Characterization of capsule genes in non-pathogenic *Neisseria* species

Marianne Elizabeth Alexandra Clemence, Martin Christopher James Maiden* and Odile Barbara Harrison

**Abstract**

The genus *Neisseria* comprises a diverse group of commensal bacteria, which typically colonize the mucosal surfaces of humans and other animals. *Neisseria meningitidis*, the meningococcus, is notable for its potential to cause invasive meningococcal disease (IMD) in humans; however, IMD is comparatively rare, and meningococci normally colonize the nasopharynx asymptomatically. Possession of a polysaccharide capsule has been shown to be a prerequisite for disease in almost all IMD cases, and was previously considered unique to *N. meningitidis*, and potentially acquired by horizontal genetic transfer (HGT). Nevertheless, the capsule must also have some role in asymptomatic colonization and/or transmission, consistent with the existence of six non-disease-associated meningococcal capsule serogroups. In this study, full complements of putative capsule genes were identified in non-pathogenic *Neisseria* species, including *Neisseria subflava* and *Neisseria elongata*. These species contained genes for capsule transport and translocation homologous to those of *N. meningitidis*, as well as novel putative capsule synthesis genes. Phylogenetic analyses were consistent with the proposal that these genes were acquired by the meningococcus through HGT. In contrast with previous evolutionary models, however, the most parsimonious explanation of these data was that capsule transport genes had been lost in the common ancestor of the meningococcus, gonococcus, and their close relatives, and then reacquired by some meningococci. The most likely donor of the meningococcal transport genes was another *Neisseria* species.

**DATA SUMMARY**

This work made use of sequencing data obtained from the pubMLST database (https://pubmlst.org/neisseria), the NCTC 3000 project (https://www.sanger.ac.uk/resources/downloads/bacteria/nctc/) and GenBank (https://www.ncbi.nlm.nih.gov/genbank/index.html). A comprehensive list of strain IDs, metadata and accession numbers can be found on FigShare at https://doi.org/10.6084/m9.figshare.6016112.v1. Annotated pubMLST isolates can be searched in the pubMLST isolate database, and allele sequences retrieved using the NEIS numbers from the pubMLST sequence database.

**INTRODUCTION**

The genus *Neisseria* is a diverse group of Gram-negative bacteria, many of which are asymptomatic colonizers of the mucosal surfaces of animals and man [1]. In humans, they have been isolated from the mouth, nose, throat and urogenital tract, but whilst many *Neisseria* species belong to the human oral microbiota, research has focused on those associated with disease: *Neisseria gonorrhoeae* and *Neisseria meningitidis*. In common with many other *Neisseria* species, *N. meningitidis* usually colonizes the nasopharynx asymptomatically; however, it occasionally invades the bloodstream, leading to life-threatening invasive meningococcal disease (IMD), comprising meningitis and/or septicaemia [2]. In contrast, there are very few case reports of other *Neisseria* species causing invasive disease. Compared to asymptomatic colonization, IMD is an extremely rare transmission-terminating event, associated with particular meningococcal genotypes that normally express a polysaccharide capsule [3].

Capsules are associated with virulence in several human pathogens, including *Escherichia coli*, *Haemophilus influenzae* and *Klebsiella pneumoniae* [4–6]. A number of successful vaccines have been developed that target capsular antigens, for example the polysaccharides forming the capsules of the meningococcal serogroups A, C, W and Y [7]. Capsules can aid evasion of immune responses, including the complement system and phagocytosis by macrophages, facilitating...
persistence in the bloodstream [8, 9]; however, capsules have also been identified in free-living bacteria and symbionts [10, 11]. In both N. meningitidis and other species, association with disease is often restricted to a subset of capsular groups or types [2, 5, 6], indicative that in general the capsule confers benefits during transmission or protects the bacteria from local inflammation in the nasopharynx [12, 13].

In N. meningitidis, the capsule is produced via ABC transporter-dependent polymerization, whereby synthesis and polymerization of the polysaccharide take place at the bacterial inner membrane, prior to transport across the membrane and translocation to the cell surface [14]. These processes are encoded by genes located in the cps locus [15], which is functionally divided into several contiguous regions. Region A contains genes involved in capsule synthesis, in particular glycosyltransferases and capsule polymerases, but also other proteins involved in additional capsule modifications, and sometimes insertion sequences are present [15, 16]. This region is highly variable, with 12 known variants corresponding to the 12 meningococcal serogroups. Of these, only six (A, B, C, W, X and Y) are associated with disease. Regions B and C are composed of the genes ctreF and ctrABCD, respectively, and are required for capsule translocation and transport. These regions are well conserved throughout N. meningitidis, unlike region A [15]. Region D of the cps contains the genes rfbABC and galE, which are thought to play a role in LPS synthesis [17]. A duplication of region D, containing a truncated galE, is designated region D’ [18]. Finally, region E contains the gene tex and two pseudo cytosine methyltransferases of unknown function. Although regions D, D’ and E are not directly involved in capsule synthesis or transport, they are generally considered part of cps due to their location within the locus. Isolates that do not contain regions A, B and C are described as capsule null, and instead possess a distinct 113–118 bp sequence located between regions D and E, the capsule null locus (cnl) [19]. The cnl locus has also been identified in N. gonorrhoeae and the non-pathogenic Neisseria (NPN) species Neisseria lactamica [19], and no encapsulated isolates from these species have been described.

