neighbor-Net algorithm cgMLST were consistent with those previously generated with MLST and a variety of other approaches [96, 178]. Together this suggests that a defined network of meningococcal carriage does not exist. However the addition of more carriage study genomes to the database could be beneficial in elucidaitng transmission fitness between chronically carried strains and those that are cleared by the host.

Chapter 6 General Discussion

Genomic studies at the population level have been made possible by the advances in nucleotide sequencing technologies and bioinformatics. These advances have markedly improved the type and quality of data available for analysis. Whole genome sequence data has the potential to provide comprehensive descriptions of genomic variation. The meningococcal core genome as described herein contains 1605 loci; however a survey (September 2015) of the *Neisseria* PubMLST.org sequence definition database contained an additional 600 accessory loci. The loci are grouped into multiple schemes, from typing profiles to pathways such as metabolic functions or genetic information processing, as well as core meningococcal lineage, species, and genus core loci. Within the meningococcal species the diversity is extensively structured, despite high levels of recombination [90]. The population structure, which is evident from only seven housekeeping genes located throughout the chromosome, is associated with particular phenotypes, including the expression of specific antigen types and the likelihood of causing invasive disease [90]. Dispite this, there is a general lack of theoretical framework to describe the structure or the evolution of lineages within the meningococcal population. The work untaken here presents an approach for population based genetic studies, which incorporates a contextual background, within which the population structure can be investigated.

Described in Chapter Three, the use of *de novo* assembly and the gene-by-gene approach utilizes generic, freely accessible and widely used tools and produced high quality well characterized isolates [96]. The proof of concept, for producing notional whole genome assemblies (or draft sequences) for analysis, was shown to be robust to a range of DNA preperations and reliable in assembling sequence with high quality. This model of population genetics has been successfully used to interogate *N. meningitidis* and other bacterial speices [87, 131]. The use of whole genome comparions enable current biological and evolutionary models to be refined and developed. While bacterial sequencing does not suffer from the production of raw sequence data, the rate-limiting step has become working with the generated data. The assembly, annotation, and computational analysis all take an

order of magnitude longer in time to complete than the sequencing procedure. Therefore exploitation of whole genome data depends on the resource being available in a structured, easy to use form.

The BIGSdb platform accommodates a range of sequence data from single gene through to multiple genes, contigs and whole genomes [34]. The gene-by-gene approach is a practical approach to cataloguing sequence diversity efficiently, and permits analyses of a single genome or multiple genome comparisons. The research chapters presented here used a combination of single pairwise genome comparisons and population pairwise comparisons, without having to reanalyse or realign the genomes being used. The use of the web interface within the *Neisseria* PubMLST.org database enables a process of annotation whereby members of the community can participate in, help maintain, and suggest improvements.

Analysis of a defined population of isolates that are known to have diverse phenotypes provides an invaluable tool for exploring the evolution of lineages within a species. The methods presented here are able to generate novel insights into the biology of the meningococcus and improve our understanding of the whole population structure, not just disease causing lineages [11, 179]. Based on this work there is evidence that the meningococcal pan genome is open, with novel loci being identified with every new genome sequenced. A lineage contains a subset of of the meningococcal core genome as well as lineage specific genes that defines and drives the evolution of the population. Within Lineage 3 this included the identification of 50 new meningococcal conserved hypothetical protiens and defining the source of the extream varation in the *tbpB* gene; mainly, the presence of both isotype I and II within the lineage. These two pieces of information further support the evolutionary differentiation of Lineage 3. In this work lineage specific core genome and the accessory genes were defined using the gene-by-gene annoation approach and enabled the identification of elements forming the population structure of Lineage 3 and the implications associated with allelic variation. This included the using whole genome analysis and multiple gene subset schemes to refine the population structure and assess the reationships of related strains.

The comparisons also demonstrate that this highly diverse speices is not divided based on a strains ability to cause disease, and yet capsule nul isolates did tend to cluster together within the Lineage 3 populaiton though not exclusively. This could indicate a separate evolutionary development path for the strains without the capsule operon, but needs further corroboration across more lineages. What is more intriguing, analysis is no longer tied to model organisms and its singular set of genes [180]. The sequence definition database is theoretically infinite and was used to search for defined mobile genetic elements identified in a single Lineage 3 genome efficiently across the entire dataset. What is more, there are many questions in bacterial biology which can be adequately addressed because of the *de novo* assembly, gene-by-gene analysis approach, and population genomics; including small rRNAs and intergenic regions such as promotors, regulators, and transcriptional DNA binidng sites. Uniquely, we are at liberty to investigate a population of interest using a reference-free analysis. Which in turn has fundamentally changed the nature of genetic experimentation; it no longer requires a prior knowledge of an organism's genome sequence and offers an unbiased approach.

It has become increasingly clear that substantial amounts of bacterial DNA is infact prophage DNA in some bacterial speicies. This has changed conventional thinking from a parasite-host relationship to a two-way coevolution, resulting in an important role for (pro)phage in the evolution for bacterial genomes [181]. The identification of multiple genes associated with mobile gentic elements (MGE) such as prophage and plasmids contributed substantially to the genetic diversity, genome plasticity, and evolution of bacterial genomes [182]; the Lineage 3 core genome is conservatively 2.5% prophage and plasmid integration, but could be as much as one-tenth of the genome. These integrated loci have been idnetifed by orthologous matches to published genetic phage compliments. The Mu-like prophages found in *Neisseria* could be the best example of longterm host residence based on prophage decay [182]. The presence of and decay of prophage within the genomes indicated multiple integrations occurred, with vertical descent of some, over a large interval of time. Unfortunalty the temporal frame of the dataset was realtively short and dating the evolution of the lineage was not possible. However the BEAST analysis of the rMLST genes was able to suggest that clonal expantion has occurred more than once within Lineage 3 which could explain

the two well definded sublineages. In addition to the prophage, the identification of plasmid associated TA systems (and a few genes of unknown function) offer new additional prospects in understanding the diversification of this and other meningococcal lineages over time. It is possible that the uncategorized hypothetical genes brought in by the MGE (and are the slowest to decay) may offer some form of increase or stableized fitness of the strains. This is a new area of research for meningococcal genome research.

Commensal bacteria, including *N. meningitidis*, colonize the mucosal regions of the human oropharynx and persist for extended time periods without causing invasive disease. However, it is unclear why the meningococcus is also responsible for a hyperinvasive phenotype that incurs significant morbidity and mortality within a short period in the absence of medical intervention. Furthermore, the disease outcome is accidental; it conveys no advantage to the organism and in all cases results in the organism's death. The adolescent age group has the highest prevalence of carriage, and they are considered to be the primary reservoir for transmission of both high and low invasive strains [152].

The propensity to cause disease appears to be arising from the circulating meningococcal carriage population, and is polygenic in nature. In other words, genomic heterogeneity within the circulating population, with respect to pathogenicity, leads to outbreaks of meningococcal disease, and only a small number of carried types represent the hyperinvasive strains [106, 143, 183]. To investigate the phenomenon of the carriage population more thourouly, large-scale whole genome studies are needed, using the collections of isolates such as those available from the UKMenCar survey. Much like the Meningitis Research Foundation – Meningococcal Genome Library has been interrogated for analysis the epidemiology of age-associated meningococcal lineages, a large collection of carriage associated isolate and their genomes are indispensable [184]. These studies are important in understanding the biology and overall population structure of *N. meningitidis*.

While the pathogenic strains cause a worldwide burden of infection, it is important to understand that the genetic basis for these strains their epidemiology and population structure can be, and are, more complex. Importantly for the study of *Neisseria meningitidis*, pathogenic strains cannot be fully understood in isolation from their greater circulating carriage populations, or even arguably their ecological context. By redressing the lack of carried genomes we can begin to examine the complexity of how the two phenotypes are related, use that knowledge to inform wider public health purposes, and study the origins of the species in an interconnected manner.

Appendices

Appendix A. Lineage 3 isolates including provenance and clinical data, Chapter Four

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
409	BZ 198	The Netherlands	1986	invasive		B	7-2	4	F1-5
421	NG H15	Norway	1988	carrier	throat swab	B	19	15-2	F1-5
427	NG E30	Norway	1988	carrier	throat swab	B	21	16	F1-7
428	NG H36	Norway	1988	carrier	throat swab	B	5-1	2-2	F1-7
646	400	Austria	1991	invasive		B	7-2	13-2	F1-5
647	AK50	Greece	1992	invasive		B	7-2	4	F1-5
649	50/94	Norway	1994	invasive		B	7-2	4	F1-5
650	M40/94	Chile	1994	invasive		B	7-2	4	F1-5
651	931905	The Netherlands	1993	invasive		B	7-2	4	F1-5
652	N45/96	Norway	1996	invasive		B	7-2	4	F1-5
653	91/40	New Zealand	1991	invasive	CSF	B	7-2	4	F1-5
654	88/03415	UK	1988	invasive		B	7-2	4	F1-5
887	OX9930357	UK	1999	carrier	throat swab	NG	19	15-1	F1-98
898	OX9930715	UK	1999	carrier	throat swab	NG	19	15-1	F1-5
918	OX9932099	UK	1999	carrier	throat swab	B	7-1	1	F3-3
975	0069/93	Czech Republic	1993	carrier	throat swab	NG	22	14-4	F1-7
977	0071/93	Czech Republic	1993	carrier	throat swab	C	21	16	F1-7
991	0091/93	Czech Republic	1993	carrier	throat swab	B	5-1	2-2	F1-7
1020	BM55	Greece	1996	invasive	CSF	B	17	9	F1-7
1057	BM95	Greece	1998	invasive	CSF	B	7-2	13	F1-88
1058	BM95a	Greece	1998	carrier	throat swab	B	7-2	13	F1-88
1059	BM95b	Greece	1998	carrier	throat swab	B	7-2	13	F1-88
1203	BM55a	Greece	1996	carrier	throat swab	B	17	9	F1-7
1585	0117/93	Czech Republic	1993	carrier	throat swab	B	22	26	F1-7
1588	0120/93	Czech Republic	1993	carrier	throat swab	B	22	14-6	F1-7
1604	0214/93	Czech Republic	1993	carrier	throat swab	NG	22	14-4	F1-7
1605	0215/93	Czech Republic	1993	carrier	throat swab	NG	19	15	F1-7
1606	0216/93	Czech Republic	1993	carrier	throat swab	B	19	15	F1-7
1607	0217/93	Czech Republic	1993	carrier	throat swab	B	19	15	F1-7
1615	0225/93	Czech Republic	1993	carrier	throat swab	B	19	15	F1-7
1616	0226/93	Czech Republic	1993	carrier	throat swab	B	19	15	F1-7
1617	0227/93	Czech Republic	1993	carrier	throat swab	B	19	15	F1-7
1618	0228/93	Czech Republic	1993	carrier	throat swab	B	19	15	F1-7
1633	0244/93	Czech Republic	1993	carrier	throat swab	B	22	14-4	F1-7
1636	0248/93	Czech Republic	1993	carrier	throat swab	C	18-1	34	F1-9

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
1640	0253/93	Czech Republic	1993	carrier	throat swab	B	5-1	10-10	F1-7
1642	0255/93	Czech Republic	1993	carrier	throat swab	B	5-1	10-10	F1-7
1645	0258/93	Czech Republic	1993	carrier	throat swab	B	19	15	F1-5
1656	0280/93	Czech Republic	1993	carrier	throat swab	B	21	4	F5-2
1955	0011/93	Czech Republic	1993	invasive	CSF	B	22	15	F4-67
1958	0307/93	Czech Republic	1993	carrier	throat swab	NG	22	14-4	F1-7
2813	OX0050004	UK	2000	carrier	throat swab	NG	21	16-5	F4-12
2818	OX0050408	UK	2000	carrier	throat swab	B	7-4	16	F1-7
2819	OX0050459	UK	2000	carrier	throat swab	B	7-4	16	F3-9
2914	OX0051332	UK	2000	carrier	throat swab	NG	18-1	3	F1-5
3193	OX0051350	UK	2000	carrier	throat swab	NG	7-2	4	F1-5
3395	OX9930681	UK	1999	carrier	throat swab	B	7-2	4	F5-8
3412	OX9931676	UK	1999	carrier	throat swab	NG	7-1	1	
3422	OX9932142	UK	1999	carrier	throat swab	NG	17-1	23	F1-5
3479	OX01060163	UK	2001	carrier	throat swab	B	17-1	23	F1-5
3490	OX01062124	UK	2001	carrier	throat swab	B	7-2	4	F5-5
3664	OX01061649	UK	2001	carrier	throat swab	B	7	2-1	F1-5
4193	OX9931563	UK	1999	carrier	throat swab	C	5	2	F1-20
4200	OX9931019	UK	1999	carrier	throat swab		7-2	4	F1-5
4201	OX9931436	UK	1999	carrier	throat swab	B	7-2	4	F1-5
4204	OX01060886	UK	2001	carrier	throat swab	B	22	26	F1-7
4314	OX9930141	UK	1999	carrier	throat swab	B	19	15	F1-7
4315	OX9932077	UK	1999	carrier	throat swab	B	19	15	F4-1
4395	OX01061972	UK	2001	carrier	throat swab	NG	7-2	4	F3-6
4398	OX01061086	UK	2001	carrier	throat swab	NG	22	14-6	F1-5
8139	0320/93	Czech Republic	1993	carrier	throat swab	B	19	15	F1-7
8144	0325/93	Czech Republic	1993	carrier	throat swab	B	17	16-3	F1-20
8156	0342/93	Czech Republic	1993	carrier	throat swab	B	17	16-3	F1-20
8164	0404/93	Czech Republic	1993	carrier	throat swab	B	17	16-3	F1-20
8168	0409/93	Czech Republic	1993	carrier	throat swab	NG	7-4	1	F1-7
8171	0412/93	Czech Republic	1993	carrier	throat swab	NG	17	16-3	F1-20
8173	0414/93	Czech Republic	1993	carrier	throat swab	B	17	16-3	F1-20
8176	0421/93	Czech Republic	1993	carrier	throat swab	NG	17	16-3	F1-20
10705	N 25/99	Norway	1999	invasive		B	7-1	1	F1-7
10735	N 56/99	Norway	1999	invasive		B	7-2	4	F1-5
10756	8/00	Norway	2000	invasive		B	19	15-1	F1-5

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
10758	10/00	Norway	2000	invasive		B	20	9	F1-7
10812	64/00	Norway	2000	invasive		B	22	4-1	F1-7
10830	MK 66/96	Norway	1996	carrier	throat swab	B	21	16	F1-7
10837	MK 75/96	Norway	1996	carrier	throat swab	B	19	15	F1-7
10847	MK 87/96	Norway	1996	carrier	throat swab	NG	19	15-1	F1-5
10848	MK 89/96	Norway	1996	carrier	throat swab	NG	19-5	15-3	F1-5
10913	16705	Norway	1991	carrier	throat swab	NG	19	15-1	F1-144
10914	16044	Norway	1991	carrier	throat swab	B	19-5	15-3	F1-5
10932	16328	Norway	1991	carrier	throat swab	Z	19	15	F1-7
14597	OX9932412	UK	1999	carrier	throat swab	NG	7	15-28	F2-9
14745	OX9930304	UK	1999	carrier	throat swab	B	22	14-6	F3-1
14746	OX9930342	UK	1999	carrier	throat swab	B	19	15	F1-7
14923	OX9932088	UK	1999	carrier	throat swab	B	21	16	F1-5
18969	M10 240478	UK	2010	invasive		B	18-1	3	F1-5
19024	M10 240480	UK	2010	invasive		B	7-2	4	F1-5
19026	M10 240482	UK	2010	invasive		B	7-2	4	F5-1
19027	M10 240484	UK	2010	invasive		B	17-1	23	F1-5
19032	M10 240498	UK	2010	invasive		B	7-2	4	F1-5
19263	NZ-05/33	New Zealand	2005	invasive			7-2	4	F1-5
19266	M01-240149	UK	2001	invasive			7-2	4	F1-5
19360	M0579	USA	1993	invasive			5	2	F1-5
19960	M10 240500	UK	2010	invasive		B	7-2	4	F1-5
19961	M10 240502	UK	2010	invasive		B	7-2	4	F1-5
19967	M10 240512	UK	2010	invasive		B	7-1	1	F1-5
19971	M10 240521	UK	2010	invasive		B	18-1	30-6	F1-15
19972	M10 240522	UK	2010	invasive		B	12-1	9	F1-5
19978	M10 240532	UK	2010	invasive		B	19	15	F1-5
19984	M10 240548	UK	2010	invasive		B	7-2	4	F1-5
19985	M10 240549	UK	2010	invasive		B	7-2	4	F5-5
19988	M10 240556	UK	2010	invasive		B	7-2	4	F5-8
19989	M10 240558	UK	2010	invasive		B	7-2	4	F1-5
19994	M10 240570	UK	2010	invasive		B	7-2	4	F1-5
20002	M10 240585	UK	2010	invasive		B	7-1	1	F1-5
20003	M10 240586	UK	2010	invasive		B	12-1	16	F1-5
20004	M10 240587	UK	2010	invasive		B	17	16-3	F5-13
20005	M10 240589	UK	2010	invasive		B	7-2	13-9	F1-5
20008	M10 240592	UK	2010	invasive		B	7-2	13-2	F4-1
20010	M10 240597	UK	2010	invasive		B	7-2	4	F1-5
20011	M10 240598	UK	2010	invasive		B	7-2	4	F1-88
20013	M10 240604	UK	2010	invasive		B	18-1	14	F1-5
20014	M10 240605	UK	2010	invasive		B	7-1	1	F1-5
20034	M10 240634	UK	2010	invasive		B	7-2	4	F1-5
20037	M10 240639	UK	2010	invasive		B	19	15	F5-5
20049	M10 240659	UK	2010	invasive		B	7-1	1	F1-14
20051	M10 240662	UK	2010	invasive		B	22	14	F1-5
20055	M10 240669	UK	2010	invasive		B	7-2	4	F1-5
20058	M10 240674	UK	2010	invasive		B	7-2	4	F1-5
20062	M10 240678	UK	2010	invasive		B	7-2	4	F1-15