Whilst a number of putative virulence genes have been found in NPN species [20], the capsule has been considered to be unique to the meningococcus [21], possibly acquired in a horizontal genetic transfer (HGT) event that gave rise to the potentially pathogenic variants of N. meningitidis [3, 22]. In this study, capsule genes have been identified and characterized in NPN species, which typically do not cause disease. These results have implications for understanding the acquisition of capsule in N. meningitidis.

METHODS

Isolate collection and species definitions

Whole-genome sequence (WGS) data from Neisseria isolates were obtained from pubmlst.org/neisseria, which is hosted on the Bacterial Isolate Genome Sequence Database (BIGSdb) genomics platform [23]. At the time of writing, the database contained WGS data from >13 000 Neisseria isolates, 235 of which were from NPN species. The pubMLST sequence database contains defined Neisseria loci and allele sequences, with each locus assigned a unique NEIS number. Isolates can be annotated with NEIS loci automatically or manually through a BLAST-based process, and new alleles are assigned an arbitrary allele number. Most of the WGSs in pubMLST are high-quality draft genomes, with sequencing reads assembled into approximately 100–300 individual contigs.

A total of 20 species had been defined in the pubmlst.org/neisseria database using the universal and high-resolution ribosomal multi-locus sequence typing (rMLST) approach [24]. These species included the human nasopharyngeal commensals Neisseria polysaccharea (16 isolates), ‘Neisseria bergeri’ (1 isolate), N. lactamica (140 isolates), Neisseria cinerea (15 isolates), Neisseria subflava (19 isolates), Neisseria oralis (4 isolates), Neisseria mucosa (10 isolates), Neisseria elongata subsp. (4 isolates) and Neisseria bacilloformis (4 isolates), and the pathogens N. meningitidis and N. gonorrhoeae, along with several species isolated from animals or animal-bite wounds. N. meningitidis region A genes were sourced from isolate sequences for 0106/93 and 0084/93 published in pubMLST, and 29 043 published in GenBank (accession no. HF562984.1).

All NPN isolates in pubMLST were surveyed for the presence of cps genes that putatively encoded a polysaccharide capsule. All isolates found to contain a cps were further characterized. Isolates from additional species were included in phylogenetic analyses. Capsule sequence data from other genera were obtained from GenBank. Reference genomes were obtained from either GenBank (http://www.ncbi.nlm.nih.gov/genbank/index.html) or the NCTC 3000 project (http://www.sanger.ac.uk/resources/downloads/bacteria/nctc/).

IMPACT STATEMENT

Neisseria meningitidis, whilst normally a harmless commensal of the human nasopharynx, can cause invasive meningococcal disease (IMD), comprising meningitis and/or septicaemia. Expression of a polysaccharide capsule is essential for IMD, but must also be involved in asymptomatic colonization. The capsule has been considered a virulence factor unique to N. meningitidis; however, here a full complement of homologous putative capsule genes was identified in non-pathogenic Neisseria (NPN) species. NPN species are important members of the human nasopharyngeal microbiota, as well as coexisting with the meningococcus in the nasopharynx. The results inform debate about the acquisition of capsule by the meningococcus, an important step in the emergence of pathogenic potential.
Annotation of cps in NPN isolates

The BIGSdb software enables BLAST searches of protein or nucleotide sequences against genomes contained within pubMLST. Region B and C cps genes ctrABCDFE, which are involved in capsule transport in the meningococcus, had previously been defined in the pubMLST.org/neisseria sequence database as NEIS0055, NEIS0056, NEIS0057, NEIS0058, NEIS0066 and NEIS0067, respectively. For each gene, the amino acid sequence of allele 1 was used as a pBLAST query against all available NPN isolates within the database for which WGSs were available. Candidate genes were annotated in Artemis [25] and G+C content determined in MEGA 7 [26]. The same approach was also used to annotate region D and E genes, and any other relevant genes, where necessary. Annotations were uploaded as novel alleles in pubMLST.

The proposed cps regions of NPN isolates were further annotated in Artemis [25]. ORFs adjacent to proposed region C genes were queried against the National Center for Biotechnology Information (NCBI) RefSeq protein database using pBLAST and the Pfam database [27], as well as the pubMLST sequence database. Support for putative region A genes was based on homology to capsule synthesis genes from N. meningitidis and/or other bacterial species, or at least for gene products consistent with a function in capsule synthesis, such as glycosyltransferases. Additional guidance was based on comparisons of syntenic with N. meningitidis and between NPN isolates and species. In this way, ORFs that did not contain significant homology to previously described capsule synthesis genes, or previously described capsule synthesis-like genes, could only be included in a putative region A if they were flanked by ORFs that did have significant homology. Region A candidates were also queried against the non-redundant sequences of the CAZy database [28], which contains data on carbohydrate-active enzyme families, using the CAZymes Analysis Toolkit [29]. CAZy families were predicted using both sequence similarity and Pfam rule-based annotations, with an E value threshold of $1 \times 10^{-10}$ and bit score threshold of 55. Annotations were uploaded as novel NEIS loci and alleles in pubMLST. The organization of the cps found in NPN isolates was compared with that of the meningococcus and visualized using genoPlotR [30].