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
20063	M10 240681	UK	2010	invasive		B	7-2	13-2	F1-5
20069	M10 240689	UK	2010	invasive		B	7-2	4	F1-5
20070	M10 240690	UK	2010	invasive		B	7-2	4	F1-5
20076	M10 240702	UK	2010	invasive		B	7-2	4	F1-5
20080	M10 240707	UK	2010	invasive		B	7-2	4	F1-5
20091	M10 240721	UK	2010	invasive		B	5-1	2-2	F5-2
20092	M10 240722	UK	2010	invasive		B	7-2	4	F1-5
20093	M10 240723	UK	2010	invasive		B	7-2	4	F1-5
20094	M10 240724	UK	2010	invasive		B	18-1	3	F1-5
20097	M10 240730	UK	2010	invasive		B	7-2	4	F1-5
20105	M10 240746	UK	2010	invasive		B	22	14	F1-5
20106	M10 240747	UK	2010	invasive		B	17-1	23	F3-9
20107	M10 240748	UK	2010	invasive		B	12-1	16	F1-5
20108	M10 240749	UK	2010	invasive		B	12-1	16	F1-5
20109	M10 240750	UK	2010	invasive		B	7-2	4	F1-5
20113	M10 240754	UK	2010	invasive		B	7-2	4	F1-5
20116	M10 240761	UK	2010	invasive		B	7-2	4	F3-4
20117	M10 240762	UK	2010	invasive		B	7-2	4	F1-5
20125	M10 240773	UK	2010	invasive		B	22	14	F1-5
20127	M10 240775	UK	2010	invasive		B	12-1	9	F1-5
20134	M10 240785	UK	2010	invasive		B	7-2	4	F1-5
20137	M10 240788	UK	2010	invasive		B	7-2	4	F1-5
20140	M10 240793	UK	2010	invasive		B	7-2	13-2	F4-1
20142	M10 240796	UK	2010	invasive		B	19	15	F3-7
20145	M10 240803	UK	2010	invasive		B	7-2	4	F1-5
20148	M10 240807	UK	2010	invasive		B	12-1	9	F1-5
20150	M10 240809	UK	2010	invasive		B	7-2	4	F1-5
20157	M10 240820	UK	2010	invasive		B	7-2	4	F3-3
20160	M10 240823	UK	2010	invasive		B	17-1	23-6	F1-5
20167	M11 240003	UK	2011	invasive		B	17-1	23	F3-3
20172	M11 240008	UK	2011	invasive		B	7-2	4	F1-5
20174	M11 240011	UK	2011	invasive		B	7-2	4	F1-5
20175	M11 240012	UK	2011	invasive		B	7-2	4	F1-5
20180	M11 240017	UK	2011	invasive		B	7-2	4	F1-5
20181	M11 240018	UK	2011	invasive		B	18-1	30	F5-8
20185	M11 240023	UK	2011	invasive		B	17-1	23-2	F1-18
20186	M11 240024	UK	2011	invasive		B	17-1	23	F5-2
20195	M11 240034	UK	2011	invasive		B	7-2	4	F1-5
20201	M11 240040	UK	2011	invasive		B	18-1	3	F1-18
20202	M11 240041	UK	2011	invasive		B	7-2	4	F1-5
20208	M11 240047	UK	2011	invasive		B	7-2	4	F1-5
20212	M11 240053	UK	2011	invasive		B	7-2	4	F1-5
20213	M11 240054	UK	2011	invasive		B	7-2	4	F1-5
20214	M11 240055	UK	2011	invasive		B	17-1	23-2	F1-18
20218	M11 240059	UK	2011	invasive		B	7-2	4	F1-5
20222	M11 240063	UK	2011	invasive		B	7-2	4	F1-5
20236	M11 240079	UK	2011	invasive		B	7-2	4	F1-5
20237	M11 240081	UK	2011	invasive		B	18-1	14	F1-5
20241	M11 240085	UK	2011	invasive		B	7-2	4	F1-5
20243	M11 240088	UK	2011	invasive		B	12-1	16	F1-5
20246	M11 240097	UK	2011	invasive		B	7-2	4	F1-5
20249	M11 240108	UK	2011	invasive		B	7-4	1	F1-5
20251	M11 240110	UK	2011	invasive		B	7-2	4	F1-5
20253	M11 240112	UK	2011	invasive		B	7-2	13-9	F1-25
20255	M11 240115	UK	2011	invasive		B	7-2	4	F1-5

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
20258	M11 240119	UK	2011	invasive		B	22-1	14	F5-2
20263	M11 240125	UK	2011	invasive		B	7-2	4	F1-47
20273	M11 240145	UK	2011	invasive		B	21	16	F1-7
20276	M11 240148	UK	2011	invasive		B	7-2	4	F1-5
20277	M11 240149	UK	2011	invasive		B	7-2	4	F1-5
20280	M11 240152	UK	2011	invasive		B	7-2	4	F4-3
20283	M11 240163	UK	2011	invasive		B	17-1	23	F5-1
20284	M11 240164	UK	2011	invasive		B	7-2	4	F1-5
20289	M11 240170	UK	2011	invasive		B	7-2	4	F1-5
20290	M11 240173	UK	2011	invasive		B	7-2	4	F1-5
20299	M11 240188	UK	2011	invasive		B	22	14	F1-5
20301	M11 240191	UK	2011	invasive		B	5-1	10-4	F1-5
20303	M11 240193	UK	2011	invasive		B	17-1	23	F1-5
20304	M11 240195	UK	2011	invasive		B	7-2	4	F3-5
20305	M11 240203	UK	2011	invasive		B	7-2	4	F1-5
20307	M11 240207	UK	2011	invasive		C	21	16	F1-15
20312	M11 240213	UK	2011	invasive		B	18-1	14	F1-5
20324	M11 240241	UK	2011	invasive		B	18-1	3	F1-5
20325	M11 240242	UK	2011	invasive		B	18-1	3	F1-5
20327	M11 240244	UK	2011	invasive		B	7-2	13-2	F1-5
20330	M11 240248	UK	2011	invasive		B	18-1	3	F1-5
20333	M11 240255	UK	2011	invasive		B	7-2	4	F1-5
20334	M11 240256	UK	2011	invasive		B	7-2	4	F1-5
20336	M11 240259	UK	2011	invasive		B	7-2	4	F1-5
20343	M11 240269	UK	2011	invasive		B	7-2	4	F1-5
20345	M11 240276	UK	2011	invasive		B	7-2	4	F3-3
20356	M11 240289	UK	2011	invasive		B	7-2	4	F1-5
20358	M11 240291	UK	2011	invasive		B	7-1	1	F1-5
20360	M11 240295	UK	2011	invasive		B	22	14	F1-5
20364	M11 240299	UK	2011	invasive		B	7-2	4	F1-5
20369	M11 240306	UK	2011	invasive		B	7-2	4	F1-5
20371	M11 240311	UK	2011	invasive		B	7-2	4	F1-5
20386	M11 240333	UK	2011	invasive		B	7-2	15	F2-9
20392	M11 240341	UK	2011	invasive		B	12-1	16	F1-5
20394	M11 240343	UK	2011	invasive		B	7-2	4	F1-5
20400	M11 240350	UK	2011	invasive		B	7-2	4	F1-7
20405	M11 240360	UK	2011	invasive		B	19	15	F1-5
20407	M11 240364	UK	2011	invasive		B	7-2	4	F1-18
20408	M11 240365	UK	2011	invasive		B	7-2	4	F1-5
20410	M11 240367	UK	2011	invasive		B	7-1	1	F1-2
20412	M11 240369	UK	2011	invasive		B	7-2	4	F1-5
20415	M11 240373	UK	2011	invasive		B	5-2	10-2	F1-19
20419	M11 240382	UK	2011	invasive		B	12-1	16	F1-5
20421	M11 240386	UK	2011	invasive		B	22	14	F1-5
20426	M11 240391	UK	2011	invasive		B	7-2	4	F1-5
20432	M11 240398	UK	2011	invasive		B	7-2	4	F5-5
20443	M11 240414	UK	2011	invasive		B	7-2	4	F5-1
20445	M11 240420	UK	2011	invasive		B	7-2	4	F1-5
20447	M11 240424	UK	2011	invasive		B	17	16-3	F4-1
20448	M11 240425	UK	2011	invasive		B	22	14	F1-5
20453	M11 240434	UK	2011	invasive		B	18-1	14	F1-5
20470	M11 240118	UK	2011	invasive		B	7-2	4	F1-5
20738	M09 240039	UK	2009	invasive	blood		18-1	3	F1-5
20742	M09 240078	UK	2009	invasive	blood		7-2	4	F1-5
20748	M09 240139	UK	2009	invasive	blood		7-2	4	F1-5

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
20755	M09 240256	UK	2009	invasive	blood		7-2	4	F1-5
20757	M09 240281	UK	2009	invasive	blood		7-2	4	F5-9
20758	M09 240283	UK	2009	invasive	blood		22	14	F5-2
20760	M09 240330	UK	2009	invasive	blood		7-2	4	F1-18
20762	M09 240439	UK	2009	invasive	blood		7-2	4	F1-5
20764	M09 240463	UK	2009	invasive	blood		7-2	4	F1-5
20766	M09 240506	UK	2009	invasive	blood		21	16	F1-15
20770	M09 240669	UK	2009	invasive	blood		22	14	F1-5
20773	M09 240831	UK	2009	invasive	blood		19	15	F1-5
20781	M10 240050	UK	2010	invasive	blood		7-2	4	F1-5
20782	M10 240052	UK	2010	invasive	throat swab		7-2	4	F1-5
20785	M10 240080	UK	2010	invasive	other		12-1	13	F1-5
20788	M10 240115	UK	2010	invasive	blood		21	16	F1-5
20792	M10 240137	UK	2010	invasive	blood		7-2	4	F5-1
20795	M10 240168	UK	2010	invasive	CSF		7-2	4	F1-15
20798	M10 240184	UK	2010	invasive	blood		7-2	4	F1-5
20800	M10 240219	UK	2010	invasive	blood		7-2	4	F1-5
20802	M10 240300	UK	2010	invasive	blood		7-2	4	F1-5
20805	M09 240174	UK	2009	invasive	blood		7-2	4	F1-5
20809	M10 240015	UK	2010	invasive	blood		7-2	4	F1-5
20884	M00 240613	UK	2000	invasive	blood	B	7-2	4	F1-5
20885	M00 240614	UK	2000		throat swab	NG	7-2	4	F1-5
20886	M01 240245	UK	2001		throat swab	NG	7-2	4	F1-5
20887	M01 240258	UK	2001	invasive	blood	B	7-2	4	F1-5
20905	M99 240945	UK	1999	invasive	CSF	B	7-2	4	F1-5
20906	M99 240975	UK	1999		throat swab	NG	7-2	4	F1-5
20907	M99 241794	Ireland	1999	invasive	blood	B	7-2	4	F1-5
20908	M99 241795	Ireland	1999		throat swab	NG	7-2	4	F1-5
20911	M05 240240	UK	2005		throat swab	B	7	9	F5-5
20915	M01 241920	UK	2001	invasive	blood	B	7-2	4	F1-5
20916	M01 241930	UK	2001		throat swab	B	7-2	4	F1-5
20941	M05 240218	UK	2005	invasive	blood	B	7	9	F5-5
20944	M05 240890	UK	2005	invasive	blood	B	7-1	1	F1-5
20945	M05 240897	UK	2005		throat swab	B	7-1	1	F1-5
20963	M07 240711	UK	2007	invasive	blood	B	7-2	4	F1-5
20964	M07 240712	UK	2007		throat swab	B	7-2	4	F1-5
20965	M07 241111	UK	2007	invasive	blood	B	7-2	4	F1-5
20966	M07 241112	UK	2007		throat swab	B	7-2	4	F1-5
21095	M11 240447	UK	2011	invasive		B	18-1	30	F1-5
21098	M11 240451	UK	2011	invasive		B	18-1	14	F1-5
21102	M11 240457	UK	2011	invasive		B	19	15	F1-5
21113	M11 240472	UK	2011	invasive		B	7-1	1	F5-8
21114	M11 240473	UK	2011	invasive		B	12-1	16	F1-5
21115	M11 240474	UK	2011	invasive		C	21-2	28	F1-5
21117	M11 240476	UK	2011	invasive		B	17-1	23	F1-5
21122	M11 240485	UK	2011	invasive		B	7-2	4	F1-5
21127	M11 240491	UK	2011	invasive		B	7-2	4	F1-5
21128	M11 240492	UK	2011	invasive		B	7-2	4	F5-5
21129	M11 240493	UK	2011	invasive		B	7-2	4	F1-5

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
21137	M11 240506	UK	2011	invasive		B	7-2	4	F1-5
21142	M11 240592	UK	2011	invasive		B	12-1	9	F1-5
21144	M11 240594	UK	2011	invasive		B	12-1	9	F1-5
21149	M11 240600	Malta	2011	invasive		B	7-2	4	F1-5
21152	M11 240707	UK	2011	invasive		B	17-1	23	F1-5
21154	M11 240712	UK	2011	invasive		B	22	14-17	F1-7
21159	M11 240721	UK	2011	invasive		B	7-2	4	F1-5
21161	M11 240724	UK	2011	invasive		B	7-2	4	F1-5
21165	M11 240728	UK	2011	invasive		B	17-1	23	F3-3
21171	M11 240736	UK	2011	invasive		B	17-1	23-2	F5-2
21184	M11 240766	UK	2011	invasive		B	22	14-6	F1-5
21188	M11 240773	UK	2011	invasive		B	7-2	4	F1-5
21195	M11 240783	UK	2011	invasive		B	5-2	10-1	F4-1
21200	M11 240790	UK	2011	invasive		B	17	16-3	F3-9
21210	M11 240945	UK	2011	invasive		B	7-2	4	F1-5
21213	M11 240949	UK	2011	invasive		B	7-2	4	F1-5
21222	M11 240981	UK	2011	invasive		B	7-2	15	F5-5
21230	M11 240992	UK	2011	invasive		B	19	15	F1-5
21234	M11 241013	UK	2011	invasive		B	7-2	4	F1-5
21235	M11 241014	UK	2011	invasive		B	7-2	4	F1-5
21236	M11 241015	UK	2011	invasive		B	12-1	9	F1-5
21238	M11 241018	UK	2011	invasive		B	18-1	14	F1-5
21245	M11 241028	UK	2011	invasive		B	12-1	9	F1-5
21247	M11 241032	UK	2011	invasive		B	12-1	16	F1-5
21254	M11 241040	UK	2011	invasive		B	7-2	4	F1-7
21257	M11 241044	UK	2011	invasive		B	18-1	14	F1-5
21258	M11 241046	UK	2011	invasive		B	22	14	F1-5
21261	M11 241050	UK	2011	invasive		B	18-1	3	F1-5
21263	M11 241054	UK	2011	invasive		B	22	14	F1-5
21264	M11 241055	UK	2011	invasive		B	7-2	4	F1-5
21270	M11 241063	UK	2011	invasive		B	7-2	4	F1-5
21273	M11 241066	UK	2011	invasive		B	7-2	4	F1-5
21277	M11 241072	UK	2011	invasive		B	7-2	4	F1-5
21278	M11 241073	UK	2011	invasive		B	12-1	9	F1-5
21286	M12 240002	UK	2012	invasive		B	7-2	4	F1-5
21291	M12 240009	UK	2012	invasive		B	7-2	4	F1-5
21292	M12 240010	UK	2012	invasive		B	7-2	4	F1-18
21296	M12 240014	UK	2012	invasive		C	18-7	9	F5-8
21299	M12 240018	UK	2012	invasive		B	7-2	4	F1-5
21301	M12 240020	UK	2012	invasive		B	7-2	4	F5-1
21310	M12 240032	UK	2012	invasive		B	7-2	4	F1-5
21313	M12 240035	UK	2012	invasive		B	18-1	14	F1-5
21314	M12 240036	UK	2012	invasive		B	7-2	4	F1-5
21316	M12 240039	UK	2012	invasive		B	7-2	4	F1-5
21318	M12 240041	UK	2012	invasive		B	7-2	4	F5-1
21319	M12 240042	UK	2012	invasive		B	18-1	3	F1-5
21321	M12 240045	UK	2012	invasive		B	7-2	4	F1-5
21337	M12 240071	UK	2012	invasive		B	7	30	F1-5
21340	M12 240077	UK	2012	invasive		B	7-2	4	F1-5
21341	M12 240078	UK	2012	invasive		B	7-2	4	F1-5
21344	M12 240081	UK	2012	invasive		B	7-2	4	F1-5
21347	M12 240085	UK	2012	invasive		B	7-2	4	F1-5
21353	M12 240094	UK	2012	invasive		B	12-1	9	F1-5
21355	M12 240097	UK	2012	invasive		B	7-2	4	F1-5
21360	M12 240103	UK	2012	invasive		B	18-1	34	F1-5

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
21368	M12 240117	UK	2012	invasive		B	7-2	4	F1-5
21369	M12 240118	UK	2012	invasive		B	7-2	4	F1-5
21374	M12 240124	UK	2012	invasive		B	7-2	4	F1-5
21376	M12 240126	UK	2012	invasive		B			F1-5
21378	M12 240128	UK	2012	invasive		B	7-1	1	F1-5
21379	M12 240131	UK	2012	invasive		NG	22	14	F1-5
21382	M12 240134	UK	2012	invasive		B	7-2	4	F1-5
21384	M12 240138	UK	2012	invasive		B	17-1	23	F1-5
21394	M12 240155	UK	2012	invasive		B	7-2	4	F1-5
21398	M12 240161	UK	2012	invasive		B	7-1	1	F1-5
21400	M12 240167	UK	2012	invasive		B	7-2	4	F1-5
21401	M12 240168	UK	2012	invasive		B	19	15-1	F1-5
21402	M12 240169	UK	2012	invasive		NG	18	25-11	F1-84
21411	M12 240186	UK	2012	invasive		B	7-2	13-2	F4-1
21412	M12 240187	UK	2012	invasive		B	5-1	10-7	F1-5
21414	M12 240191	UK	2012	invasive		B	18-1	34	F1-5
21423	M12 240207	UK	2012	invasive		B	7-2	4	F1-5
21426	M12 240213	UK	2012	invasive		B	12-1	9	F1-5
21431	M12 240220	UK	2012	invasive		B	12-1	16	F1-5
21434	M12 240224	UK	2012	invasive		B	7-2	4	F1-5
21437	M12 240227	UK	2012	invasive		B	12-1	16	F1-5
21450	M12 240245	UK	2012	invasive		B	7-2	4	F1-5
21452	M12 240249	UK	2012	invasive		B	7-2	4	F3-3
21455	M12 240252	UK	2012	invasive		B	7-2	4	F5-8
21460	M12 240257	UK	2012	invasive		B	18-7	9	F1-7
21463	M12 240263	UK	2012	invasive		B	7-2	4	F1-5
21464	M12 240264	UK	2012	invasive		B	7-2	4	F1-5
21468	M12 240274	UK	2012	invasive		B	5-2	10-4	F3-3
21470	M12 240284	UK	2012	invasive		B	18-1	14	F1-5
21474	M12 240290	UK	2012	invasive		B	7-2	4	F1-5
21478	M12 240296	UK	2012	invasive		B	18-7	9	F3-7
21479	M12 240299	UK	2012	invasive		B	7-2	4	F1-47
21481	M12 240301	UK	2012	invasive		B	18-1	3	F1-5
21482	M12 240302	UK	2012	invasive		B	22	14-6	F1-15
21484	M12 240305	UK	2012	invasive		B	7-2	4	F1-5
21487	M12 240308	UK	2012	invasive		B	22	14	F1-5
21488	M12 240309	UK	2012	invasive		B	7-2	4	F3-9
21490	M12 240314	UK	2012	invasive		B	7-2	4	F1-5
21491	M12 240315	UK	2012	invasive		B	7-2	4	F1-5
21494	M12 240319	UK	2012	invasive		B	22	14	F1-5
21496	M12 240321	UK	2012	invasive		B	7-2	4	F1-5
21500	M12 240325	UK	2012	invasive		B	7-2	4	F1-5
21506	M12 240333	UK	2012	invasive		B	7-2	4	F1-5
21507	M12 240334	UK	2012	invasive		B	7-2	4	F1-5
26254	BB123	UK	2012	carrier	throat swab		19	15	F1-5
26268	BB77	UK	2012	carrier	throat swab		17-1	23	F1-5
26288	BB89	UK	2012	carrier	throat swab		17	16-3	F5-5
26293	BB238	UK	2012	carrier	throat swab		7-2	2	F1-5
26298	z77	UK	2012	carrier	throat swab		17-1	23	F1-5
26311	V57	UK	2011	carrier	throat swab		18	25-1	F1-5
26331	BB44	UK	2012	carrier	throat swab		17-1	23	F1-5

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
26333	R176	UK	2011	carrier	throat swab		12-1	9-20	F1-5
26353	BB64	UK	2012	carrier	throat swab		21	16	F1-15
26378	BB291	UK	2012	carrier	throat swab		7-2	4	F1-5
26433	T57	UK	2011	carrier	throat swab		18	25-1	F1-5
26463	R291	UK	2011	carrier	throat swab		7-2	4	F1-5
26469	BB52	UK	2012	carrier	throat swab		18	25-10	
26484	R89	UK	2011	carrier	throat swab		17	16-3	F5-5
26485	R304	UK	2011	carrier	throat swab		7-2	4	F1-5
26490	z64	UK	2012	carrier	throat swab		21	16	F1-15
26492	T291	UK	2011	carrier	throat swab		7-2	4	F1-5
26823	12024_2010	Ireland	2010	invasive	blood	B	22	14	F3-7
26831	12032_2010	Ireland	2010	invasive	blood	B	7-2	4	F3-6
26832	12033_2010	Ireland	2010	invasive	CSF	B	7-2	4	F1-5
26833	12034_2010	Ireland	2010	invasive	sputum	B	7-2	4	F5-12
26834	12035_2010	Ireland	2010	invasive	other	B	7-2	4	F1-21
26836	12003_2011	Ireland	2011	invasive	other	B	7-2	4	F1-5
26837	12002_2011	Ireland	2011	invasive	blood	B	18-1	3	F1-5
26838	12004_2011	Ireland	2011	invasive	blood	B	7-2	4	F1-5
26840	12006_2011	Ireland	2011	invasive	blood	B	7-2	4	F1-5
26842	12008_2011	Ireland	2011	invasive	blood	B	7-2	4	F5-12
26843	12009_2011	Ireland	2011	invasive	blood	B	7-2	4	F1-5
26846	12017_2011	Ireland	2011	invasive	blood	B	7-2	4	F3-6
26848	12019_2011	Ireland	2011	invasive	blood	B	22	14	F1-5
26849	12020_2011	Ireland	2011	invasive	CSF	B	18-1	3	F1-5
26861	12031_2011	Ireland	2011	invasive	CSF	B	7-2	4	F1-5
26862	12032_2011	Ireland	2011	invasive	blood	B	7-2	4	F5-1
26864	12001_2012	Ireland	2012	invasive	blood	B	7-2	4	F1-5
26865	12002_2012	Ireland	2012	invasive	blood	B	7-2	4	F1-5
26866	12003_2012	Ireland	2012	invasive	blood	B	7-2	4	F5-5
26867	12004_2012	Ireland	2012	invasive	blood	B	22	14	F3-6
26875	12012_2012	Ireland	2012	invasive	joint fluid	B	7-46	13	F5-2
26884	12024_2012	Ireland	2012	invasive	blood	B	7-2	4	F5-8
26886	12026_2012	Ireland	2012	invasive	blood	B	7-2	4	F1-5
26890	12031_2012	Ireland	2012	invasive	blood	B	22	14-6	F1-5
26893	12036_2012	Ireland	2012	invasive	CSF	B	7-2	4	F1-5
26897	12001_2013	Ireland	2013	invasive	blood	B	7-2	4	F4-12
26901	12005_2013	Ireland	2013	invasive	blood	B	7-2	4	F3-9
26904	12008_2013	Ireland	2013	invasive	blood	B	18-1	3	F1-5
26909	12013_2013	Ireland	2013	invasive	blood	B	7-2	4	F1-5
26915	12019_2013	Ireland	2013	invasive	blood	B	7-2	4	F3-6
27510	20588	UK	2009			Y	18-1	34	F1-5
27778	M12 240336	UK	2012	invasive		B	7-2	4	F5-5
27783	M12 240349	UK	2012	invasive		B	17-1	23	F5-2
27786	M12 240650	UK	2012	invasive		B	22	14	F1-5
27787	M12 240651	UK	2012	invasive		B	18-1	3	F1-5
27788	M12 240652	UK	2012	invasive		B	7-2	4	F1-5
27790	M12 240654	UK	2012	invasive		B	7-2	4-37	F1-5
27791	M12 240656	UK	2012	invasive		B	18-1	3	F1-5
27792	M12 240658	UK	2012	invasive		B	7-2	4	F1-5

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
27794	M12 240664	UK	2012	invasive		B	7-2	4	F1-5
27795	M12 240667	UK	2012	invasive		B	18-1	14	F1-5
27796	M12 240668	UK	2012	invasive		B	7-2	4-15	F1-5
27798	M12 240672	UK	2012	invasive		B	5-1	10-4	F3-7
27802	M12 240681	UK	2012	invasive		B	18-7	9	F1-7
27806	M12 240685	UK	2012	invasive		B	17	16-93	F1-20
27811	M12 240697	UK	2012	invasive		B	7-2	4	F1-5
27812	M12 240699	UK	2012	invasive		B	7-2	4	F1-5
27816	M12 240706	UK	2012	invasive		B	12-1	9	F1-5
27818	M12 240712	UK	2012	invasive		B	7-2	4	F5-12
27822	M12 240720	UK	2012	invasive		B	7-2	4	F1-5
27833	M12 240741	UK	2012	invasive		B	7-2	4	F1-25
27836	M12 240750	UK	2012	invasive		B	12-1	9	F1-5
27837	M12 240752	UK	2012	invasive		B	7-2	4	F1-5
27846	M12 240783	UK	2012	invasive		B	21-7	16	F1-5
27847	M12 240785	UK	2012	invasive		B	7-2	4	F1-5
27856	M12 240807	UK	2012	invasive		B	7-2	4	F1-5
27865	M12 240825	UK	2012	invasive		B	12-1	16	F1-5
27872	M12 240846	UK	2012	invasive		B	7-2	4	F1-5
27873	M12 240847	UK	2012	invasive		B	7-2	4	F3-3
27875	M12 240849	UK	2012	invasive		B	7-2	4	F1-5
27876	M12 240850	UK	2012	invasive		B	7-2	4	F1-5
27885	M12 240868	UK	2012	invasive		B	7-2	4	F1-5
27886	M12 240872	UK	2012	invasive		B	21	16	F1-15
27889	M12 240876	UK	2012	invasive		B	19	15-1	F1-24
27895	M12 240887	UK	2012	invasive		B	7-2	4	F1-5
27900	M12 240901	UK	2012	invasive		B	19	15-1	F5-15
27901	M12 240902	UK	2012	invasive		B	19	16	F1-5
27903	M13 240001	UK	2013	invasive		B	7-2	4	F1-5
27908	M13 240007	UK	2013	invasive		B	18-1	3	F1-5
27909	M13 240009	UK	2013	invasive		B	7-2	4	F1-5
27910	M13 240012	UK	2013	invasive		B	12-1	9	F1-5
27913	M13 240019	UK	2013	invasive		B	18	25	F1-5
27918	M13 240031	UK	2013	invasive		B	18-3	1	F5-2
27924	M13 240041	UK	2013	invasive		B	7-2	4	F1-5
27929	M13 240047	UK	2013	invasive		B	22	14-6	F1-5
27943	M13 240084	UK	2013	invasive		B	7-2	4	F5-8
27944	M13 240085	UK	2013	invasive		B	7	30	F1-5
27946	M13 240088	UK	2013	invasive		B	7-2	4	F1-5
27947	M13 240090	UK	2013	invasive		B	7-2	4	F1-5
27950	M13 240097	UK	2013	invasive		B	17-1	23-2	F3-7
27951	M13 240098	UK	2013	invasive		B	19	15	F1-5
27954	M13 240111	UK	2013	invasive		B	18-1	14	F1-5
27956	M13 240115	UK	2013	invasive		B	18-4	25	F1-5
27958	M13 240120	UK	2013	invasive		B	7-2	4	F1-5
27960	M13 240123	UK	2013	invasive		B	7-2	4-15	F1-5
27966	M13 240135	UK	2013	invasive		B	18-4	25	F1-5
27967	M13 240142	UK	2013	invasive		B	17	16-3	F1-20
27971	M13 240148	UK	2013	invasive		B	12-1	13-1	F1-17
27974	M13 240157	UK	2013	invasive		B	7-2	4	F5-8
27978	M13 240166	UK	2013	invasive		B	18-1	14	F1-5
27981	M13 240171	UK	2013	invasive		B	7-2	4	F1-5
27983	M13 240173	UK	2013	invasive		B	22	14-10	F1-5
27984	M13 240174	UK	2013	invasive		B	7-2	4	F5-8
27985	M13 240175	UK	2013	invasive		B	7-1	1	F1-14

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
27986	M13 240181	UK	2013	invasive		B	7-2	4	F1-5
27994	M13 240204	UK	2013	invasive		B	18-1	3	F1-5
27995	M13 240205	UK	2013	invasive		B	12-1	16	F1-5
27996	M13 240206	UK	2013	invasive		B	22	14-6	F1-5
27998	M13 240210	UK	2013	invasive		B	7-2	4	F1-5
28000	M13 240212	UK	2013	invasive		B	17	16-3	F1-7
28003	M13 240217	UK	2013	invasive		B	7-2	4	F1-5
28006	M13 240228	UK	2013	invasive		B	7-2	4	F3-3
28009	M13 240231	UK	2013	invasive		B	7-2	4	F1-5
28012	M13 240234	UK	2013	invasive		B	7-2	4	F1-5
28013	M13 240235	UK	2013	invasive		B	12-1	9	
28014	M13 240237	UK	2013	invasive		B	7-2	4	F1-5
28015	M13 240239	UK	2013	invasive		B	7-2	4	F1-5
28024	M13 240261	UK	2013	invasive		B	7-2	4	F1-5
28026	M13 240263	UK	2013	invasive		B	5-2	10-4	F3-3
28028	M13 240271	UK	2013	invasive		B	7-1	1	F1-14
28030	M13 240273	UK	2013	invasive		B	18-4	25	F1-5
28032	M13 240275	UK	2013	invasive		B	7-2	4	F1-5
28033	M13 240277	UK	2013	invasive		B	7-2	4	F1-5
28036	M13 240292	UK	2013	invasive		B	22	14-6	F1-5
28039	M13 240385	UK	2013	invasive		B	7-2	4	F1-5
28044	M13 240409	UK	2013	invasive		B	7-2	4	F1-5
28046	M13 240413	UK	2013	invasive		B	7-2	4	F1-5
28049	M13 240418	UK	2013	invasive		B	7-1	1	F1-5
28059	M13 240438	UK	2013	invasive		B	7-1	1-1	F1-7
28061	M13 240449	UK	2013	invasive		B	7-2	4	F1-18
28064	M13 240452	UK	2013	invasive		B	7-2	4	F1-5
28072	M13 240472	UK	2013	invasive		B	7-2	4	F1-5
28073	M13 240475	UK	2013	invasive		B	7-2	4	F1-7
28076	M13 240479	UK	2013	invasive		B	7-2	4	F1-5
28088	M13 240002	UK	2013	invasive		C	18-1	34	F1-5
28110	M13 240198	UK	2013	invasive		NG	5-2	10-4	F3-3
28112	M13 240399	UK	2013	invasive		NG	7	30	F1-5
28850	OX9930030	UK	1999	carrier	throat swab	E	7-2	4	F1-5
28851	OX9930099	UK	1999	carrier	throat swab	B	22	13-1	F1-5
28852	OX9930434	UK	1999	carrier	throat swab	NG	7-2	4	F1-5
28853	OX9930486	UK	1999	carrier	throat swab	B	19	15	F1-5
28854	OX9930499	UK	1999	carrier	throat swab	B	22-1	14	F1-5
28855	OX9930539	UK	1999	carrier	throat swab	B	22	14-6	F3-1
28856	OX9930579	UK	1999	carrier	throat swab	B	19	15	F1-5
28857	OX9930716	UK	1999	carrier	throat swab	NG	18-1	3	F3-1
28858	OX9930765	UK	1999	carrier	throat swab	B	17-1	23-2	F5-36
28859	OX9930826	UK	1999	carrier	throat swab	B	19	15-1	F1-5
28860	OX9930864	UK	1999	carrier	throat swab	B	7-2	13-1	F1-5
28861	OX9930906	UK	1999	carrier	throat swab	B	17	16-3	F1-20
28862	OX9930945	UK	1999	carrier	throat swab	B	18-1	3	F1-5
28863	OX9930951	UK	1999	carrier	throat swab	B	19	15	F4-1

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
28864	OX9930963	UK	1999	carrier	throat swab	B	19	15-1	F1-5
28865	OX9930970	UK	1999	carrier	throat swab	B	17-1	23	F5-36
28866	OX9930996	UK	1999	carrier	throat swab	B	17-1	23-2	F1-5
28867	OX9931018	UK	1999	carrier	throat swab	B	7-2	4	F1-5
28868	OX9931037	UK	1999	carrier	throat swab	NG	7-2	13-1	F5-17
28869	OX9931039	UK	1999	carrier	throat swab	B	7-2	4	F1-5
28870	OX9931049	UK	1999	carrier	throat swab	B	21	16-32	F1-5
28871	OX9931066	UK	1999	carrier	throat swab	NG	7-2	4	F1-5
28872	OX9931111	UK	1999	carrier	throat swab	NG	7-2	4	F1-7
28873	OX9931152	UK	1999	carrier	throat swab	B	19	15-1	F1-5
28874	OX9931165	UK	1999	carrier	throat swab	B	19	15-1	F1-5
28876	OX9931218	UK	1999	carrier	throat swab	B	21	16	F1-5
28878	OX9931242	UK	1999	carrier	throat swab	B	17-1	23	F1-5
28879	OX9931249	UK	1999	carrier	throat swab	B	17-1	23	F1-5
28880	OX9931268	UK	1999	carrier	throat swab	B	17-1	23-2	F1-5
28881	OX9931283	UK	1999	carrier	throat swab	B	17-1	23-2	F1-5
28882	OX9931352	UK	1999	carrier	throat swab	NG	22	14-3	F3-4
28883	OX9931550	UK	1999	carrier	throat swab	B	5-1	10-8	F1-5
28884	OX9931553	UK	1999	carrier	throat swab	E	19-1	26	F1-7
28885	OX9931715	UK	1999	carrier	throat swab	E	7-2	4-13	F1-5
28886	OX9931875	UK	1999	carrier	throat swab	B	19	15-1	F1-5
28887	OX9931911	UK	1999	carrier	throat swab	B	19	15-1	F3-6
28889	OX9931967	UK	1999	carrier	throat swab	NG	18-1	34	F1-5
28890	OX9931990	UK	1999	carrier	throat swab	B	19	15-1	F1-5
28891	OX9932003	UK	1999	carrier	throat swab	B	18-1	3	F1-5
28892	OX9932015	UK	1999	carrier	throat swab	B	18-1	3	F1-5
28893	OX9932020	UK	1999	carrier	throat swab	NG	7-2	4	F1-5
28894	OX9932093	UK	1999	carrier	throat swab	B	21	16	F1-5
28895	OX9932180	UK	1999	carrier	throat swab	B	17-1	23	F1-5
28898	OX9932389	UK	1999	carrier	throat swab	B	19	15	
28899	OX0050007	UK	2000	carrier	throat swab	B	17-1	23	F1-5
28900	OX0050057	UK	2000	carrier	throat swab	B	17	16-3	F5-7
28901	OX0050209	UK	2000	carrier	throat swab	NG	7-2	4-13	F1-5
28902	OX0050225	UK	2000	carrier	throat swab	B	18	25	F1-5
28903	OX0050304	UK	2000	carrier	throat swab	NG	7-2	4	F1-5

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
28904	OX0050570	UK	2000	carrier	throat swab	B	7-2	deleted	F1-5
28905	OX0050677	UK	2000	carrier	throat swab	NG	7-2	4	F1-7
28906	OX0050872	UK	2000	carrier	throat swab	B	7-2	4	F1-5
28907	OX0050873	UK	2000	carrier	throat swab	B	7-2	4	F1-5
28908	OX0050898	UK	2000	carrier	throat swab	NG	19-1	26	F1-7
28909	OX0050926	UK	2000	carrier	throat swab	NG	19-1	26	F1-7
28911	OX0050941	UK	2000	carrier	throat swab	B	17-1	23	F1-5
28912	OX0051021	UK	2000	carrier	throat swab	B	5-1	2-2	F1-28
28913	OX0051136	UK	2000	carrier	throat swab	B	19	15	F1-5
28914	OX0051387	UK	2000	carrier	throat swab	B	22	14-6	F3-1
28915	OX0051404	UK	2000	carrier	throat swab	NG	17-1	23	F1-5
28916	OX0051423	UK	2000	carrier	throat swab	NG	19	15-1	F1-98
28917	OX0052019	UK	2000	carrier	throat swab	B	7-2	4	F1-5
28919	OX0052120	UK	2000	carrier	throat swab	B	17	16-3	F4-3
28920	OX0052129	UK	2000	carrier	throat swab	NG	18-1	34	F4-24
28921	OX0052162	UK	2000	carrier	throat swab	NG	7-2	4	F1-5
28922	OX0052296	UK	2000	carrier	throat swab	B	19	15	F1-5
28923	OX0052474	UK	2000	carrier	throat swab	B	17-1	23	F1-5
28924	OX0052529	UK	2000	carrier	throat swab	B	7-2	4	F1-5
28925	OX0052958	UK	2000	carrier	throat swab	B	21	16	F1-5
28926	OX01060046	UK	2001	carrier	throat swab	B	7-2	4	F1-5
28927	OX01060062	UK	2001	carrier	throat swab	NG	7	15	F5-75
28929	OX01060279	UK	2001	carrier	throat swab	B	19	4	F5-8
28930	OX01060297	UK	2001	carrier	throat swab	B	22	14	F5-2
28932	OX01060472	UK	2001	carrier	throat swab	NG	7	30-1	F1-5
28933	OX01060561	UK	2001	carrier	throat swab	NG	7-2	4	F1-5
28934	OX01060596	UK	2001	carrier	throat swab	B	19	15	F1-7
28935	OX01060652	UK	2001	carrier	throat swab	B	17-1	23	F1-5
28936	OX01060915	UK	2001	carrier	throat swab	B	17-1	23-2	F1-5
28937	OX01060954	UK	2001	carrier	throat swab	NG	18		F3-7
28938	OX01060960	UK	2001	carrier	throat swab	NG	7-2	4	F1-5
28939	OX01061044	UK	2001	carrier	throat swab	NG	18-1	3	F1-7
28940	OX01061115	UK	2001	carrier	throat swab	NG	7-2	13-1	F5-8
28941	OX01061123	UK	2001	carrier	throat swab	NG	7-2	13-1	F5-8
28942	OX01061131	UK	2001	carrier	throat swab	B	19	15-1	F1-5