Identification of homologous candidate region A genes

Potentially homologous genes shared by isolates from the same species were identified based on gene order, sequence length and predicted function. The nucleotide sequences for each gene were aligned in Clustal Omega [31], and pairwise identity matrices (PIMs) were generated using Clustal 2.1. Suspected homology between different species was identified in the same way and investigated using pairwise comparisons of amino acid sequences generated by Clustal, and pairwise tBLASTx comparisons between the proposed region A of each species were made using the Artemis Comparison Tool [32] and visualized using genoPlotR.

Phylogenetic analyses

A recombination-corrected phylogenetic tree of Neisseria isolates, with Moraxella catarrhalis as an outgroup, was generated based on rMLST loci [24]. The nucleotide sequences of 51 of the 53 genes that constitute the protein subunits of the ribosome (excluding rpmE and rpmF, as they are paralogous in some Neisseria), were extracted from each isolate using the BIGSdb genome comparator, and aligned using MAFFT [33]. A maximum-likelihood (ML) tree was generated in PhyML v3.1 [34] using the GTR+I+G substitution model, determined to be the best-fit model by jModelTest v2.1.10 [35], with 100 bootstrap replicates. The tree was corrected for recombination in ClonalFrameML [36], and rendered and annotated using the ETE 3 toolkit [37]. The phylogeny included all isolates belonging to those species in which capsule genes were identified, and representatives of N. cinerea, N. lactamica, ‘N. bergeri’, N. polysaccharea and N. gonorrhoeae, in none of which were capsule genes identified, and representatives of N. meningitidis.

Where present, region B and C genes were also extracted from isolates in the dataset. Additionally, homologous capsule genes, as determined by NCBI BLAST queries, were extracted from isolates of Mannheimia haemolytica, Actinobacillus pleuropneumoniae, Actinobacillus suis, Bibersteinia trehalosi, H. influenzae, Pasteurella multocida and Kingella kingae. Amino acid sequences from all species were concatenated, aligned with MUSCLE [38] and trimmed in trimAL v1.2 [39] to remove columns with gaps in more than 20% of sequences or with a similarity score lower than 0.001. An unrooted ML tree was generated in PhyML using the LG+I+G substitution model, selected to be the best fit by ProtTest v3.4.2 [40]. The tree was rendered in FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

RESULTS

Identification of cps in NPN species

Capsule gene homologues were identified in isolates from a total of 13 Neisseria species contained within the pubmlst.org/neisseria database [23], including N. bacilliformis (4 of 4 isolates); N. elongata subsp. (2 of 4 isolates); Neisseria musculi (1 isolate); Neisseria dentiae (1 isolate); Neisseria animaloris (1 isolate); Neisseria zoodegmatis (1 isolate); Neisseria weaveri (1 isolate); Neisseria canis (1 isolate); Neisseria wadsworthii (1 isolate); Neisseria animalis (1 isolate); N. oralis (4 of 4 isolates); N. mucosa (1 of 9 isolates); and N. subflava (17 of 19 isolates). Regions A, B and C were annotated in isolates of all of these species (Fig. 1), with the exception of N. wadsworthii and N. animaloris, in which region A was interrupted in the genome assembly.

With the exception of 11 N. subflava isolates, isolates in which capsule genes were identified were found to contain homologues for all six of the N. meningitidis region B and C genes (Table S1, available in the online version of this article), sharing 52–99% amino acid sequence identity with the relevant query. Alignment length covered at least 92% of...
the relevant query, except the homologues of ctrF in CCUG 50858 (85%) and NJ9703 (71%). This reduced query coverage was due to an incomplete gene located at the end of a contig in WGS data for isolate CCUG 50858, and a frame-shift mutation in the sequences for isolate NJ9703.

Annotation using Artemis [25] showed that ctrABCD genes of region C were contiguous and in the same order as those found in N. meningitidis. The ctrE and ctrF genes were contiguous in all isolates except those from N. canis and N. bacilliformis, although in these two species ctrE was adjacent to region C. The remaining 11 N. subflava isolates were found to contain homologues for one or both of ctrE and ctrF, but no region C genes were identified (Table S1).

**Annotation of putative novel region A genes in NPN species**

With the exception of N. bacilliformis, isolates that contained a complete set of region B and C genes possessed a region adjacent to region C that had not been defined in the pubMLST.org/neisseria database at the time of analysis. A total of 59 ORFs were annotated as putative region A genes. Of these, 33 were homologous, with capsule genes described in both N. meningitidis and in other non-Neisseria species, including A. pleuropneumoniae and Mannheimia haemolytica, and a further 16 were homologous with genes commonly involved in capsule synthesis, including glycosyltransferases and acetyltransferases (Table 1). Remaining genes were included as part of a putative region A based on synteny. In many cases, genes had been annotated previously in the relevant species in RefSeq, albeit not directly attributed to a NPN cps, and so BLAST hits were almost identical to the BLAST query. A total of 40 of the region A candidates belonged to a CAZy [28] glycosyltransferase family, based on either sequence similarity of Pfam rule-based annotation (Table 1). With the exception of GT61 (to which a gene in N. weaveri belonged), these were all glycosyltransferase families to which N. meningitidis region A genes belong.

In all but four species, the proposed region A was flanked on both sides by region B, region C, region D or some other gene with an unrelated function. In N. animalis, between the last putative capsule synthesis gene and galE was an ORF predicted to belong to the DUF1016 family; since this family is predicted to code for nuclease genes, it was considered unlikely to have a role in capsule synthesis. In N. elongata subsp. elongata, an IS565 insertion sequence was identified between region A and galE, but no evidence was found to suggest that this was interrupting a putative capsule synthesis gene. In N. canis, the proposed region A was preceded by an IS481 insertion sequence, but again no evidence was found showing that a putative capsule synthesis gene had been interrupted; ORFs adjacent to the insertion sequence had been previously annotated.

**BLAST** querying the N. bacilliformis isolates with all novel region A candidates identified putative region A genes homologous to those found in N. elongata subsp.
<table>
<thead>
<tr>
<th>Species</th>
<th>NEIS locus</th>
<th>Product of highest non-hypothetical pBLAST hit/CAZy family</th>
<th>Identity (%)</th>
<th>E value</th>
<th>G+C (mol%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>N. animalis</em> (CCUG 808)</td>
<td>2173</td>
<td>Glycerol-3-phosphate cytidylytransferase GT9*</td>
<td>99</td>
<td>0</td>
<td>43.6</td>
</tr>
<tr>
<td></td>
<td>2947</td>
<td>Phosphotransferase</td>
<td>98</td>
<td>0</td>
<td>43.8</td>
</tr>
<tr>
<td></td>
<td>2939</td>
<td>CDP-glycerol–glycerophosphate glycerophosphotransferase GT4</td>
<td>41</td>
<td>6&lt;sup&gt;−31&lt;/sup&gt;</td>
<td>44.2</td>
</tr>
<tr>
<td></td>
<td>2940</td>
<td>CDP-glycerol–glycerophosphate glycerophosphotransferase GT2</td>
<td>33</td>
<td>2&lt;sup&gt;−37&lt;/sup&gt;</td>
<td>40.6</td>
</tr>
<tr>
<td></td>
<td>2966</td>
<td>Galactosyl transferase GT32</td>
<td>33</td>
<td>3&lt;sup&gt;−33&lt;/sup&gt;</td>
<td>41.1</td>
</tr>
<tr>
<td></td>
<td>2967</td>
<td>α-/β-Hydrolase</td>
<td>31</td>
<td>5&lt;sup&gt;−44&lt;/sup&gt;</td>
<td>45.8</td>
</tr>
<tr>
<td><em>N. bacilliformis</em> (CCUG 38158)</td>
<td>2173</td>
<td>Glycerol-3-phosphate cytidylytransferase GT9*</td>
<td>100</td>
<td>1&lt;sup&gt;−94&lt;/sup&gt;</td>
<td>37.6</td>
</tr>
<tr>
<td></td>
<td>2938</td>
<td>Phosphotransferase</td>
<td>47</td>
<td>7&lt;sup&gt;−118&lt;/sup&gt;</td>
<td>36.7</td>
</tr>
<tr>
<td></td>
<td>2939</td>
<td>CDP-glycerol–glycerophosphate glycerophosphotransferase GT4</td>
<td>41</td>
<td>3&lt;sup&gt;−24&lt;/sup&gt;</td>
<td>37.9</td>
</tr>
<tr>
<td></td>
<td>2940</td>
<td>CDP-glycerol–glycerophosphate glycerophosphotransferase GT2</td>
<td>36</td>
<td>2&lt;sup&gt;−36&lt;/sup&gt;</td>
<td>39.4</td>
</tr>
<tr>
<td><em>N. cantis</em> (CCUG 56775 T)</td>
<td>2157</td>
<td>UDP-N-acetylglucosamine 2-epimerase GT28*</td>
<td>74</td>
<td>0</td>
<td>40.3</td>
</tr>
<tr>
<td></td>
<td>2737</td>
<td>UDP-N-acetyl-D-mannosamine dehydrogenase GT2*</td>
<td>87</td>
<td>0</td>
<td>43.4</td>
</tr>
<tr>
<td></td>
<td>2738</td>
<td>Glycosyl transferase GT2</td>
<td>40</td>
<td>1&lt;sup&gt;−92&lt;/sup&gt;</td>
<td>37.1</td>
</tr>
<tr>
<td></td>
<td>2739</td>
<td>Glycosyl transferase GT4</td>
<td>45</td>
<td>0</td>
<td>38.9</td>
</tr>
<tr>
<td></td>
<td>2740</td>
<td>Riboflavin synthase subunit β</td>
<td>28</td>
<td>4&lt;sup&gt;−13&lt;/sup&gt;</td>
<td>31.3</td>
</tr>
<tr>
<td></td>
<td>2968</td>
<td>α-/β-Hydrolase</td>
<td>34</td>
<td>3&lt;sup&gt;−35&lt;/sup&gt;</td>
<td>30.2</td>
</tr>
<tr>
<td></td>
<td>2741</td>
<td>N-Acetyltransferase</td>
<td>31</td>
<td>6&lt;sup&gt;−40&lt;/sup&gt;</td>
<td>34.2</td>
</tr>
<tr>
<td></td>
<td>2742</td>
<td>Spore coat protein GT4</td>
<td>81</td>
<td>0</td>
<td>38.5</td>
</tr>
<tr>
<td><em>N. dentiae</em> (CCUG 53898)</td>
<td>2173</td>
<td>Glycerol-3-phosphate cytidylytransferase GT9*</td>
<td>99</td>
<td>0</td>
<td>43.6</td>
</tr>
<tr>
<td></td>
<td>2941</td>
<td>Phosphotransferase</td>
<td>87</td>
<td>0</td>
<td>43.4</td>
</tr>
<tr>
<td></td>
<td>2942</td>
<td>CDP-glycerol–glycerophosphate glycerophosphotransferase GT4</td>
<td>41</td>
<td>3&lt;sup&gt;−24&lt;/sup&gt;</td>
<td>37.9</td>
</tr>
<tr>
<td></td>
<td>2943</td>