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
28944	OX01061328	UK	2001	carrier	throat swab	NG	22	14	F1-5
28945	OX01061416	UK	2001	carrier	throat swab	B	19-5	15-15	F1-5
28946	OX01061431	UK	2001	carrier	throat swab	NG	22	13-1	F1-5
28947	OX01061451	UK	2001	carrier	throat swab	NG	18-1	34	F1-5
28948	OX01061511	UK	2001	carrier	throat swab	NG	22	14-6	F1-5
28950	OX01061587	UK	2001	carrier	throat swab	B	7-2	4	F1-7
28951	OX01061635	UK	2001	carrier	throat swab	NG	18-3	1	F5-5
28952	OX01061707	UK	2001	carrier	throat swab	NG	7-2	4	F1-5
28953	OX01061748	UK	2001	carrier	throat swab	B	7-2	4	F1-5
28954	OX01061840	UK	2001	carrier	throat swab	NG	19-1	26	F1-7
28955	OX01062035	UK	2001	carrier	throat swab	B	7-2	4	F3-9
28956	OX01062098	UK	2001	carrier	throat swab	NG	18-1	3-7	F1-5
28957	OX01062156	UK	2001	carrier	throat swab	B	19	15-1	F1-5
28958	OX01062220	UK	2001	carrier	throat swab	B	7-2	4	F1-5
28959	OX01062252	UK	2001	carrier	throat swab	B	7-2	4	F1-5
28960	OX01062298	UK	2001	carrier	throat swab	NG	7-2	4	F1-5
28961	OX01062341	UK	2001	carrier	throat swab	NG	7-2	4	F1-5
28962	OX01062350	UK	2001	carrier	throat swab	B	18-1	3	F1-5
28963	OX01062357	UK	2001	carrier	throat swab	NG	7-2	4	F1-5
28964	OX01062462	UK	2001	carrier	throat swab	B	17-1	23	F1-5
28966	OX9930855	UK	1999	carrier	throat swab		7-2	13-1	F1-5
28984	IBD-611	Poland	2011	invasive	blood	B	7-2	4	F1-5
28988	IBD-615	Poland	2011	invasive	blood	B	21	16	F1-15
28997	IBD-624	Poland	2011	invasive	CSF	B	12-1	13-1	F1-5
29003	IBD-630	Poland	2011	invasive	blood	B	7-2	4	F1-5
29008	IBD-635	Poland	2011	invasive	CSF	C	17	16-4	F3-9
29013	IBD-640	Poland	2011	invasive	blood	B	5-1	10-8	F1-5
29015	IBD-642	Poland	2011	invasive	CSF	C	17	16-4	F3-9
29022	IBD-649	Poland	2011	invasive	CSF	B	21-2	28	F1-7
29028	IBD-655	Poland	2011	invasive	CSF	B	5-1	10-1	F1-37
29030	IBD-657	Poland	2011	invasive	CSF	B	19	15	F1-5
29031	IBD-658	Poland	2011	invasive	CSF	B	7-2	4	F1-5
29033	IBD-660	Poland	2011	invasive	CSF	B	7-2	14	F1-5
29046	IBD-673	Poland	2012	invasive	CSF	B	7-2	4	F1-5
29047	IBD-674	Poland	2012	invasive	blood	B	12-1	13-43	F1-5
29048	IBD-675	Poland	2012	invasive	CSF	B	18-7	9	
29054	IBD-681	Poland	2012	invasive	blood	B	7-2	4	F1-5
29057	IBD-684	Poland	2012	invasive	blood	B	7-2	4	F1-5
29058	IBD-685	Poland	2012	invasive	CSF	C	7-4	15	F5-5
29066	IBD-694	Poland	2012	invasive	blood	C	17	16-4	F3-9
29068	IBD-696	Poland	2012	invasive	CSF	B	7-1	1	F1-24
29072	IBD-700	Poland	2012	invasive	CSF	C	12-3	4	F3-5
29076	IBD-704	Poland	2012	invasive	blood	C	17	16-4	F3-4

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
29077	IBD-705	Poland	2012	invasive	CSF	C	18-1	34	F5-16
29079	IBD-707	Poland	2012	invasive	blood	C	21-2	28	F5-5
29086	IBD-714	Poland	2012	invasive	CSF	B	17	16-4	F4-1
29089	IBD-717	Poland	2012	invasive	blood	B	17	16-4	F3-9
29093	IBD-721	Poland	2012	invasive	blood	C	7-2	14	F1-7
29108	IBD-736	Poland	2012	invasive	blood	B	12-1	13-1	F1-5
29116	IBD-745	Poland	2012	invasive	CSF	B	5-2	10	F1-24
29125	IBD-754	Poland	2012	invasive	blood	C	21-2	28	F5-5
29127	IBD-756	Poland	2012	invasive	CSF	C	18-1	34	F1-7
29131	IBD-802	The Netherlands	2012	invasive	blood	B	7-2	30	F1-5
29134	IBD-805	The Netherlands	2012	invasive	blood	B	7-2	4	F1-5
29139	IBD-810	The Netherlands	2012	invasive	blood	B	7-2	13-9	F1-5
29143	IBD-814	The Netherlands	2012	invasive	blood	B	21	16	F1-15
29150	IBD-821	The Netherlands	2012	invasive	blood	B	7-2	4	F1-5
29154	IBD-825	The Netherlands	2012	invasive	joint fluid	B	7-2	4	F1-5
29165	IBD-836	The Netherlands	2011	invasive	blood	B	18	25	F1-5
29174	IBD-417	Finland	2011	invasive	blood	B	7-2	10-7	F1-23
29178	IBD-421	Finland	2011	invasive	blood	B	7-2	4	F1-5
29180	IBD-423	Finland	2011	invasive	blood	B	7-2	4	F1-5
29181	IBD-424	Finland	2011	invasive	blood	B	7-2	4	F1-5
29182	IBD-425	Finland	2011	invasive	blood	B	7-2	4	F1-5
29185	IBD-428	Finland	2011	invasive	blood	B	7-2	4	F1-5
29194	IBD-438	Finland	2012	invasive	blood	B	7-2	4	F1-5
29196	IBD-440	Finland	2012	invasive	CSF	B	7-2	4	F1-5
29198	IBD-442	Finland	2012	invasive	blood	B	7-2	4	F1-5
29207	IBD-510	Norway	2011	invasive	CSF	B	12-20	13-1	F5-7
29215	IBD-518	Norway	2011	invasive	CSF	B	7-2	4	F3-3
29219	IBD-522	Norway	2011	invasive	blood	B	7-2	13-2	F4-1
29223	IBD-526	Norway	2011	invasive	blood	B	7-2	4	F1-5
29231	IBD-534	Norway	2012	invasive		B	22	14-6	
29237	IBD-580	Poland	2011	invasive	CSF	B	5-1	10-8	F1-5
29261	IBD-604	Poland	2011	invasive	CSF	C	17	16-4	F3-9
29310	38465	South Africa	2013	invasive	blood	NG	7-1	1	F3-20
29741	M97 252240	UK	1997	invasive	CSF	B	7-2	4	F1-5
29742	M97 252241	UK	1997	invasive	blood	B	7-2	4	F1-5
29743	M97 252242	UK	1997	invasive	throat swab	B	7-2	4	F1-5
29744	M98 250082	UK	1998	invasive	CSF	B	7-2	4	F1-18
29745	M98 250087	UK	1998	invasive	blood	B	7-2	4	F1-18
29746	M98 250088	UK	1998	invasive	throat swab	B	7-2	4	F1-18
29757	M98 252086	UK	1998	invasive	CSF	B	17	16-3	F1-20
29759	M98 252088	UK	1998	invasive	throat swab	B	17	16-3	F1-20
29763	M00 240046	UK	2000	invasive	blood	B	7-2	4	F1-5
29764	M00 240163	UK	2000	invasive	CSF	B	7-2	4	F3-3
29765	M00 240207	UK	2000	invasive	nasal swab	NG	7-2	4	F3-3
29766	M00 240208	UK	2000	invasive	throat swab	B	7-2	4	F3-3
29767	M00 240858	UK	2000	invasive	nasal swab	B	7-2	4	F1-5
29768	M00 240859	UK	2000	invasive	throat	B	7-2	4	F1-5

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
					swab				
29769	M00 240860	UK	2000	invasive	blood	B	7-2	4	F1-5
29770	M00 240863	UK	2000	invasive	CSF	B	7-2	4	F1-5
29771	M00 240877	UK	2000	invasive	blood	B	7-2	4	F1-5
29772	M00 240878	UK	2000	invasive	throat swab	B	7-2	4	F1-5
29773	M00 240993	UK	2000	invasive	blood	B	7-2	4	F1-5
29774	M00 240995	UK	2000	invasive	CSF	B	7-2	4	F1-5
29800	M03 241002	UK	2003	invasive	CSF	B	7-2	4	F1-5
29802	M03 241017	UK	2003	invasive	CSF	B	7-2	4	F1-5
29823	M98 253067	UK	1998	invasive	CSF	B	7-2	4	F4-12
29824	M98 253068	UK	1998	invasive	blood	B	7-2	4	F4-12
29825	M98 253069	UK	1998	invasive	throat swab	B	7-2	4	F4-12
30045	M01 242572	UK	2001	invasive	blood	B	7-2	4	F1-5
30062	M14 240064	Turkey	2005	invasive		W	7-2	13-1	F1-5
30063	M14 240065	Turkey	2009	invasive		W	7-2	13-1	F1-5
30112	M99 240981	UK	1999	invasive	nasal swab	B	5-1	10-4	F1-5
30113	M99 240982	UK	1999	invasive	throat swab	B	5-1	10-4	F1-5
30115	M99 242419	UK	1999	invasive	blood	B	7-2	4	F5-1
30116	M99 242434	UK	1999	invasive	CSF	B	7-2	4	F5-1
30117	M99 242435	UK	1999	invasive	throat swab	B	7-2	4	F5-1
30120	M99 243964	UK	1999	invasive	CSF	B	7-2	4	F1-5
30121	M99 243965	UK	1999	invasive	nasal swab	B	7-2	4	F1-5
30277	M00 242922	UK	2000	invasive	blood	B	7-2	4	F1-5
30279	2012 107	Portugal	2012	carrier	throat swab	C	21	16	F1-15
30281	2012 167	Portugal	2012	carrier	throat swab	B	18-1	34	F1-5
30359	IBD-844	The Netherlands	2011	invasive	CSF	B	7-1	1	F5-3
30360	IBD-845	The Netherlands	2011	invasive	blood	B	7-1	1	F5-3
30362	IBD-847	The Netherlands	2011	invasive	blood	B	18	25	F1-5
30366	IBD-851	The Netherlands	2011	invasive	blood	B	7-2	4	F1-5
30370	IBD-855	The Netherlands	2011	invasive	blood	B	7-1	1	F1-5
30371	IBD-856	The Netherlands	2011	invasive	CSF	B	7-1	1	F1-5
30374	IBD-859	The Netherlands	2011	invasive	blood	B	18-7	9	F1-7
30378	IBD-863	The Netherlands	2011	invasive	blood	B	7-2	4	F1-5
30380	IBD-865	The Netherlands	2011	invasive	blood	B	7-2	4	F1-5
30387	IBD-872	The Netherlands	2011	invasive	CSF	B	5-2	10-2	F1-5
30393	IBD-878	The Netherlands	2011	invasive	CSF	B	7-2	4	F1-5
30394	IBD-879	The Netherlands	2011	invasive	CSF	B	17-1	10-2	F3-7
30395	IBD-880	The Netherlands	2012	invasive	blood	B	7-2	13-1	F1-5
30410	IBD-896	The Netherlands	2012	invasive	CSF	B	7-2	30-2	F1-5
30411	IBD-897	The Netherlands	2012	invasive	blood	B	5-1	10-4	F1-5
30413	IBD-900	The Netherlands	2012	invasive	blood	B	7-2	30-4	F1-5

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
30418	IBD-762	France	2011	invasive	blood	B	7-2	4	F1-5
30430	IBD-774	France	2011	invasive	blood	B	12-1	13-2	F5-37
30432	IBD-776	France	2011	invasive	blood	B	7-2	4	F1-5
30437	IBD-781	France	2012	invasive	CSF	B	7-2	4	F1-5
30440	IBD-784	France	2012	invasive	blood	B	19-2	13-1	F1-21
30506	IBD-901	Germany	2011	invasive	blood	B	7-2	4	F1-5
30515	IBD-910	Germany	2011	invasive	blood	B	22	13	F1-5
30518	IBD-913	Germany	2011	invasive	blood	B	7-2	4	F1-5
30526	IBD-921	Germany		invasive	blood	B	5-1	2-2	F1-5
30529	IBD-924	Germany	2011	invasive	CSF	B	19	15	F1-5
30530	IBD-925	Germany	2011	invasive	blood	B	7-2	4	F1-5
30532	IBD-927	Germany	2011	invasive	CSF	B	7-2	13-2	F1-5
30533	IBD-928	Germany	2011	invasive	blood	B	5-1	10-4	F1-5
30544	IBD-939	Germany	2011	invasive	CSF	B	18-1	3	F1-5
30554	IBD-949	Germany	2011	invasive	CSF	B	12-1	13-1	F1-5
30555	IBD-950	Germany	2011	invasive	CSF	B	21	16	F1-15
30564	IBD-959	Germany	2011	invasive	blood	B	7-2	4	F1-5
30570	IBD-965	Germany	2011	invasive	CSF	B	7-2	4	F1-5
30577	IBD-972	Germany	2011	invasive	CSF	B	7-2	4	F1-5
30583	IBD-978	Germany	2011	invasive	blood	B	7-2	4	F1-5
30593	IBD-988	Germany	2011	invasive	blood	B	7-2	4	F1-5
30596	IBD-991	Germany	2011	invasive	CSF	B	12-1	13-1	F1-5
30597	IBD-992	Germany	2011	invasive	blood	B	7-2	15	F1-5
30598	IBD-993	Germany	2011	invasive	CSF	B	7-2	4	F1-5
30599	IBD-994	Germany	2011	invasive	CSF	B	7-2	4	F1-5
30601	IBD-996	Germany	2011	invasive	blood	B	7-2	13-9	F1-5
30602	IBD-462	Austria	2011	invasive	blood	B	17	16-3	F3-3
30648	IBD-537	Slovakia	2011	invasive	blood	B	7-1	1	F5-12
30656	IBD-548	Slovakia	2012	invasive	CSF	B	17	16-3	F1-20
30665	IBD-557	Slovakia	2012	invasive	blood	B	7-2	4	F1-5
30668	IBD-429	Finland	2011	invasive	blood	B	7-2	4	F1-5
30684	12003-14	Ireland	2014	invasive	blood	B	12-1	16	F1-5
30686	12004-14	Ireland	2014	invasive	blood	B	18-1	34-4	F1-5
30687	12005-10	Ireland	2010	invasive	blood	B	7-2	4	F1-5
30688	12005-14	Ireland	2014	invasive	blood	B	18-1	3	F1-5
30689	12006-14	Ireland	2014	invasive	blood	B	12-1	9	F1-5
30691	12007-14	Ireland	2014	invasive	blood	B	7-2	4	F1-7
30692	12008-14	Ireland	2014	invasive	CSF	B	18	25-11	F1-84
30693	12009-10	Ireland	2010	invasive	blood	B	7-2	13-1	F1-5
30696	12011-09	Ireland	2009	invasive	CSF	B	7-2	4	F5-12
30701	12015-10	Ireland	2010	invasive	blood	B	7-2	4	F1-5
30706	12020-10	Ireland	2010	invasive	blood	B	7-2	4	F1-7
30707	12021-10	Ireland	2010	invasive	blood	B	7-2	4	F1-5
30709	12022-10	Ireland	2010	invasive	blood	B	22	14	F1-5
30714	12027-13	Ireland	2013	invasive	blood	B	7	9	F5-5
30716	12029-13	Ireland	2013	invasive	blood	B	7-2	4	F5-12
30719	12032-13	Ireland	2013	invasive	blood	B	12-1	16	F1-5
30720	12033-13	Ireland	2013	invasive	CSF	B	12-1	16	F1-5
30722	12043-13	Ireland	2013	invasive	blood	B	7-2	4	F4-12
30728	12039-09	Ireland	2009	invasive	blood	B	7-2	4	F1-5
30730	12040-09	Ireland	2009	invasive	CSF	B	7-2	4	F1-5
30732	12041-09	Ireland	2009	invasive	blood	B	7-2	4	F1-5
30737	12044-09	Ireland	2009	invasive	blood	B	7-2	4	F5-12
30738	12044-13	Ireland	2013	invasive	blood	B	22	14	F3-7
30739	12045-09	Ireland	2009	invasive	blood	B	7-2	4	F5-8

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
30741	12046-09	Ireland	2009	invasive	blood	B	7-2	4	F5-1
30743	12047-13	Ireland	2013	invasive	blood	B	7-2	4	F5-12
30746	12051-09	Ireland	2009	invasive	blood	B	7-2	4	F1-5
30756	12087-06	Ireland	2006	invasive	blood	B	7-2	4	F1-5
30757	3193-99	Ireland	1999	invasive	blood	B	7-2	4	F1-5
30758	M1.1	Ireland	2013	carrier	throat swab	B	7-2	4	F5-12
30759	M13.1	Ireland	2013	carrier	throat swab	B	7-2	4	F5-12
30762	M2.1	Ireland	2013	carrier	throat swab	B	7-2	4	F5-12
30765	M31.1	Ireland	2013	carrier	throat swab	B	7-2	4	F5-12
30767	M34.1	Ireland	2013	carrier	throat swab	B	7-2	4	F5-12
30768	M35.1	Ireland	2013	carrier	throat swab	B	7-2	4	F5-12
30772	M46.3	Ireland	2013	carrier	throat swab	B	7-2	4	F5-43
30774	NZNMI00008	New Zealand	2000	invasive	blood		7-2	4	F1-5
30775	NZNMI00034	New Zealand	2000	invasive	blood		7-2	4	F1-5
30776	NZNMI00113	New Zealand	2000	invasive	blood		7-2	4	F5-18
30777	NZNMI00149	New Zealand	2000	invasive	blood		7-2	4	F1-5
30778	NZNMI00234	New Zealand	2000	invasive	blood		7-2	4	F1-5
30779	NZNMI01001	New Zealand	2001	invasive	blood		7-2	4	F5-18
30781	NZNMI01051	New Zealand	2001	invasive	blood		7-2	4	F1-5
30782	NZNMI01113	New Zealand	2001	invasive	blood		7-2	4	F1-5
30783	NZNMI01324	New Zealand	2001	invasive	blood		7-2	4	F1-5
30784	NZNMI02008	New Zealand	2002	invasive	blood		7-2	4	F1-5
30785	NZNMI02011	New Zealand	2002	invasive	blood		7-2	4	F1-5
30786	NZNMI02016	New Zealand	2002	invasive	blood		7-2	4	F5-5
30787	NZNMI02085	New Zealand	2002	invasive	blood		7-2	4	F1-5
30788	NZNMI03246	New Zealand	2003	invasive	blood		7-2	4	F1-5
30789	NZNMI03267	New Zealand	2003	invasive	blood		7-2	4	F1-5
30790	NZNMI03273	New Zealand	2003	invasive	blood		7-2	4	F1-5
30791	NZNMI03289	New Zealand	2003	invasive	blood		7-2	4	F1-5
30792	NZNMI04019	New Zealand	2004	invasive	CSF		7-2	4-12	F1-5
30793	NZNMI04020	New Zealand	2004	invasive	blood		7-2	4-12	F1-5
30794	NZNMI04152	New Zealand	2004	invasive	blood		7-2	4	F1-5
30795	NZNMI04168	New Zealand	2004	invasive	blood		7-2	4	F1-5
30796	NZNMI04170	New Zealand	2004	invasive	blood		7-2	4	F1-5
30797	NZNMI04188	New Zealand	2004	invasive	blood		7-2	4	F1-5
30798	NZNMI05053	New Zealand	2005	invasive	blood		7-2	4	F1-5
30799	NZNMI05067	New Zealand	2005	invasive	blood		7-2	4	F1-5
30800	NZNMI05094	New Zealand	2005	invasive	blood		7-2	4	F1-5
30801	NZNMI05099	New Zealand	2005	invasive	blood		7-2	4	F1-5
30802	NZNMI05139	New Zealand	2005	invasive	blood		7-2	4	F1-5
30803	NZNMI05146	New Zealand	2005	invasive	blood		7-2	4	F1-5
30804	NZNMI05152	New Zealand	2005	invasive	blood		7-2	4	F1-5
30805	NZNMI05154	New Zealand	2005	invasive	blood		7-2	4	F5-5
30806	NZNMI06007	New Zealand	2006	invasive	blood		7-2	4	F1-5
30807	NZNMI06031	New Zealand	2006	invasive	blood		7-2	4	F1-25
30808	NZNMI06037	New Zealand	2006	invasive	blood		7-2	4	F5-8
30809	NZNMI06046	New Zealand	2006	invasive	blood		7-2	4	F1-5
30810	NZNMI07030	New Zealand	2007	invasive	blood		7-2	4	F1-5
30811	NZNMI07033	New Zealand	2007	invasive	blood		7-2	4	F5-18
30812	NZNMI07045	New Zealand	2007	invasive	blood		7-2	4	F1-5
30813	NZNMI07078	New Zealand	2007	invasive	blood		7-2	4	F1-5