<td>CDP-glycerol–glycerophosphate glycerophosphotransferase GT2</td>
<td>36</td>
<td>2&lt;sup&gt;−36&lt;/sup&gt;</td>
<td>39.4</td>
</tr>
<tr>
<td><em>N. elongata subsp. elongata</em> (CCUG 2043T)</td>
<td>2173</td>
<td>Capsule biosynthesis protein GT4</td>
<td>100</td>
<td>0</td>
<td>31.4</td>
</tr>
<tr>
<td></td>
<td>2941</td>
<td>Glycosyl transferase GT2</td>
<td>100</td>
<td>0</td>
<td>36.1</td>
</tr>
<tr>
<td></td>
<td>2942</td>
<td>Acetyltransferase</td>
<td>100</td>
<td>0</td>
<td>37.9</td>
</tr>
<tr>
<td></td>
<td>2943</td>
<td>α-/β-Hydrolase</td>
<td>38</td>
<td>1&lt;sup&gt;−15&lt;/sup&gt;</td>
<td>34.6</td>
</tr>
<tr>
<td></td>
<td>2974</td>
<td>d-Alanine–d-alanine ligase GT2</td>
<td>100</td>
<td>0</td>
<td>37.7</td>
</tr>
<tr>
<td><em>N. elongata subsp. nitroreducens</em> (CCUG 30802T)</td>
<td>2173</td>
<td>Glycerol-3-phosphate cytidylytransferase GT9*</td>
<td>99</td>
<td>1&lt;sup&gt;−89&lt;/sup&gt;</td>
<td>39.2</td>
</tr>
<tr>
<td></td>
<td>2948</td>
<td>Phosphotransferase</td>
<td>51</td>
<td>1&lt;sup&gt;−116&lt;/sup&gt;</td>
<td>33.7</td>
</tr>
<tr>
<td></td>
<td>2939</td>
<td>CDP-glycerol–glycerophosphate glycerophosphotransferase GT4</td>
<td>31</td>
<td>1&lt;sup&gt;−22&lt;/sup&gt;</td>
<td>36.5</td>
</tr>
<tr>
<td></td>
<td>2940</td>
<td>CDP-glycerol–glycerophosphate glycerophosphotransferase GT2</td>
<td>35</td>
<td>9&lt;sup&gt;−37&lt;/sup&gt;</td>
<td>37.4</td>
</tr>
<tr>
<td><em>N. mucosa</em> (CCUG 431)</td>
<td>2972</td>
<td>Spore coat protein GT2</td>
<td>59</td>
<td>0</td>
<td>29.9</td>
</tr>
<tr>
<td></td>
<td>2973</td>
<td>Acetyltransferase</td>
<td>45</td>
<td>8&lt;sup&gt;−58&lt;/sup&gt;</td>
<td>33.5</td>
</tr>
<tr>
<td><em>N. musculi</em> (AP2031)</td>
<td>2173</td>
<td>UDP-N-acetylglucosamine 2-epimerase GT28*</td>
<td>78</td>
<td>0</td>
<td>50.2</td>
</tr>
<tr>
<td></td>
<td>2737</td>
<td>UDP-N-acetyl-d-mannosamine dehydrogenase GT2*</td>
<td>91</td>
<td>0</td>
<td>49.9</td>
</tr>
<tr>
<td></td>
<td>2738</td>
<td>Glycosyl transferase GT2</td>
<td>59</td>
<td>6&lt;sup&gt;−92&lt;/sup&gt;</td>
<td>35.4</td>
</tr>
<tr>
<td></td>
<td>2739</td>
<td>Glycosyl transferase GT4</td>
<td>48</td>
<td>0</td>
<td>38.5</td>
</tr>
<tr>
<td></td>
<td>2950</td>
<td>Asparagine synthase</td>
<td>22</td>
<td>1&lt;sup&gt;−22&lt;/sup&gt;</td>
<td>31.6</td>
</tr>
<tr>
<td></td>
<td>2741</td>
<td>N-Acetyltransferase</td>
<td>34</td>
<td>3&lt;sup&gt;−37&lt;/sup&gt;</td>
<td>35.1</td>
</tr>
<tr>
<td></td>
<td>2742</td>
<td>Spore coat protein GT4</td>
<td>90</td>
<td>0</td>
<td>46.4</td>
</tr>
<tr>
<td><em>N. oralis</em> (F0314)</td>
<td>2941</td>
<td>Glycosyl transferase GT2</td>
<td>100</td>
<td>0</td>
<td>39.6</td>
</tr>
<tr>
<td></td>
<td>2942</td>
<td>Acetyltransferase</td>
<td>100</td>
<td>0</td>
<td>37.9</td>
</tr>
<tr>
<td></td>
<td>2943</td>
<td>α-/β-Hydrolase (variant in CCUG 804)</td>
<td>36</td>
<td>9&lt;sup&gt;−14&lt;/sup&gt;</td>
<td>36.3</td>
</tr>
<tr>
<td><em>N. subflava</em> (NJ9703)</td>
<td>2173</td>
<td>Capsule biosynthesis protein GT4</td>
<td>100</td>
<td>0</td>
<td>39.7</td>
</tr>
<tr>
<td></td>
<td>2941</td>
<td>Glycosyl transferase GT2 (variant in C102, C6A, CCUG 7826 and CCUG 24918)</td>
<td>100</td>
<td>0</td>
<td>39.7</td>
</tr>
<tr>
<td></td>
<td>2942</td>
<td>Acetyltransferase (missing in C6A, CCUG 7826 and CCUG 24918)</td>
<td>100</td>
<td>0</td>
<td>37.2</td>
</tr>
<tr>
<td></td>
<td>2943</td>
<td>α-/β-Hydrolase</td>
<td>38</td>
<td>6&lt;sup&gt;−15&lt;/sup&gt;</td>
<td>36.1</td>
</tr>
</tbody>
</table>
In *N. bacilliformis*, these four region A candidates were contiguous, but were located on different contigs from regions B and C. Region A candidates were not identified in *N. subflava* isolates that contained region B, but not region C. The G+C content of region A candidates was found to be lower than those typical for *Neisseria* genomes (49–54%, but 60% for *N. bacilliformis*) (Table 1). tBLASTx queries additionally identified homologues of meningococcal serogroup B/C/W/Y sialic acid synthesis genes cssABC in *N. weaveri* isolate CCUG 4007 T. These three genes were in a separate region of the genome of CCUG 4007 T and were distinct to the other candidate region A genes identified flanking regions B and C.