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
30814	NZNMI08038	New Zealand	2008	invasive	blood		7-2	4	F1-5
30815	NZNMI08057	New Zealand	2008	invasive	blood		7-2	4	F1-18
30816	NZNMI08089	New Zealand	2008	invasive	blood		7-2	4	F3-6
30817	NZNMI08091	New Zealand	2008	invasive	blood		7-2	4	F1-18
30818	NZNMI09017	New Zealand	2009	invasive	blood		7-2	4	F1-5
30819	NZNMI09023	New Zealand	2009	invasive	blood		7-2	4	F1-5
30820	NZNMI09045	New Zealand	2009	invasive	blood		7-2	4	F1-5
30821	NZNMI09053	New Zealand	2009	invasive	blood		7-2	4	F1-5
30822	NZNMI10023	New Zealand	2010	invasive	blood		7-2	4	F1-5
30823	NZNMI10035	New Zealand	2010	invasive	blood		7-2	4	F1-5
30824	NZNMI10042	New Zealand	2010	invasive	blood		7-2	4	F1-5
30825	NZNMI10074	New Zealand	2010	invasive	blood		7-2	4	F1-5
30826	NZNMI11024	New Zealand	2011	invasive	blood		7-2	4	F1-5
30827	NZNMI11051	New Zealand	2011	invasive	blood		7-2	4	F1-5
30828	NZNMI11068	New Zealand	2011	invasive	blood		7-2	4	F5-5
30829	NZNMI11089	New Zealand	2011	invasive	blood		7-2	4	F1-5
30830	NZNMI91049	New Zealand	1991	invasive	blood		7-2	4	F1-5
30831	NZNMI91058	New Zealand	1991	invasive	blood		7-2	4	F1-5
30832	NZNMI91059	New Zealand	1991	invasive	blood		7-2	13	F5-14
30833	NZNMI91060	New Zealand	1991	invasive	CSF		7-2	4	F1-5
30834	NZNMI92076	New Zealand	1992	invasive	blood		7-2	4	F1-5
30835	NZNMI92126	New Zealand	1992	invasive	blood		7-2	4	F1-5
30836	NZNMI92130	New Zealand	1992	invasive	blood		7-2	4	F1-5
30837	NZNMI93061	New Zealand	1993	invasive	blood		7-2	4	F1-5
30838	NZNMI93063	New Zealand	1993	invasive	blood		7-2	4	F1-5
30839	NZNMI93101	New Zealand	1993	invasive	blood		7-2	4	F1-5
30840	NZNMI93108	New Zealand	1993	invasive	blood		7-2	4	F1-5
30841	NZNMI94026	New Zealand	1994	invasive	blood		7-2	4	F1-5
30842	NZNMI94063	New Zealand	1994	invasive	blood		7-2	4	F1-5
30843	NZNMI94091	New Zealand	1994	invasive	blood		7-2	4	F1-5
30844	NZNMI94115	New Zealand	1994	invasive	blood		7-2	4	F1-5
30845	NZNMI94212	New Zealand	1994	invasive	blood		7-2	4	F1-5
30846	NZNMI94215	New Zealand	1994	invasive	blood		7-2	4	F1-5
30847	NZNMI94225	New Zealand	1994	invasive	blood		7-2	4	F1-5
30848	NZNMI95049	New Zealand	1995	invasive	blood		7-2	4	F1-5
30849	NZNMI95054	New Zealand	1995	invasive	blood		7-2	4	F1-5
30850	NZNMI95210	New Zealand	1995	invasive	blood		7-2	4	F1-5
30851	NZNMI95222	New Zealand	1995	invasive	blood		7-2	4	F1-5
30852	NZNMI95247	New Zealand	1995	invasive	blood		7-2	4	F1-5
30853	NZNMI95290	New Zealand	1995	invasive	CSF		7-2	4	F1-5
30854	NZNMI96086	New Zealand	1996	invasive	blood		7-2	4	F1-5
30855	NZNMI96168	New Zealand	1996	invasive	blood		7-2	4	F1-5
30856	NZNMI96290	New Zealand	1996	invasive	blood		7-2	4	F1-5
30857	NZNMI96320	New Zealand	unknown	invasive	unknown		7-2	4	F1-5
30858	NZNMI97005	New Zealand	1997	invasive	blood		7-2	4	F1-5
30859	NZNMI97042	New Zealand	1997	invasive	blood		7-2	4	F1-5
30860	NZNMI97070	New Zealand	1997	invasive	blood		7-2	4	F1-5
30861	NZNMI97218	New Zealand	1997	invasive	blood		7-2	4	F1-5
30862	NZNMI98053	New Zealand	1998	invasive	blood		7-2	4	F1-5
30863	NZNMI98058	New Zealand	1998	invasive	blood		7-2	4	F1-5
30864	NZNMI98153	New Zealand	1998	invasive	blood		7-2	4	F1-5
30865	NZNMI98302	New Zealand	1998	invasive	blood		7-2	4	F1-5
30866	NZNMI99035	New Zealand	1999	invasive	blood		7-2	4	F1-5
30867	NZNMI99133	New Zealand	1999	invasive	blood		7-2	4	
30868	NZNMI99152	New Zealand	1999	invasive	blood		7-2	4	F1-5

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
30869	NZNMI99293	New Zealand	1999	invasive	blood		7-2	4	F1-5
30872	IBD-999	Germany	2011	invasive	blood	B	7-2	4	F1-5
30873	IBD-1000	Germany	2012	invasive	blood	B	7-2	4	F1-5
30874	IBD-1001	Germany	2012	invasive	CSF	B	7-2	13-2	F1-5
30875	IBD-1002	Germany	2012	invasive	blood	B	5-1	2-2	F1-14
30880	IBD-1007	Germany	2012	invasive	CSF	B	7-2	4	F1-5
30886	IBD-1013	Germany	2011	invasive	blood	B	12-1	13-2	F1-5
30888	IBD-1015	Germany	2012	invasive	blood	B	18-1	3	F1-5
30889	IBD-1016	Germany	2012	invasive	blood	B	7-2	4	F1-5
30891	IBD-1018	Germany	2012	invasive	blood	B	7-2	4	F1-5
30899	IBD-1026	Germany	2012	invasive	blood	B	7-2	4	F1-5
30901	IBD-1028	Germany		invasive	blood	B	19	15	F1-5
30902	IBD-1029	Germany	2012	invasive	CSF	B	7-2	4	F1-5
30905	IBD-1032	Germany	2012	invasive	blood	B	18-1	3	F1-5
30914	IBD-1041	Germany	2012	invasive	CSF	B	7-2	4	F1-5
30915	IBD-1042	Germany	2012	invasive	blood	B	12-1	13-2	F1-5
30917	IBD-1044	Germany	2012	invasive	blood	B	7-2	4	F1-5
30928	IBD-1055	Germany	2012	invasive	CSF	B	18-1	3	F1-5
30936	IBD-1063	Germany	2012	invasive	CSF	B	7-2	4	F1-5
30937	IBD-1064	Germany	2012	invasive	blood	B	19	15	F5-5
30940	IBD-1067	Germany	2012	invasive	CSF	B	18-1	3-8	F1-5
30963	IBD-1090	Germany	2012	invasive	blood	B	7-2	4	F3-6
30965	IBD-1092	Germany		invasive	blood	B	12-1	13-1	F1-5
30974	IBD-1138	Germany	2012	invasive	blood	B	7-2	4	F1-5
30983	IBD-377	Czech Republic		invasive		B	19	15	F1-7
30988	IBD-382	Czech Republic		invasive		B	19	15	F1-14
31001	IBD-396	Czech Republic		invasive		B	17	16-4	F3-4
31003	IBD-398	Czech Republic		invasive		B	7-2	13-1	F1-5
31022	IBD-365	Slovenia	2012	invasive	blood	B	7-2	4	F1-5
31025	IBD-368	Slovenia	2012	invasive	blood	B	7	30	F1-7
31071	IBD-1097	Germany	2012	invasive	blood	B	7-2	4	F1-5
31072	IBD-1098	Germany	2012	invasive	blood	B	7-2	4	F1-5
31073	IBD-1099	Germany	2012	invasive	blood	B	7-2	4	F1-5
31100	IBD-1126	Germany	2012	invasive	blood	B	18-1	3-8	F1-5
31176	NM8250	UK	2011	invasive	blood	B	7-2	13-9	F1-25
31189	NM8736	UK	2011	carrier	rectal swab		17-1	23	F1-5
31190	NM9062	UK	2012	invasive	blood	B	7-2	4	F1-5
31196	NM9853	UK	2012	carrier	urethral swab		7-2	4	F1-5
31207	NM10833	UK	2013	carrier	rectal swab		5-2	10-1	F5-9
31209	NM10863	UK	2013	invasive	blood	B	7-2	4	F3-9
31210	NM10864	UK	2013	invasive	blood	B	7-2	4	F1-5
31212	NM11067	UK	2013	invasive	CSF	B	17	16-3	F5-7
34507	M13 240495	UK	2013	invasive		B	12-1	9	F1-5
34508	M13 240497	UK	2013	invasive		B	17-1	23	F1-5
34510	M13 240501	UK	2013	invasive		B	7-2	4	F1-5
34517	M13 240525	UK	2013	invasive		B	7-2	4	F1-5
34542	NZ98/254	New Zealand	1998			B	7-2	4	F1-5
34615	NM003	USA	1992			B	22-1	14	F5-2
34642	NM3139	USA	2009			B	22-1	14	F5-2
34661	NM518	USA	2002			B	7-1	1	F1-7
34910	IBD-741	Poland	2012	invasive	blood	C	7	30-5	F3-4

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
34912	IBD-755	Poland	2012	invasive	CSF	B	18-1	3	F1-5
34914	IBD-804	The Netherlands	2012	invasive	CSF	B	7-2	4	F1-5
34918	IBD-531	Norway	2012	invasive	blood	B	7-2	4	F1-5
34919	IBD-581	Poland	2011	invasive	blood	C	17	16-4	F1-7
34922	IBD-866	The Netherlands	2011	invasive	blood	B	7-2	4	F1-5
34924	IBD-795	France	2012	invasive	blood	B	7-2	4	F1-5
34926	IBD-919	Germany	2011	invasive	blood	B	18-1	3	F1-5
34945	IBD-499	Hungary	2012	invasive	CSF	C	5-2	10-2	F1-5
34946	IBD-506	Hungary	2011	invasive	blood	B	19	15	F1-5
34949	IBD-542	Slovakia	2012	invasive	CSF	B	7-2	13-1	F2-20
34955	IBD-1003	Germany	2012	invasive	blood	B	22	14-6	F5-5
34956	IBD-1022	Germany	2012	invasive	blood	B	7-2	4	F1-5
34957	IBD-1048	Germany	2012	invasive	CSF	B	22	9	F5-2
34961	IBD-1072	Germany	2012	invasive	blood	C	21-2	28	F5-7
34963	IBD-1142	Germany	2012	invasive	CSF	B	12-1	13-1	F3-9
34967	IBD-366	Slovenia	2011	invasive	blood	B	17	16-93	F5-5
34975	IBD-562	Greece	2011	invasive	blood	B	18	25	F1-5
34976	IBD-568	Greece	2012	invasive	blood	B	12-1		F1-5
34984	IBD-1105	Germany	2012	invasive	CSF	B	7-2	4	F1-5
35054	DAC-46	Norway		invasive			18	25	F1-7
35097	OX9930110	UK	1999	carrier	throat swab				
35195	Sl_MC_38	Slovenia	2013	carrier	throat swab	B	17-7	23	F1-5
35196	Sl_MC_39	Slovenia	2013	carrier	throat swab	B	17-7	23	F1-5
35197	Sl_MC_40	Slovenia	2013	carrier	throat swab	B	17-7	23	F1-5
35198	Sl_MC_41	Slovenia	2013	carrier	throat swab	B	18-1	3	F1-21
35200	Sl_MC_43	Slovenia	2012	carrier	throat swab	B	18	25	F1-5
35231	09.1331.Y	UK	2009	invasive	blood	B	7-2	4	F1-5
35232	09.1368.F	UK	2009	invasive	blood	B	22	14-6	F1-5
35235	09.1453.F	UK	2009	invasive	blood	B	7-2	4	F1-5
35236	09.1507.Q	UK	2009	invasive	CSF	B	7-2	4	F1-5
35241	09.1619.R	UK	2009	invasive	blood	B	7-2	4	F1-5
35247	09.2014.G	UK	2009	invasive	CSF	B	17-1	23-2	F5-2
35248	09.2045.J	UK	2009	invasive	blood	B	5-1	10-4	F5-56
35250	09.2082.C	UK	2009	invasive	blood	B	7-2	4	F1-7
35252	09.2270.V	UK	2009	invasive	blood	B	7-2	4	F1-19
35253	09.2386.V	UK	2009	invasive	blood	B	7-2	4	F1-5
35255	09.2522.Q	UK	2009	invasive	blood	B	7-2	4	F1-5
35257	09.2658.E	UK	2009	invasive	blood	B	7-2	4	F1-2
35261	09.2961.S	UK	2009	invasive	blood	B	7-2	4	F1-5
35267	09.3271.E	Iceland	2009	invasive	blood	B	7-2	4	F1-5
35270	09.3368.H	UK	2009	invasive	blood	B	7-2	4	F1-5
35280	10.2012.H	UK	2010	invasive	blood	B	7-2	4	F1-25
35287	10.2422.K	UK	2010	invasive	CSF	B	7-2	4	F1-5
35290	10.2763.M	UK	2010	invasive	blood	B	7-2	4	F1-5
35291	10.2850.R	UK	2010	invasive	blood	B	7-2	4	F1-5
35294	10.3051.M	UK	2010	invasive	blood	B	7	30	F1-15
35296	11.1071.G	UK	2011	invasive	blood	B	7-2	4	F1-5
35297	11.1190.L	UK	2010	invasive	blood	B	18-1	14	F1-5
35298	11.1286.D	UK	2011	invasive	blood	B	18-1	14	F1-5
35299	11.1312.X	UK	2011	invasive	blood	B	7-2	4	F1-5

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
35302	11.1426.J	UK	2011	invasive	blood	B	7-2	4	F1-5
35303	11.1479.T	UK	2011	invasive	blood	B	7-2	4	F3-28
35308	11.1825.A	UK	2011	invasive	blood	B	22	14-6	F1-5
35309	11.1826.C	UK	2011	invasive	CSF	B	22	14-6	F1-5
35312	11.2463.D	UK	2011	invasive	blood	B	7-2	4	F5-8
35319	11.4999.T	UK	2011	invasive	blood	B	7-2	4	F1-5
35324	11.5842.B	UK	2011	invasive	blood	B	7-2	4	F5-36
35335	12.8534.X	UK	2012	invasive	blood	B	7-2	4	F1-5
35354	13.4278.S	UK	2013	invasive	blood	B	7-2	4	F1-5
35362	12.3021.W	UK	2012	invasive	blood	B	7-1	1	F1-5
35378	09.1936.F	UK	2009	invasive	CSF	B	22	14	F5-56
35381	13.4765.S	UK	2013	invasive	blood	B	7	30-1	F1-5
35388	13.6067.H	UK	2013	invasive	blood	B	7-2	4	F1-5
35390	13.6161.J	UK	2013	invasive	blood	B	7-2	4	F1-5
35392	13.6756.W	UK	2013	invasive	blood	B	19	15-7	F1-7
35394	13.7552.F	UK	2013	invasive	blood	B	7-2	4	F1-5
35426	M13 240532	UK	2013	invasive		B	5-1	10-1	F1-5
35434	M13 240548	UK	2013	invasive		B	17-1	23-2	F1-7
35437	M13 240560	UK	2013	invasive		B	12-6	13-4	F1-7
35438	M13 240561	UK	2013	invasive		B	7-2	4	F1-5
35440	M13 240564	UK	2013	invasive		B	7-2	4	F1-5
35455	M13 240598	UK	2013	invasive		Y	17-1	23-2	F1-7
35457	M13 240601	UK	2013	invasive		B	7-2	4	F1-5
35459	M13 240608	UK	2013	invasive		B	19	15-1	F1-5
35462	M13 240613	UK	2013	invasive		B	7-2	4	F1-5
35463	M13 240614	UK	2013	invasive		B	7-2	4	F3-9
35467	M13 240620	UK	2013	invasive		B	18-1	3	F1-5
35470	M13 240624	UK	2013	invasive		B	7-2	4	F5-5
35471	M13 240626	UK	2013	invasive		B	22	14	F1-5
35474	M13 240636	UK	2013	invasive		B	5-2	10-1	F3-3
35478	M13 240641	UK	2013	invasive		B	7-2	4	F3-9
35482	M13 240647	UK	2013	invasive		B	7-2	4	F1-5
35483	M13 240648	UK	2013	invasive		B	12-1	16	F1-5
35489	M13 240670	UK	2013	invasive		B	12-1	16	F1-5
35492	M13 240675	UK	2013	invasive		B	7-2	4	F5-8
35497	M13 240692	UK	2013	invasive		B	7-2	4	F3-3
35503	M13 240698	UK	2013	invasive		B	7-2	4	F1-5
35504	M13 240699	UK	2013	invasive		B	7-2	4	F1-5
35509	M13 240708	UK	2013	invasive		B	7-2	4	F1-5
35510	M13 240709	UK	2013	invasive		B	17	16-3	F5-7
35526	M13 240738	UK	2013	invasive		B	12-1	16	F1-5
35532	M13 240745	UK	2013	invasive		B	12-1	16	F1-5
35533	M13 240746	UK	2013	invasive		B	7-2	4	F1-80
35539	M13 240753	UK	2013	invasive		B	18-1	14	F1-5
35541	M14 240003	UK	2014	invasive		B	22	26	F1-5
35542	M14 240008	UK	2014	invasive		B	7-2	4	F3-6
35548	M14 240016	UK	2014	invasive		B	7-2	4	F1-5
35549	M14 240017	UK	2014	invasive		B	7-2	4	F1-5
35550	M14 240018	UK	2014	invasive		B	7-2	13-2	F4-1
35552	M14 240021	UK	2014	invasive		B	7-2	4	F1-5
35556	M14 240027	UK	2014	invasive		B	7-2	4	F1-5
35562	M14 240045	UK	2014	invasive		B	7-2	4	F1-5
35563	M14 240046	UK	2014	invasive		B	12-1	9	F1-5
35566	M14 240055	UK	2014	invasive		B	7-2	4	F1-5
35573	M14 240074	UK	2014	invasive		B	7-2	4	F1-5