**Arrangement of cps in *N. meningitidis* and NPN species**

None of the NPN *cps* were syntenic with the gene order seen in *N. meningitidis* (Fig. 1): all NPN lacked the duplicated region D, and only contained tex from region E, with the pseudo cytosine methyltransferases not identified during BLAST searches. *N. subflava* was the only NPN species in which the putative regions A, B, C and D were contiguous, although the putative regions A, B and C were contiguous or nearly contiguous in all species, apart from *N. oralis*, *N. animalis*, *N. canis* and *N. bacilliformis*. In the case of *N. canis*, *N. animalis*, *N. elongata* subsp. elongata and *N. bacilliformis*, the different regions were not located on a single contig, but the separation of these regions being an artefact was rejected based on comparisons to closed genome sequences of these species. The gene encoding galE from region D was not present in all NPN found to contain a *cps*, and in *N. animalis*, *N. oralis* and both *N. elongata* isolates, it was found to be near or adjacent to region A, rather than contiguously with the other region D genes as is the case in *N. meningitidis*.

**Homologous region A genes among species**

In some instances, several isolates from each species possessed a *cps*: *N. bacilliformis* (four isolates); *N. subflava* (six isolates); and *N. oralis* (four isolates). Each gene found in *N. bacilliformis* shared >98% nucleotide identity with the corresponding gene in all other isolates, indicating that all four isolates shared a highly similar *cps*. In *N. subflava*, four of the five candidate genes shared >97% nucleotide identity among isolates, although three isolates were missing a predicted acetyltransferase. The identity scores for the other gene, a predicted glycosyltransferase, were either 71–73 or 97–100%, which indicated that there were two versions of this gene (Fig. 2b). *N. oralis* had three region A candidates, two of which shared >98% nucleotide identity among isolates. The third only had identity scores of >98% among three of the four isolates, with the version in CCUG 804 only sharing 81% identity with the others (Fig. 2b).

Based on similarities in *cps* organization, and the results of BLAST searches during region A annotation, it also became clear that there was shared region A homology between NPN species. Homologous genes were consistently grouped, such that if a group of species shared one gene, they were likely to share another gene, with three homology groups identified in total (Fig. 2). Pairwise comparisons of amino acid sequences generated by Clustal were used to analyse these groups further.

Group 1 contained putative region A from *N. animalis*, *N. elongata* subsp. *nitroreducens* and *N. bacilliformis*, which shared four homologous genes with 78–83, 42–47, 56–58 and 61–64% aa identity between species (Fig. 2a). The first two genes also shared 73–81 and 42–47% aa identity with *cssA/*cshA and *cssB/*cshB, respectively, which are found at the beginning of the region A of serogroups H and Z.