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
35578	M14 240085	UK	2014	invasive		B	7-2	4	F1-18
35585	M14 240096	UK	2014	invasive		B	7-2	4	F1-5
35586	M14 240098	UK	2014	invasive		B	18-1	3	F3-3
35594	M14 240110	UK	2014	invasive		B	7-2	4	F5-8
35606	M14 240124	UK	2014	invasive		B	7-2	4	F1-18
35609	M14 240127	UK	2014	invasive		B	7-2	4	F1-5
35613	M14 240134	UK	2014	invasive		B	7-2	4	F5-17
35625	M14 240153	UK	2014	invasive		B	7-2	4	F1-5
35629	M14 240174	UK	2014	invasive		B	7-2	4	F1-5
35646	M14 240217	UK	2014	invasive		B	17-1	23	F1-5
35656	M14 240232	UK	2014	invasive		B	18-1	34	F1-5
35657	M14 240233	UK	2014	invasive		B	19-3	15	F1-5
35673	M14 240256	UK	2014	invasive		B	18-1	3	F1-5
35681	M14 240267	UK	2014	invasive		B	7-2	4	F5-25
35704	M14 240306	UK	2014	invasive		B	7-2	4	F1-5
35705	M14 240309	UK	2014	invasive		B	7-2	4	F1-5
35708	M14 240312	UK	2014	invasive		B	12-1	13-1	F5-5
35719	M14 240331	UK	2014	invasive		B	7-2	4	F1-5
35732	M14 240353	UK	2014	invasive		B	7-2	4	F1-5
35735	M14 240357	UK	2014	invasive		B	7-2	*deleted*	F1-5
35752	M14 240387	UK	2014	invasive		B	7-2	4-10	F1-5
35758	M14 240396	UK	2014	invasive		B	19	15-1	F1-5
35759	M14 240400	UK	2014	invasive		B	18-1	3	F1-5
35766	M14 240410	UK	2014	invasive		B	12-1	16	F1-24
35768	M14 240425	UK	2014	invasive		B	22	14	F1-5
35771	M14 240429	UK	2014	invasive		B	12-1	16	F1-5
35780	M14 240455	UK	2014	invasive		B	12-1	16-117	F1-5
35789	M14 240471	UK	2014	invasive		B	22	14-6	F1-5
35790	M14 240472	UK	2014	invasive		B	22	14-6	F1-5
35793	M14 240475	UK	2014	invasive		B	12-3	4	F3-9
35794	M14 240476	UK	2014	invasive		B	7-2	4	F1-5
35795	M14 240477	UK	2014	invasive		B	12-1	16	F1-5
35803	M14 240488	UK	2014	invasive		B	17-1	23	F1-5
35805	M14 240490	UK	2014	invasive		B	7-2	4	F1-5
35834	IH167461	Finland	2011			B	7-2	4	F1-5
36204	BM59	Greece	1996	invasive	throat swab	B	7-2	4	F1-5
36205	BM59a	Greece	1996	carrier	throat swab	B	7-2	4	F1-5
36207	BM59c	Greece	1996	carrier	throat swab	B	7-2	4	F1-5
36208	BM65	Greece	1997	invasive	CSF	B	18-1	3	F1-5
36209	BM65a	Greece	1997	carrier	throat swab	B	18-1	3	F1-5
36210	BM65b	Greece	1997	carrier	throat swab	B	18-1	3	F1-5
36211	BM65c	Greece	1997	carrier	throat swab	B	18-1	3	F1-5
36212	OX0050940	UK	2000	carrier	throat swab	B	7-2	4-34	F1-5
36213	OX0052076	UK	2000	carrier	throat swab	B	17	16-3	F4-3
36214	OX01060217	UK	2001	carrier	throat swab	B	21	16	F1-5
36215	OX01060367	UK	2001	carrier	throat swab	B	17	16-3	F3-9
36216	OX01060896	UK	2001	carrier	throat swab	B	18-7	9	F1-5
36217	OX01061140	UK	2001	carrier	throat swab	NG	18-1	3	F1-7

Id	Isolate	country	year	disease	source	sero-group	PorA VR1	PorA VR2	FetA VR
36218	OX01061310	UK	2001	carrier	throat swab	B	19-1	15-1	F1-5
36219	OX01061577	UK	2001	carrier	throat swab	NG	7-2	4	F1-5
36220	OX01061579	UK	2001	carrier	throat swab	B	21	16	F1-5
36221	OX01062311	UK	2001	carrier	throat swab	B	7-2	4	F1-5
36222	OX01062468	UK	2001	carrier	throat swab	NG	18-1	3	F1-7
36223	OX9931197	UK	1999	carrier	throat swab	B	17	16-3	F4-3
36224	OX9931238	UK	1999	carrier	throat swab	B	17	16-3	F4-3
36225	OX9931966	UK	1999	carrier	throat swab	B	7-2	4	F1-5
36226	OX9932316	UK	1999	carrier	throat swab	B	7-2	4	F1-5
36227	OX9932386	UK	1999	carrier	throat swab	NG	7-2	4	F1-5

Appendix B.

Major Publications

Bound with Chapter One: Introduction

- Bratcher HB, Bennett JS, Maiden MCJ. Evolutionary and genomic insights into meningococcal biology. **Future Microbiology** 2012, 7:873-85.

Bound with Chapter Three: Proof of Concept

- Bratcher HB, Corton C, Jolley KA, Parkhill J, Maiden MCJ. A gene-by-gene population genomics platform: *de novo* assembly, annotation and genealogical analysis of 108 representative *Neisseria meningitidis* genomes. **BMC Genomics** 2014, **15:**1138

Public Communication

- Bratcher HB, MCJ Maiden. Analysis of the population structure and evolution of the highly diverse meningococcal Lineage 3. 13th Congress of the European Meningococcal Disease Society, Amsterdam, The Netherlands, September 2015. SPEAKER
- Embracing 'next generation' technologies in Outbreak Investigations. Centers for Disease Control and Prevention, Atlanta, GA, USA September 2011. PANEL DISUSSION SPEAKER

Abstracts and Posters

- Bratcher HB, Kesanopoulos K, Tzanakaki G, Maiden MCJ. Whole genome comparison of *Neisseria meningitidis* isolates from patients and their close family contacts using gene-by-gene analysis. International Pathogenic *Neisseria* Conference, North Carolina, USA, October 2014. POSTER
- Bratcher HB, Corton C, Bentley SD, Parkhill J, Maiden MCJ. Validation of high throughput draft genome sequence production for a diverse collection of *Neisseria meningitidis* isolates. International Pathogenic *Neisseria* Conference, Wurzburg, Germany, September 2012. POSTER

- Bratcher HB, MacLennan J, Maiden MCJ. Temporally paired isolates of *Neisseria meningitidis* compared by whole genome sequencing. International Pathogenic *Neisseria* Conference, Wurzburg, Germany, September 2012. POSTER
- Bratcher HB, Corton C, Jolley KA, Bennett J, Harrison OB, Brehony C, Bentley SD, Maiden MCJ. Genome Analysis of 108 meningococci: a General Approach to Bacterial Population Genomics. Society for General Microbiology, York, UK, September 2011. POSTER

Co-authored Publication and Abstracts

Book Chapter

- Bennett JS, Bratcher HB, Brehony C, Harrison OB, Maiden MCJ (2014). The Genus *Neisseria* in Rosenberg, DeLong, Lory, Stackebrandt, Thompson (Eds.), ***The Prokaryotes,* 4th Edition, Volume 8: Alphaproteobacteria and Betaproteobacteria.** Springer-Verlag Berlin Heidelberg.

Journal Publications

- Lucidarme J, Hill DM, Bratcher HB, Gray SJ, du Plessis M, Tsang RS, Vazquez JA, Taha MK, Ceyhan M, Efron AM, Gorla MC, Findlow J, Jolley KA, Maiden MC, Borrow R. Genomic resolution of an aggressive, widespread, diverse and expanding meningococcal serogroup B, C and W lineage. **Journal of Infectious Disease**, 2015 xx, 1-9, ARTICLE IN PRESS
- Hill DMC, Lucidarme J, Gray SJ, Newbold LS, Ure R, Brehoney C, Harrison OB, Bray JE, Jolley KA, Bratcher HB, Parkhill J, Tang CM, Borrow R and Maiden MCJ. Genomic epidemiology of age-associated meningococcal lineages in national surveillance. **Lancet ID** 2015, ARTICLE IN PRESS
- Alexandra Jackson, María Victoria Humbert, Anish Pandey, Holly Bratcher, Myron Christodoulides. Draft Genome Sequence of *Dichelobacter nodosus* ATCC 25549, Strain VPI 2340 [11342], a Bacterium Causing Footrot in Sheep. **Genome Announcement** 2015 vol.3 issue 5

- Törös B, Hedberg ST, Unemo M, Jacobsson S, Hill DM, Olcén P, Fredlund H, Bratcher HB, Jolley KA, Maiden MC, Mölling P. Genome-Based Characterization of Emergent Invasive *Neisseria meningitidis* Serogroup Y Isolates in Sweden from 1995 to 2012. **Journal of Clinical Microbiology** 2015, **53:**2154-62
- Deasy AM, Guccione E, Dale AP, Andrews N, Evans CM, Bennett JS, Bratcher HB, Maiden MC, Gorringe AR, Read RC. Nasal Inoculation of the Commensal *Neisseria lactamica* Inhibits Carriage of *Neisseria meningitidis* by Young Adults: A Controlled Human Infection Study. **Clinical Infectious Diseases** 2015, 60:1512-20
- Harrison OB, Claus H, Jiang Y, Bennett JS, Bratcher HB, Jolley KA, Corton C, Care R, Poolman JT, Zollinger WD, Frasch CE, Stephens DS, Feavers I, Frosch M, Parkhill J, Vogel U, Quail MA, Bentley SD, Maiden MC. Description and nomenclature of *Neisseria meningitidis* capsule locus. **Emerging Infectious Diseases** 2013, 19:566-73
- Jolley KA, Hill DM, Bratcher HB, Harrison OB, Feavers IM, Parkhill J, Maiden MC. Resolution of a meningococcal disease outbreak from whole-genome sequence data with rapid Web-based analysis methods. **Journal of Clinical Microbiology** 2012, 50:3046-53
- Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, Maiden MC. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. **Microbiology** 2012, 158:1005-15

Abstracts and Posters

- Harrison OB, Bratcher HB, Eales J, Gray S, Corton C, Jolley KA, Bentley SD, Parkhill J and Maiden MCJ. Genetic stability of the *Neisseria meningitidis* capsule locus: implications for capsule switching. International Pathogenic *Neisseria* Conference, Wurzburg, Germany, September 2012.
- Caugant DA, Harrison OB, Bratcher HB, Jolley KA, Parkhill J, Maiden MCJ. Forty years of evolution of a meningococcal clone analysed by whole genome sequencing. International Pathogenic *Neisseria* Conference, Wurzburg, Germany, September 2012.

Appendix C.

Bacterial domain (BACT)* and *Neisseria* species (NEIS) locus tag identifiers for 56 ribosomal protein genes

* The 'BACT' locus tag identifiers are used exclusively in the ribosomal multilocus sequence typing sequence definition database and the 'NEIS' locus tag identifiers are used in the *Neisseria* PubMLST.org sequence definition database.

Neisseria species	Bacterial domain	gene name	full name/product
NEIS0120	BACT000030	*rplA*	50S ribosomal protein L1
NEIS0135	BACT000031	*rplB*	50S ribosomal protein L2
NEIS0132	BACT000032	*rplC*	50S ribosomal protein L3
NEIS0133	BACT000033	*rplD*	50S ribosomal protein L4
NEIS0144	BACT000034	*rplE*	50S ribosomal protein L5
NEIS0147	BACT000035	*rplF*	50S ribosomal protein L6
NEIS1257	BACT000038	*rplI*	50S ribosomal protein L9
NEIS0121	BACT000039	*rplJ*	50S ribosomal protein L10
NEIS0119	BACT000040	*rplK*	50S ribosomal protein L11
NEIS0122	BACT000036	*rplL*	50S ribosomal protein L7/L12
NEIS2038	BACT000042	*rplM*	50S ribosomal protein L13
NEIS0145	BACT000043	*rplN*	30S ribosomal protein S14
NEIS0151	BACT000044	*rplO*	50S ribosomal protein L15
NEIS0139	BACT000045	*rplP*	50S ribosomal protein L16
NEIS0159	BACT000046	*rplQ*	50S ribosomal protein L17
NEIS0148	BACT000047	*rplR*	50S ribosomal protein L18
NEIS0531	BACT000048	*rplS*	50S ribosomal protein L19
NEIS0675	BACT000049	*rplT*	50S ribosomal protein L20
NEIS1847	BACT000050	*rplU*	50S ribosomal protein L21
NEIS0137	BACT000051	*rplV*	50S ribosomal protein L22
NEIS0134	BACT000052	*rplW*	50S ribosomal protein L23
NEIS0143	BACT000053	*rplX*	50S ribosomal protein L24
NEIS0817	-	*rplY*	50S ribosomal protein L25
NEIS1848	BACT000056	*rpmA*	50S ribosomal protein L27
NEIS1851	BACT000057	*rpmB*	50S ribosomal protein L28
NEIS0140	BACT000058	*rpmC*	50S ribosomal protein L29
NEIS0150	BACT000059	*rpmD*	50S ribosomal protein L30
NEIS1928	BACT000060	*rpmE1*	50S ribosomal protein L31
NEIS0920	BACT000060	*rpmE2*	50S ribosomal protein L31
NEIS0312	BACT000061	*rpmF*	50S ribosomal protein L32
NEIS1850	BACT000062	*rpmG*	50S ribosomal protein L33
NEIS0319	BACT000063	*rpmH*	50S ribosomal protein L34
NEIS0674	BACT000064	*rpmI*	50S ribosomal protein L35
NEIS0154	BACT000065	*rpmJ1*	50S ribosomal protein L36
NEIS0919	BACT000065	*rpmJ2*	50S ribosomal protein L36
NEIS1238	BACT000001	*rpsA*	30S ribosomal protein S1
NEIS2080	BACT000002	*rpsB*	30S ribosomal protein S2
NEIS0138	BACT000003	*rpsC*	30S ribosomal protein S3
NEIS0157	BACT000004	*rpsD*	30S ribosomal protein S4

Neisseria species	Bacterial domain	gene name	full name/product
NEIS0149	BACT000005	*rpsE*	30S ribosomal protein S5
NEIS1260	BACT000006	*rpsF*	30S ribosomal protein S6
NEIS0126	BACT000007	*rpsG*	30S ribosomal protein S7
NEIS0146	BACT000008	*rpsH*	30S ribosomal protein S8
NEIS2037	BACT000009	*rpsI*	30S ribosomal protein S9
NEIS0129	BACT000010	*rpsJ*	30S ribosomal protein S10
NEIS0156	BACT000011	*rpsK*	30S ribosomal protein S11
NEIS0125	BACT000012	*rpsL*	30S ribosomal protein S12
NEIS0155	BACT000013	*rpsM*	30S ribosomal protein S13
NEIS0142	BACT000014	*rpsN*	50S ribosomal protein L14
NEIS0552	BACT000015	*rpsO*	30S ribosomal protein S15
NEIS0534	BACT000016	*rpsP*	30S ribosomal protein S16
NEIS0141	BACT000017	*rpsQ*	30S ribosomal protein S17
NEIS1258	BACT000018	*rpsR*	30S ribosomal protein S18
NEIS0136	BACT000019	*rpsS*	30S ribosomal protein S19
NEIS1688	BACT000020	*rpsT*	30S ribosomal protein S20
NEIS1921	BACT000021	*rpsU*	30S ribosomal protein S21

Appendix D.

Observed carriage and disease association in clonal complexes

The *Neisseria* pubMLST.org webpage was accessed on 07/2015 and queried for clonal complex and disease.

clonal complex		total isolate count	carrier	%	disease	%
found in UK study and found in Oxford region	**ST-41/44**	4506	815	18.09	3691	81.91
	ST-11	3722	177	4.76	3545	95.24
	ST-32	2413	271	11.23	2142	88.77
	ST-269	1421	141	9.92	1280	90.08
	ST-8	1223	12	0.98	1211	99.02
	ST-23	898	260	28.95	638	71.05
	ST-213	561	157	27.99	404	72.01
	ST-22	635	285	44.88	350	55.12
	ST-18	387	52	13.44	335	86.56
	ST-60	466	165	35.41	301	64.59
	ST-35	558	258	46.24	300	53.76
	ST-334	233	15	6.44	218	93.56
	ST-461	233	33	14.16	200	85.84
	ST-162	260	85	32.69	175	67.31
	ST-167	302	145	48.01	157	51.99
	ST-103	214	63	29.44	151	70.56
	ST-364	151	36	23.84	115	76.16
	ST-254	236	130	55.08	106	44.92
	ST-865	131	43	32.82	88	67.18
	ST-37	78	23	29.49	55	70.51
	ST-1157	121	75	61.98	46	38.02
	ST-282	45	14	31.11	31	68.89
	ST-750	47	20	42.55	27	57.45
	ST-198	244	228	93.44	16	6.56
	ST-53	329	313	95.14	16	4.86
	ST-92	92	81	88.04	11	11.96
	ST-212	41	31	75.61	10	24.39
	ST-178	75	66	88.00	9	12.00
	ST-1136	60	52	86.67	8	13.33
	ST-1117	60	58	96.67	2	3.33
found in UK study but not found in Oxford region	**ST-174**	214	60	28.04	154	71.96
	ST-116	50	35	70.00	15	30.00
	ST-4240/6688	35	3	8.57	32	91.43
	ST-106	34	33	97.06	1	2.94
	ST-376	75	63	84.00	12	16.00
not found in UK study or Oxford region	**ST-5**	819	99	12.09	720	87.91
	ST-1	412	46	11.17	366	88.83
	ST-175	317	71	22.40	246	77.60
	ST-181	98	34	34.69	64	65.31

clonal complex		total isolate count	carrier	%	disease	%
	ST-4821	132	88	66.67	44	33.33
	ST-231	53	21	39.62	32	60.38
	ST-226	57	29	50.88	28	49.12
	ST-4	25	0	0.00	25	100.00
	ST-549	85	70	82.35	15	17.65
	ST-292	29	25	86.21	4	13.79
	unassigned	5207	1772	34.03	3435	65.97
	Total	27406	6574	23.99	20832	76.01

Appendix E.

Computational programs and bioinformatics tools

VELVET [32, 33]

BIGSdb / *Neisseria* PubMLST database [34, 185]

http://pubmlst.org/neisseria/

MEGA / MEGA-CC [36, 46, 115, 186]

STADEN: PREGAP | GAP4 [52]

Bowtie [53]

Tablet [187, 188]

SAM tools [189]

BEAST [48]

Pfam [41]

http://pfam.xfam.org/

UniProtKB [190]

http://www.uniprot.org/

Appendix F.

Abbreviations

indel	insertion or deletion of sequence
MLEE	multilocus enzyme electrophoresis
MLST	multilocus sequence type
cnl	capsule null locus
cc	clonal complex
41/44cc	ST-41/44 clonal complex
ST	sequence type
MRF	Meningitis Research Foundation
CSF	cerebral spinal fluid
BIGSdb	Bacterial Isolate Genome Sequence Database
NEIS	*Neisseria* species (annotation locus tag identifier)
PATRIC	Pathosystems Resource Integration Center
KEGG	Kyoto Encyclopedia of Genes and Genomes
PROKKA	Prokaryotic Genomes Automatic Annotation Pipeline
BEAST	Bayesian Evolutionary Analysis Sampling Trees
Pfam	A comprehensive database of protein domain families based on seed alignments
UniProtKB	UniProt Knowledge Base; a web based resource of protein and functional information
contig	contiguous nucleotide sequence
OLC	overlap layout consensus assembly
SBH	sequence by hybridization
rMLST	ribosomal multilocus sequence typing
rST	ribosomal sequence type
BACT	bacterial domain (annotation locus tag identifier)

References

1. Sturtevant AH: *A history of genetics.* New York,: Harper & Row; 1965.
2. Morgan TH: **Chromosomes and Heredity.** *The American Naturalist* 1910, **44:**449-496.
3. Avery OT, Macleod CM, McCarty M: **Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii.** *J Exp Med* 1944, **79:**137-158.
4. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors.** *Proceedings of the National Acadamy of Sciences of the USA* 1977, **74:**5463-5467.
5. Beckwith J: *Making Genes, Making Waves.* Harvard University Press; 2002.
6. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H: **Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction.** *Cold Spring Harb Symp Quant Biol* 1986, **51 Pt 1:**263-273.
7. Min Jou W, Haegeman G, Ysebaert M, Fiers W: **Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein.** *Nature* 1972, **237:**82-88.
8. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE: **Fluorescence detection in automated DNA sequence analysis.** *Nature* 1986, **321:**674-679.
9. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* RD.** *Science* 1995, **269:**496-512.
10. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, et al.: **The minimal gene complement of *Mycoplasma genitalium*.** *Science* 1995, **270:**397-403.
11. Bennett JS, Jolley KA, Earle SG, Corton C, Bentley SD, Parkhill J, Maiden MC: **A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus *Neisseria*.** *Microbiology* 2012, **158:**1570-1580.
12. Schielke S, Kurzai O, Frosch M: **Virulence determinants involved in differential host niche adaptation of *Neisseria meningitidis* and *Neisseria gonorrhoeae*.** *Medical Microbiology and Immunology* 2010, **199:**185-196.
13. Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A, Carniel E: ***Yersinia pestis,* the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*.** *PNAS* 1999, **96:**14043-14048.
14. Pearson T, Okinaka RT, Foster JT, Keim P: **Phylogenetic understanding of clonal populations in an era of whole genome sequencing.** *Infection, Genetics and Evolution* 2009, **9:**1010-1019.
15. Achtman M: **Insights from genomic comparisons of genetically monomorphic bacterial pathogens.** *Philos Trans R Soc Lond B Biol Sci* 2012, **367:**860-867.
16. Binnewies TT, Motro Y, Hallin PF, Lund O, Dunn D, La T, Hampson DJ, Bellgard M, Wassenaar TM, Ussery DW: **Ten years of bacterial genome sequencing: comparative-genomics-based discoveries.** *Funct Integr Genomics* 2006, **6:**165-185.
17. Lederberg J: **Cell genetics and hereditary symbiosis.** *Physiol Rev* 1952, **32:**403-430.
18. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation.** *Nature* 2000, **405:**299-304.
19. Ochman H: **Neutral mutations and neutral substitutions in bacterial genomes.** *Mol Biol Evol* 2003, **20:**2091-2096.
20. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M, et al: **Evidence for several waves of global transmission in the seventh cholera pandemic.** *Nature* 2011, **477:**462-465.
21. Cramer N, Klockgether J, Wrasman K, Schmidt M, Davenport CF, Tummler B: **Microevolution of the major common Pseudomonas aeruginosa clones C and PA14 in cystic fibrosis lungs.** *Environ Microbiol* 2011, **13:**1690-1704.

22. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, et al: **Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology.** *PLoS One* 2011, **6:**e22751.
23. Köser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR, et al: **Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak.** *New England Journal of Medicine* 2012, **366:**2267-2275.
24. Catlin BW, Cunningham LS: **Transforming activities and base contents of deoxyribonucleate preparations from various *Neisseriae*.** *JGM* 1961, **26:**303-312.
25. Maiden MC, Frosch M: **Molecular techniques for the investigation of meningococcal disease epidemiology.** *Molecular Biotechnology* 2001, **18:**119-134.
26. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, et al: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.** *PNAS* 1998, **95:**3140-3145.
27. Selander RK, Caugant DA, Ochman H, Musser JM, Gilmour MN, Whittam TS: **Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics.** *Appl Environ Microbiol* 1986, **51:**837-884.
28. **MRF Meningococcus Genome Library** [http://www.meningitis.org/research/genome]
29. Maiden MC, Stuart JM, UK_Meningococcal_Carriage_Group: **Carriage of serogroup C meningococci 1 year after meningococcal C conjugate polysaccharide vaccination.** *Lancet* 2002, **359:**1829-1831.
30. Syed F, Grunenwald H, Caruccio N: **Optimized library preparation method for next-generation sequencing.** *Nature Methods* 2009, **6:**I-Ii.
31. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ: **A large genome center's improvements to the Illumina sequencing system.** *Nat Methods* 2008, **5:**1005-1010.
32. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Research* 2008, **18:**821-829.
33. Zerbino D: **Using the Velvet *de novo* Assembler for Short-Read Sequencing Technologies.** *Current Protocols in Bioinformatics* 2010, **11:**1-12.
34. Jolley KA, Maiden MC: **BIGSdb: Scalable analysis of bacterial genome variation at the population level.** *BMC Bioinformatics* 2010, **11:**595.
35. Harrison OB, Bray, J.E., Maiden, M. C. and Caugant, D. A.: **Genomic analysis of the evolution and global spread of hyper-invasive meningococcal lineage 5.** *EBioMedicine* 2015, **2:**234-243.
36. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28:**2731-2739.
37. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Research* 2000, **28:**27-30.
38. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, et al: **PATRIC, the bacterial bioinformatics database and analysis resource.** *Nucleic Acids Research* 2014, **42:**D581-591.
39. Seemann T: **Prokka: rapid prokaryotic genome annotation.** *Bioinformatics* 2014.
40. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification.** *Bmc Bioinformatics* 2010, **11**.
41. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al: **The Pfam protein families database.** *Nucleic Acids Research* 2012, **40:**D290-D301.
42. Bentley SD, Vernikos GS, Snyder LA, Churcher C, Arrowsmith C, Chillingworth T, Cronin A, Davis PH, Holroyd NE, Jagels K, et al: **Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18.** *PLoS Genet* 2007, **3:**e23.

43. Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR, Morelli G, Basham D, Brown D, Chillingworth T, et al: **Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491.** *Nature* 2000, **404:**502-506.
44. Budroni S, Siena E, Hotopp JCD, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR, Daugherty SC, Angiuoli SV, et al: ***Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination.** *PNAS* 2011, **108:**4494-4499.
45. Bryant D, Moulton V: **Neighbor-net: an agglomerative method for the construction of phylogenetic networks.** *Mol Biol Evol* 2004, **21:**255-265.
46. Kumar S, Stecher G, Peterson D, Tamura K: **MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis.** *Bioinformatics* 2012, **28:**2685-2686.
47. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies.** *Mol Biol Evol* 2006, **23:**254-267.
48. Drummond AJ, Suchard MA, Xie D, Rambaut A: **Bayesian phylogenetics with BEAUti and the BEAST 1.7.** *Mol Biol Evol* 2012, **29:**1969-1973.
49. Russell JE, Jolley KA, Feavers IM, Maiden MC, Suker J: **PorA variable regions of *Neisseria meningitidis*.** *Emerg Infect Dis* 2004, **10:**674-678.
50. Thompson EAL, Feavers IM, Maiden MCJ: **Antigenic diversity of meningococcal enterobactin receptor FetA, a vaccine component.** *Microbiology* 2003, **149:**1849-1858.
51. Brehony C, Wilson DJ, Maiden MC: **Variation of the factor H-binding protein of *Neisseria meningitidis*.** *Microbiology* 2009, **155:**4155-4169.
52. Staden R: **The Staden sequence analysis package.** *Molecular Biotechnology* 1996, **5:**233-241.
53. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biology* 2009, **10:**R25.
54. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nature Methods* 2012, **9:**357-U354.
55. Wu DY, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, et al: **A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea.** *Nature* 2009, **462:**1056-1060.
56. Whitworth DE: **Genomes and knowledge - a questionable relationship?** *Trends Microbiol* 2008, **16:**512-519.
57. Galperin MY, Koonin EV: **From complete genome sequence to 'complete' understanding?** *Trends in Biotechnology* 2010, **28:**398-406.
58. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome.** *Current Opinion in Genetics & Development* 2005, **15:**589-594.
59. Shendure J, Ji H: **Next-generation DNA sequencing.** *Nat Biotechnol* 2008, **26:**1135-1145.
60. Orsi RH, Bowen BM, Wiedmann M: **Homopolymeric tracts represent a general regulatory mechanism in prokaryotes.** *BMC Genomics* 2010, **11:**102.
61. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.
62. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18:**1851-1858.
63. Hatem A, Bozdag D, Toland AE, Catalyurek UV: **Benchmarking short sequence mapping tools.** *BMC Bioinformatics* 2013, **14:**184.
64. Idury RM, Waterman MS: **A new algorithm for DNA sequence assembly.** *J Comput Biol* 1995, **2:**291-306.
65. Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly.** *Proc Natl Acad Sci U S A* 2001, **98:**9748-9753.
66. de la Bastide M, McCombie WR: **Assembling genomic DNA sequences with PHRAP.** *Curr Protoc Bioinformatics* 2007, **Chapter 11:**Unit11 14.
67. Mullikin JC, Ning Z: **The phusion assembler.** *Genome Res* 2003, **13:**81-90.
68. Sutton GG, White O, Adams MD, Kerlavage AR: **TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects.** *Genome Science and Technology* 1995, **1:**9-19.

69. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing.** *Genome Research* 2007, **17:**1697-1706.
70. Warren RL, Sutton GG, Jones SJ, Holt RA: **Assembling millions of short DNA sequences using SSAKE.** *Bioinformatics* 2007, **23:**500-501.
71. Bryant DW, Jr., Wong WK, Mockler TC: **QSRA: a quality-value guided de novo short read assembler.** *BMC Bioinformatics* 2009, **10:**69.
72. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, et al: **A whole-genome assembly of Drosophila.** *Science* 2000, **287:**2196-2204.
73. Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J: **De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer.** *Genome Research* 2008, **18:**802-809.
74. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data.** *Genome Res* 2009, **19:**1117-1123.
75. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, et al: **De novo assembly of human genomes with massively parallel short read sequencing.** *Genome Res* 2010, **20:**265-272.
76. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.** *J Comput Biol* 2012, **19:**455-477.
77. Chaisson MJ, Pevzner PA: **Short read fragment assembly of bacterial genomes.** *Genome Res* 2008, **18:**324-330.
78. Cerdeira LT, Carneiro AR, Ramos RTJ, de Almeida SS, D'Afonseca V, Schneider MPC, Baumbach J, Tauch A, McCulloch JA, Azevedo VAC, Silva A: **Rapid hybrid de novo assembly of a microbial genome using only short reads: Corynebacterium pseudotuberculosis I19 as a case study.** *Journal of Microbiological Methods* 2011, **86:**218-223.
79. Silva A, Schneider MPC, Cerdeira L, Barbosa MS, Ramos RTJ, Carneiro AR, Santos R, Lima M, D'Afonseca V, Almeida SS, et al: **Complete Genome Sequence of *Corynebacterium pseudotuberculosis* I19, a Strain Isolated from a Cow in Israel with Bovine Mastitis.** *J Bacteriol* 2011, **193:**323-324.
80. Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR, Chisholm SW: **Whole Genome Amplification and De novo Assembly of Single Bacterial Cells.** *PLoS One* 2009, **4:**e6864.
81. Farrer RA, Kemen E, Jones JDG, Studholme DJ: **De novo assembly of the Pseudomonas syringae pv. syringae B728a genome using Illumina/Solexa short sequence reads.** *Fems Microbiology Letters* 2009, **291:**103-111.
82. Nishito Y, Osana Y, Hachiya T, Popendorf K, Toyoda A, Fujiyama A, Itaya M, Sakakibara Y: **Whole genome assembly of a natto production strain *Bacillus subtilis* natto from very short read data.** *BMC Genomics* 2010, **11:**243.
83. Nagarajan H, Butler JE, Klimes A, Qiu Y, Zengler K, Ward J, Young ND, Methe BA, Palsson BO, Lovley DR, Barrett CL: **De Novo Assembly of the Complete Genome of an Enhanced Electricity-Producing Variant of Geobacter sulfurreducens Using Only Short Reads.** *PLoS One* 2010, **5:**e10922.
84. Bratcher HB, Bennett JS, Maiden MCJ: **Evolutionary and genomic insights into meningococcal biology.** *Future Microbiology* 2012, **7:**873-885.
85. Whitman WB, Coleman DC, Wiebe WJ: **Prokaryotes: the unseen majority.** *PNAS* 1998, **95:**6578-6583.
86. Staley JT: **The bacterial species dilemma and the genomic-phylogenetic species concept.** *Philos Trans R Soc Lond B Biol Sci* 2006, **361:**1899-1909.
87. Maiden MC, van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND: **MLST revisited: the gene-by-gene approach to bacterial genomics.** *Nat Rev Microbiol* 2013, **11:**728-736.
88. Bennett JS, Watkins ER, Jolley KA, Harrison OB, Maiden MC: **Identifying *Neisseria* species using the 50S ribosomal protein L6 (*rplF*) gene.** *J Clin Microbiol* 2014, **52:**1375-1381.

89. Konstantinidis KT, Tiedje JM: **Genomic insights that advance the species definition for prokaryotes.** *PNAS* 2005, **102:**2567-2572.
90. Caugant DA, Maiden MC: **Meningococcal carriage and disease - population biology and evolution.** *Vaccine* 2009, **27 Suppl 2:**B64-70.
91. Caugant DA, Tzanakaki G, Kriz P: **Lessons from meningococcal carriage studies.** *FEMS Microbiology Reviews* 2007, **31:**52-63.
92. Harrison OB, Claus H, Jiang Y, Bennett JS, Bratcher HB, Jolley KA, Corton C, Care R, Poolman JT, Zollinger WD, et al: **Description and nomenclature of *Neisseria meningitidis* capsule locus.** *Emerg Infect Dis* 2013, **19:**566-573.
93. Caugant DA, Bovre K, Gaustad P, Bryn K, Holten E, Høiby EA, Frøholm LO: **Multilocus genotypes determined by enzyme electrophoresis of *Neisseria meningitidis* isolated from patients with systemic disease and from healthy carriers.** *JGM* 1986, **132:**641-652.
94. Caugant DA, Mocca LF, Frasch CE, Frøholm LO, Zollinger WD, Selander RK: **Genetic structure of *Neisseria meningitidis* populations in relation to serogroup, serotype, and outer membrane protein pattern.** *J Bacteriol* 1987, **169:**2781-2792.
95. Tzanakaki G, Urwin R, Musilek M, Kriz P, Kremastinou J, Pangalis A, Blackwell CC, Maiden MC: **Phenotypic and genotypic approaches to characterization of isolates of *Neisseria meningitidis* from patients and their close family contacts.** *J Clin Microbiol* 2001, **39:**1235-1240.
96. Bratcher HB, Corton C, Jolley KA, Parkhill J, Maiden MC: **A gene-by-gene population genomics platform: *de novo* assembly, annotation and genealogical analysis of 108 representative *Neisseria meningitidis* genomes.** *BMC Genomics* 2014, **15:**1138.
97. Maiden MC, Ibarz-Pavon AB, Urwin R, Gray SJ, Andrews NJ, Clarke SC, Walker AM, Evans MR, Kroll JS, Neal KR, et al: **Impact of Meningococcal Serogroup C Conjugate Vaccines on Carriage and Herd Immunity.** *Journal of Infectious Diseases* 2008, **197:**737-743.
98. Schoen C, Tettelin H, Parkhill J, Frosch M: **Genome flexibility in *Neisseria meningitidis*.** *Vaccine* 2009, **27 Suppl 2:**B103-111.
99. Caugant DA, Bol P, Høiby EA, Zanen HC, Frøholm LO: **Clones of serogroup B *Neisseria meningitidis* causing systemic disease in the Netherlands, 1958-1986.** *Journal of Infectious Diseases* 1990, **162:**867-874.
100. Scholten RJ, Bijlmer HA, Poolman JT, Kuipers B, Caugant DA, van Alphen L, Dankert J, Valkenburg HA: **Meningococcal disease in The Netherlands, 1958-1990: a steady increase in the incidence since 1982 partially caused by new serotypes and subtypes of *Neisseria meningitidis*.** *Clinical Infectious Diseases* 1993, **16:**237-246.
101. Scholten RJPM, Poolman JT, Valkenburg HA, Bijlmer HA, Dankert J, Caugant DA: **Phenotypic and genotypic changes in a new clone complex of *Neisseria meningitidis* causing disease in The Netherlands, 1958-1990.** *Journal of Infectious Diseases* 1994, **169:**673-676.
102. Van Looveren M, Vandamme P, Hauchecorne M, Wijdooghe M, Carion F, Caugant DA, Goossens H: **Molecular epidemiology of recent belgian isolates of *Neisseria meningitidis* serogroup B.** *J Clin Microbiol* 1998, **36:**2828-2834.
103. Elias J, Schouls LM, van de Pol I, Keijzers WC, Martin DR, Glennie A, Oster P, Frosch M, Vogel U, van der Ende A: **Vaccine preventability of meningococcal clone, Greater Aachen Region, Germany.** *Emerg Infect Dis* 2010, **16:**465-472.
104. Russell JE, Urwin R, Gray SJ, Fox AJ, Feavers IM, Maiden MC: **Molecular epidemiology of meningococcal disease in England and Wales 1975-1995, before the introduction of serogroup C conjugate vaccines.** *Microbiology* 2008, **154:**1170-1177.
105. Gray SJ, Trotter CL, Ramsay ME, Guiver M, Fox AJ, Borrow R, Mallard RH, Kaczmarski EB: **Epidemiology of meningococcal disease in England and Wales 1993/94 to 2003/04: contribution and experiences of the Meningococcal Reference Unit.** *Journal of Medical Microbiology* 2006, **55:**887-896.
106. Yazdankhah SP, Kriz P, Tzanakaki G, Kremastinou J, Kalmusova J, Musilek M, Alvestad T, Jolley KA, Wilson DJ, McCarthy ND, et al: **Distribution of serogroups and genotypes among disease-associated and carried isolates of *Neisseria meningitidis* from the Czech Republic, Greece, and Norway.** *J Clin Microbiol* 2004, **42:**5146-5153.