Group 2 contained putative region A from *N. elongata* subsp. *elongata*, *N. subflava* and *N. oralis*, which shared up to five homologous genes with 58, 60–95, 97, 86–95 and 75–99% aa identity between species, although *N. oralis* was missing the first and third genes, and three isolates of *N. subflava* lacked the third gene (Fig. 2b). The first two genes also shared 61–72 and 32–33% aa identity with *cslA* and *cslB*, respectively, which are found at the beginning of the region A of serogroup I.

### Table 1. cont.

<table>
<thead>
<tr>
<th>Species</th>
<th>NEIS locus</th>
<th>Product of highest non-hypothetical pBLAST hit/CAZy family</th>
<th>Identity (%)</th>
<th>E value</th>
<th>G+C (mol%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>N. weaveri</em> (CCUG 4007 T)</td>
<td>2974</td>
<td>d-Alanine–d-alanine ligase GT2</td>
<td>100</td>
<td>0</td>
<td>40.55</td>
</tr>
<tr>
<td></td>
<td>2944</td>
<td>Glycosyl transferase GT4</td>
<td>100</td>
<td>0</td>
<td>30.2</td>
</tr>
<tr>
<td></td>
<td>2945</td>
<td>Capsule biosynthesis protein</td>
<td>99</td>
<td>0</td>
<td>27.5</td>
</tr>
<tr>
<td></td>
<td>2946</td>
<td>Glycosyl transferase GT61*</td>
<td>29</td>
<td>2–17</td>
<td>28.6</td>
</tr>
<tr>
<td><em>N. zoodegmatis</em> (NCTC 12230 T)</td>
<td>2736</td>
<td>UDP-N-acetylglucosamine 2-epimerase GT28*</td>
<td>77</td>
<td>0</td>
<td>40.7</td>
</tr>
<tr>
<td></td>
<td>2737</td>
<td>UDP-N-acetyl-d-mannosamine dehydrogenase G2*</td>
<td>85</td>
<td>0</td>
<td>42.3</td>
</tr>
<tr>
<td></td>
<td>2949</td>
<td>Glycosyl transferase GT4</td>
<td>34</td>
<td>3–126</td>
<td>37.6</td>
</tr>
<tr>
<td></td>
<td>2971</td>
<td>Hypothetical</td>
<td>48</td>
<td>0</td>
<td>34.7</td>
</tr>
<tr>
<td></td>
<td>2742</td>
<td>Spore coat protein GT4</td>
<td>81</td>
<td>0</td>
<td>37.8</td>
</tr>
</tbody>
</table>

*Indicates CAZy family predictions determined using Pfam rule-based annotations rather than sequence similarity.
Fig. 2. Pairwise tBLASTx comparisons of candidate region A genes between representative *Neisseria* species proposed to contain region A genes. Isolates within each group, 1 (a), 2 (b) or 3 (c) share homologous region A genes. Red and blue indicate shared amino acid identity between isolates, with higher intensity indicating higher sequence similarity. Blue indicates an inversion. Between species variation in nucleotide sequence identity was >97%; hence, one representative is provided for each species, with three exceptions: two *N. elongata* isolates belong to totally different groups 1 (a)/ 3 (b); one of the four *N. oralis* isolates had a variant of NEIS2943 that shared only 81% nucleotide identity with other *N. oralis* isolates, and two of the six *N. subflava* isolates had a variant of NEIS2941 that shared only 71–73% nucleotide identity with other *N. subflava* isolates (b).
gene with 59% aa identity with csiB, differing in length by only 3 bp, as well as a homologue with 56% aa identity to csiC.

Group 3 contained putative region A as from N. dentiae, N. musculi, N. canis, N. zoodegmatis and N. animaloris, although N. animaloris was not annotated further due to its incomplete assembly. These isolates shared up to seven homologous genes with 56–96%, 70–82%, 73–85%, 65%, 68–83% and 79–90% aa identity between species (Fig. 2c). N. zoodegmatis only contained the first two and last one of these genes, whilst N. dentiae lacked the fifth gene. The first two genes and last gene were also 71–74%, 83–89% and 76–86% homologous to csiA, csiB and csiE from serogroup I N. meningitidis, respectively, with the exception of the first gene in N. canis, which had only 56% aa identity with csiE and 68% aa identity with csaA from serogroup A N. meningitidis.