107. Levy C, Taha MK, Weil Olivier C, Quinet B, Lecuyer A, Alonso JM, Aujard Y, Bingen E, Cohen R: **Association of meningococcal phenotypes and genotypes with clinical characteristics and mortality of meningitis in children.** *Pediatr Infect Dis J* 2010, **29:**618-623.
108. Elias J, Schouls LM, van de Pol I, Keijzers WC, Martin DR, Glennie A, Oster P, Frosch M, Vogel U, van der Ende A: **Vaccine Preventability of Meningococcal Clone, Greater Aachen Region, Germany.** *Emerg Infect Dis* 2010, **16:**464-472.
109. Harrison Lee H, Shutt Kathleen A, Schmink Susanna E, Marsh Jane W, Harcourt Brian H, Wang X, Whitney Anne M, Stephens David S, Cohn Amanda A, Messonnier Nancy E, Mayer Leonard W: **Population Structure and Capsular Switching of Invasive *Neisseria meningitidis* Isolates in the Pre–Meningococcal Conjugate Vaccine Era—United States, 2000–2005.** *The Journal of Infectious Diseases* 2010, **201:**1208-1224.
110. Martin DR, Walker SJ, Baker MG, Lennon DR: **New Zealand epidemic of meningococcal disease identified by a strain with phenotype B:4:P1.4.** *Journal of Infectious Diseases* 1998, **177:**497-500.
111. Dyet KH, Martin DR: **Clonal analysis of the serogroup B meningococci causing New Zealand's epidemic.** *Epidemiol Infect* 2006, **134:**377-383.
112. Trotter CL, Chandra M, Cano R, Larrauri A, Ramsay ME, Brehony C, Jolley KA, Maiden MC, Heuberger S, Frosch M: **A surveillance network for meningococcal disease in Europe.** *FEMS Microbiology Reviews* 2007, **31:**27-36.
113. O'Connor L, Ward M, Bennett D, Mulhall R, O'Lorcain P, Cunney R, McDermott R, Neville E, Heslin J, FitzGerald R, et al: **A prolonged outbreak of invasive meningococcal disease in an extended Irish Traveller family across three Health Service Executive (HSE) areas in Ireland, 2010 to 2013.** *Eurosurveillance* 2015, **20**.
114. Jolley KA, Hill DM, Bratcher HB, Harrison OB, Feavers IM, Parkhill J, Maiden MC: **Resolution of a meningococcal disease outbreak from whole genome sequence data with rapid web-based analysis methods.** *J Clin Microbiol* 2012, **50:**3046-3053.
115. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: **MEGA6: Molecular Evolutionary Genetics Analysis version 6.0.** *Mol Biol Evol* 2013, **30:**2725-2729.
116. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J: **ACT: the Artemis Comparison Tool.** *Bioinformatics* 2005, **21:**3422-3423.
117. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16:**944-945.
118. Roberts RJ: **Restriction enzymes and their isoschizomers.** *Nucleic Acids Res* 1990, **18 Suppl:**2331-2365.
119. Drozdz M, Piekarowicz A, Bujnicki JM, Radlinska M: **Novel non-specific DNA adenine methyltransferases.** *Nucleic Acids Res* 2012, **40:**2119-2130.
120. Roberts E, Sethi A, Montoya J, Woese CR, Luthey-Schulten Z: **Molecular signatures of ribosomal evolution.** *PNAS* 2008, **105:**13953-13958.
121. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony CM, Colles FM, Wimalarathna HM, Harrison OB, Sheppard SK, Cody AJ, Maiden MC: **Ribosomal Multi-Locus Sequence Typing: universal characterization of bacteria from domain to strain.** *Microbiology* 2012, **158:**1005-1015.
122. Read DS, Woodcock DJ, Strachan NJ, Forbes KJ, Colles FM, Maiden MC, Clifton-Hadley F, Ridley A, Vidal A, Rodgers J, et al: **Evidence for phenotypic plasticity amongst multi-host *Campylobacter jejuni* and *C. coli* lineages using ribosomal MLST and Raman spectroscopy.** *Appl Environ Microbiol* 2013, **79:**965-973.
123. Forsythe SJ, Dickins B, Jolley KA: **Cronobacter, the emergent bacterial pathogen Enterobacter sakazakii comes of age; MLST and whole genome sequence analysis.** *BMC Genomics* 2014, **15:**1121.
124. Facey PD, Meric G, Hitchings MD, Pachebat JA, Hegarty MJ, Chen X, Morgan LV, Hoeppner JE, Whitten MM, Kirk WD, et al: **Draft Genomes, Phylogenetic Reconstruction, and Comparative Genomics of Two Novel Cohabiting Bacterial Symbionts Isolated from Frankliniella occidentalis.** *Genome Biol Evol* 2015, **7:**2188-2202.

125. Gorringe AR, Pajon R: **Bexsero: a multicomponent vaccine for prevention of meningococcal disease.** *Hum Vaccin Immunother* 2012, **8:**174-183.
126. Bille E, Zahar JR, Perrin A, Morelle S, Kriz P, Jolley KA, Maiden MC, Dervin C, Nassif X, Tinsley CR: **A chromosomally integrated bacteriophage in invasive meningococci.** *Journal of Experimental Medicine* 2005, **201:**1905-1913.
127. Masignani V, Giuliani MM, Tettelin H, Comanducci M, Rappuoli R, Scarlato V: **Mu-like prophage in serogroup B *Neisseria meningitidis* coding for surface-exposed antigens.** *Infect Immun* 2001, **69:**2580-2588.
128. Agarwal S, Mishra NK, Bhatnagar S, Bhatnagar R: **PemK Toxin of Bacillus anthracis Is a Ribonuclease AN INSIGHT INTO ITS ACTIVE SITE, STRUCTURE, AND FUNCTION.** *Journal of Biological Chemistry* 2010, **285:**7254-7270.
129. Gazit E, Sauer RT: **The Doc toxin and Phd antidote proteins of the bacteriophage P1 plasmid addiction system form a heterotrimeric complex.** *J Biol Chem* 1999, **274:**16813-16818.
130. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, et al: **Evolution of MRSA during hospital transmission and intercontinental spread.** *Science* 2010, **327:**469-474.
131. Cody AJ, McCarthy ND, Jansen van Rensburg M, Isinkaye T, Bentley S, Parkhill J, Dingle KE, Bowler IC, Jolley KA, Maiden MC: **Real-time genomic epidemiology of human *Campylobacter* isolates using whole genome multilocus sequence typing.** *J Clin Microbiol* 2013, **51:**2526-2534.
132. Eppinger M, Pearson T, Koenig SS, Pearson O, Hicks N, Agrawal S, Sanjar F, Galens K, Daugherty S, Crabtree J, et al: **Genomic epidemiology of the Haitian cholera outbreak: a single introduction followed by rapid, extensive, and continued spread characterized the onset of the epidemic.** *MBio* 2014, **5:**e01721.
133. Walker TM, Clp CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, et al: **Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study.** *Lancet Infectious Diseases* 2013, **13:**137-146.
134. Harrison LH, Trotter CL, Ramsay ME: **Global epidemiology of meningococcal disease.** *Vaccine* 2009, **27:**B51-B63.
135. Brehony C, Trotter CL, Ramsay ME, Chandra M, Jolley KA, van der Ende A, Carion F, Berthelsen L, Hoffmann S, Harðardóttir H, et al: **Differential age distribution of disease-associated meningococcal lineages-Implications for vaccine development.** *Clinical and Vaccine Immunology* 2014.
136. Bijlsma MW, Brouwer MC, Spanjaard L, Van de Beek D, Van der Ende A: **A decade of herd protection after introduction of meningococcal serogroup C conjugate vaccination.** *Clinical Infectious Disases* 2014, **In the press**.
137. Didelot X, Urwin R, Maiden MC, Falush D: **Genealogical typing of *Neisseria meningitidis*.** *Microbiology* 2009, **155:**3176-3186.
138. Jolley KA, Maiden MC: **Using MLST to study bacterial variation: prospects in the genomic era.** *Future Microbiol* 2014, **9:**623-630.
139. Schoen C, Blom J, Claus H, Schramm-Gluck A, Brandt P, Muller T, Goesmann A, Joseph B, Konietzny S, Kurzai O, et al: **Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in *Neisseria meningitidis*.** *PNAS* 2008, **105:**3473-3478.
140. Joseph B, Schwarz RF, Linke B, Blom J, Becker A, Claus H, Goesmann A, Frosch M, Muller T, Vogel U, Schoen C: **Virulence evolution of the human pathogen *Neisseria meningitidis* by recombination in the core and accessory genome.** *PLoS One* 2011, **6:**e18441.
141. Gray RR, Tatem AJ, Johnson JA, Alekseyenko AV, Pybus OG, Suchard MA, Salemi M: **Testing Spatiotemporal Hypothesis of Bacterial Evolution Using Methicillin-Resistant Staphylococcus aureus ST239 Genome-wide Data within a Bayesian Framework.** *Mol Biol Evol* 2011, **28:**1593-1603.

142. Snyder LA, Butcher SA, Saunders NJ: **Comparative whole-genome analyses reveal over 100 putative phase-variable genes in the pathogenic *Neisseria* spp.** *Microbiology* 2001, **147:**2321-2332.
143. Watkins ER, Maiden MC: **Persistence of hyperinvasive meningococcal strain types during global spread as recorded in the PubMLST database.** *PLoS ONE* 2012, **7:**e45349.
144. Vogel U, Taha MK, Vazquez JA, Findlow J, Claus H, Stefanelli P, Caugant DA, Kriz P, Abad R, Bambini S, et al: **Predicted strain coverage of a meningococcal multicomponent vaccine (4CMenB) in Europe: a qualitative and quantitative assessment.** *Lancet Infectious Diseases* 2013, **13:**416-425.
145. Gerdes K, Christensen SK, Lobner-Olesen A: **Prokaryotic toxin-antitoxin stress response loci.** *Nat Rev Microbiol* 2005, **3:**371-382.
146. Zhang J, Zhang Y, Zhu L, Suzuki M, Inouye M: **Interference of mRNA function by sequence-specific endoribonuclease PemK.** *J Biol Chem* 2004, **279:**20678-20684.
147. Hazan R, Sat B, Reches M, Engelberg-Kulka H: **Postsegregational killing mediated by the P1 phage "addiction module" phd-doc requires the Escherichia coli programmed cell death system mazEF.** *J Bacteriol* 2001, **183:**2046-2050.
148. Caugant DA, Høiby EA, Rosenqvist E, Frøholm LO, Selander RK: **Transmission of *Neisseria meningitidis* among asymptomatic military recruits and antibody analysis.** *Epidemiology and Infection* 1992, **109:**241-253.
149. Goldschneider I, Gotschlich EC, Artenstein MS: **Human Immunity to the Meningococcus. II. Development of Natural Immunity.** *Journal of Experimental Medicine* 1969, **129:**1327-1348.
150. Cartwright KAV, Stuart JM, Jones DM, Noah ND: **The Stonehouse survey: nasopharyngeal carriage of meningococci and *Neisseria lactamica*.** *Epidemiology and Infection* 1987, **99:**591-601.
151. Stephens DS: **Uncloaking the meningococcus: dynamics of carriage and disease.** *Lancet* 1999, **353:**941-942.
152. Claus H, Maiden MC, Wilson DJ, McCarthy ND, Jolley KA, Urwin R, Hessler F, Frosch M, Vogel U: **Genetic analysis of meningococci carried by children and young adults.** *Journal of Infectious Diseases* 2005, **191:**1263-1271.
153. Trotter CL, Gay NJ, Edmunds WJ: **The natural history of meningococcal carriage and disease.** *Epidemiology and Infection* 2006, **134:**556-566.
154. Yazdankhah SP, Caugant DA: ***Neisseria meningitidis*: an overview of the carriage state.** *Journal of Medical Microbiology* 2004, **53:**821-832.
155. Jolley KA, Kalmusova J, Feil EJ, Gupta S, Musilek M, Kriz P, Maiden MC: **Carried meningococci in the Czech Republic: a diverse recombining population.** *J Clin Microbiol* 2000, **38:**4492-4498.
156. Caugant DA, Kristiansen BE, Frøholm LO, Bovre K, Selander RK: **Clonal diversity of *Neisseria meningitidis* from a population of asymptomatic carriers.** *Infection and Immunity* 1988, **56:**2060-2068.
157. Caugant DA, Høiby EA, Magnus P, Scheel O, Hoel T, Bjune G, Wedege E, Eng J, Frøholm LO: **Asymptomatic carriage of *Neisseria meningitidis* in a randomly sampled population.** *J Clin Microbiol* 1994, **32:**323-330.
158. Davila S, Wright VJ, Khor CC, Sim KS, Binder A, Breunis WB, Inwald D, Nadel S, Betts H, Carrol ED, et al: **Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease.** *Nat Genet* 2010, **42:**772-776.
159. Kugelberg E, Gollan B, Tang CM: **Mechanisms in *Neisseria meningitidis* for resistance against complement-mediated killing.** *Vaccine*

Complement: The First Barrier of Innate Immunity 2008, **26:**I34-I39.
160. Ibarz Pavón AB: **Molecular Characterization and Evolutionary Analysis of a Carried Meningococcal Population Sample.** *D.Phil.* University of Oxofrd, Department of Zoology; 2007.
161. Campbell H, Borrow R, Salisbury D, Miller E: **Meningococcal C conjugate vaccine: The experience in England and Wales.** *Vaccine* 2009, **27:**B20-B29.

162. Maiden MCJ, Spratt BG: **Meningococcal conjugate vaccines: new opportunities and new challenges.** *Lancet* 1999, **354:**615-616.
163. Claus H, Stummeyer K, Batzilla J, Muhlenhoff M, Vogel U: **Amino acid 310 determines the donor substrate specificity of serogroup W-135 and Y capsule polymerases of Neisseria meningitidis.** *Mol Microbiol* 2009, **71:**960-971.
164. Huson DH: **SplitsTree: analyzing and visualizing evolutionary data.** *Bioinformatics* 1998, **14:**68-73.
165. Moulton V: **SplitsTree: A network-based tool for exploring evolutionary relationships in molecular data.** In *The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny.* Edited by Salemi M, Vandamme AM. Cambridge, UK: Cambridge University Press; 2003: 312-328
166. **Inkscape: A Vector Drawing Tool.**
167. Neal KR: **Changing carriage rate of *Neisseria meningitidis* among university students during the first week of term: cross sectional study.** *Bmj* 2000, **320:**846-849.
168. Ala'Aldeen DAA, Neal KR, Ait-Tahar K, Nguyen-Van-Tam JS, English A, Falla TJ, Hawkey PM, Slack RCB: **Dynamics of Meningococcal Long-term Carriage among University Students and Their Implications for Mass Vaccination.** *J Clin Microbiol* 2000, **38:**2311-2316.
169. Trotter CL, Gay NJ: **Analysis of longitudinal bacterial carriage studies accounting for sensitivity of swabbing: an application to *Neisseria meningitidis*.** *Epidemiology and infection* 2003, **130:**201-205.
170. Gupta S, Anderson RM: **Population structure of pathogens: the role of immune selection.** *Parasitology Today* 1999, **15:**497-501.
171. Buckee CO, Jolley K, Recker M, Penman B, Kriz P, Gupta S, Maiden MC: **Role of selection in the emergence of lineages and the evolution of virulence in *Neisseria meningitidis*.** *PNAS* 2008, **105:**15082-15087.
172. Ali O, Aseffa A, Bedru A, Lemma T, Moti T, Worku A, Xabher HG, Yamuah L, Boukary RM, Collard JM, et al: **Meningococcal carriage in the African meningitis belt.** *Tropical Medicine & International Health* 2013, **18:**968-978.
173. Glitza IC, Ehrhard I, Muller-Pebody B, Reintjes R, Breuer T, Ammon A, Sonntag HG: **Longitudinal study of meningococcal carrier rates in teenagers.** *International Journal of Hygiene and Environmental Health* 2008, **211:**263-272.
174. MacLennan J, Kafatos G, Neal K, Andrews N, Cameron JC, Roberts R, Evans MR, Cann K, Baxter DN, Maiden MC, Stuart JM: **Social behavior and meningococcal carriage in British teenagers.** *Emerg Infect Dis* 2006, **12:**950-957.
175. Stuart JM, Cartwright KA, Robinson PM, Noah ND: **Effect of smoking on meningococcal carriage.** *Lancet* 1989, **2:**723-725.
176. El Ahmer OR, Essery SD, Saadi AT, Raza MW, Ogilvie MM, Weir DM, Blackwell CC: **The effect of cigarette smoke on adherence of respiratory pathogens to buccal epithelial cells.** *FEMS Immunol Med Microbiol* 1999, **23:**27-36.
177. Arcavi L, Benowitz NL: **Cigarette smoking and infection.** *Arch Intern Med* 2004, **164:**2206-2216.
178. Brehony C, Jolley KA, Maiden MC: **Multilocus sequence typing for global surveillance of meningococcal disease.** *FEMS Microbiology Reviews* 2007, **31:**15-26.
179. Deasy AM, Guccione E, Dale AP, Andrews N, Evans CM, Bennett JS, Bratcher HB, Maiden MC, Gorringe AR, Read RC: **Nasal Inoculation of the Commensal Neisseria lactamica Inhibits Carriage of Neisseria meningitidis by Young Adults: A Controlled Human Infection Study.** *Clin Infect Dis* 2015.
180. Parkhill J: **Time to remove the model organism blinkers.** *Trends Microbiol* 2008, **16:**510-511.
181. Canchaya C, Proux C, Fournous G, Bruttin A, Brussow H: **Prophage genomics.** *Microbiol Mol Biol Rev* 2003, **67:**238-276, table of contents.
182. Canchaya C, Fournous G, Brussow H: **The impact of prophages on bacterial chromosomes.** *Mol Microbiol* 2004, **53:**9-18.
183. Stollenwerk N, Maiden MC, Jansen VA: **Diversity in pathogenicity can cause outbreaks of meningococcal disease.** *PNAS* 2004, **101:**10229-10234.

184. Hill DMC, Lucidarme J, Gray SJ, Newbold LS, Ure R, Brehony C, Harrison OB, Bray JE, Jolley KA, Bratcher HB, et al: **Genomic epidemiology of age-associated meningococcal lineages in national surveillance: an observational cohort study** *Lancet Infectious Diseases* 2015, **15:**1420-1428.
185. **PubMLST website - Publicly-accessible MLST databases and software.** [http://pubmlst.org]
186. **Molecular Evolutionary Genetics Analysis** [http://evolgen.biol.metro-u.ac.jp/MEGA/]
187. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D: **Tablet--next generation sequence assembly visualization.** *Bioinformatics* 2010, **26:**401-402.
188. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D: **Using Tablet for visual exploration of second-generation sequencing data.** *Brief Bioinform* 2013, **14:**193-202.
189. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
190. Consortium TU: **Activities at the Universal Protein Resource (UniProt).** *Nucleic Acids Research* 2014, **42:**D191-D198.