**Distribution of cps homologues among Neisseria species**

Mapping the presence of cps onto the phylogeny of Neisseria species reconstructed from rMLST [24] sequences indicated that cps genes were common and widely distributed among Neisseria (Fig. 3). Species sharing homologous region A genes did not necessarily belong to a monophyletic group. N. cinerea, N. lactamica, N. polysacchara and N. gonorhoeae all belonged to a monophyletic group with N. meningitidis, and, with the exception of N. meningitidis, no isolates from these species were found to possess a

---

**Fig. 3.** An ML phylogeny generated from an alignment of concatenated rMLST nucleotide sequences from Neisseria, with Moraxella catarrhalis as an outgroup. Branch supports are based on 100 bootstrap replicates. Corrected for recombination in ClonalFrameML. Coloured circles indicate the presence of capsule genes. Isolates sharing homology in region A, as shown in Fig. 2, are filled with the same colour. Other isolates are black. Grey circles indicate partial region B genes identified only.
complete \textit{cps} locus. The closest encapsulated relative of \textit{N. meningitidis} was \textit{N. subflava}. Absences of \textit{cps} genes in other species were more sporadic, with 2 of 4 \textit{N. elongata} subsp. isolates, \textit{Neisseria shayegani} (1 isolate), 9 of 10 \textit{N. mucosa} and 13 of 19 \textit{N. subflava} lacking \textit{cps}, although 11 of the \textit{N. subflava} isolates contained remnants of region \textit{B} genes.

NCBI pBLAST searches indicated that \textit{Kingella kingae}, \textit{B. trehalosi}, \textit{Actinobacillus sp.}, \textit{Mannheimia haemolytica} and \textit{H. influenzae} possessed homologues of \textit{N. meningitidis} region \textit{B} and \textit{C} genes. An unrooted ML phylogeny generated from aligned amino acid sequences of region \textit{B} and \textit{C} homologues in these species, \textit{P. multocida} and \textit{Neisseria} indicated that genes in \textit{N. meningitidis} were more closely related to those from \textit{N. subflava} and most other \textit{Neisseria} than any other genera (Fig. 4).

DISCUSSION

Among the \textit{Neisseria}, the polysaccharide capsule has been considered to be a virulence factor unique to \textit{N. meningitidis}. Although region \textit{B} and \textit{C} genes had been identified previously in an isolate of \textit{N. subflava}, in the absence of further evidence at the time, this was attributed to an isolated HGT event facilitated by a DNA uptake sequence in \textit{ctrA} [21]. In this study, homologues of all the conserved region \textit{B} and \textit{C} genes in multiple NPN species from across the genus have been identified, with accompanying putative capsule synthesis loci. On the balance of evidence from genomic data, including the comparable synteny between most NPN species and the meningococcal \textit{cps} (Fig. 1), and the high homology of several putative capsule synthesis genes to those of \textit{N. meningitidis} and other species (Fig. 2, Table 1), the candidate region \textit{A} genes identified were most likely to function in capsule synthesis. The discovery of capsule genes in non-pathogenic bacteria is not unprecedented, with a similar finding in the \textit{H. influenzae}, \textit{S. mitis} group streptococci overturning the assumption that capsule production was unique to the pathogenic \textit{S. pneumoniae} [41]. In common with many virulence factors [42], including the type IV pilus [20, 21], the capsule might be better described as a ‘host adaptation factor’ [21], with effects on pathogenic potential being incidental.

Region A annotations were consistent with the potential for more than one capsule group within \textit{N. elongata}, and possibly \textit{N. subflava} and \textit{N. oralis} (Fig. 2). Differences observed among \textit{N. subflava} isolates were comparable to the divergence between the polysialyltransferase-encoding \textit{csb} and \textit{csc}, which give rise to the structural differences between meningococcal serogroups \textit{B} and \textit{C}, respectively [15, 43]. The presence of multiple groups or types is commonplace among Gram-negative bacteria, including \textit{E. coli}, \textit{H. influenzae} and \textit{Mannheimia haemolytica} [4–6]. In \textit{E. coli}, over 80 structurally different capsules exist, some of which are associated with specific pathologies, or are only expressed at certain temperatures [4]. The range of niches exploited by \textit{E. coli}, including different hosts and tissues, as well as free-living environments, may be responsible for this diversity [44]. \textit{Neisseria} do not demonstrate such a wide exploitation of niches, but it has been demonstrated that different species have tropisms for specific nasopharyngeal sites [45]. In \textit{N. meningitidis}, isolates expressing a capsule from serogroups \textit{A}, \textit{B}, \textit{C}, \textit{W}, \textit{X} or \textit{Y} are associated with IMD [2], leading to an interest in the evolutionary history of the capsule.

Models presented previously, hypothesizing that \textit{N. meningitidis} must have acquired a capsule by HGT [3, 22], can be re-examined in light of the data presented here. The identification of capsules in 13 NPN species does not preclude the acquisition of capsule genes in \textit{N. meningitidis} by HGT. Notably, capsule genes have still not been identified in any isolates belonging to the monophyletic group that contains \textit{N. cinerea}, \textit{N. lactamica}, ‘\textit{N. bergeri}’, \textit{N. polysaccharaea}, \textit{N. gonorrhoeae} and \textit{N. meningitidis}, with the exception of \textit{N. meningitidis} itself (Fig. 3). This uneven distribution of capsule genes must be explained by acquisition and/or loss [46, 47]. Given the predicted common ancestor of these species, for the capsule to be present only in certain \textit{N. meningitidis} isolates, the capsule genes must either have been lost independently as many as six times, or lost once in a common ancestor of the monophyle and re-acquired in \textit{N. meningitidis}. The latter is the most evolutionarily parsimonious explanation, although a scenario between these two extremes is also possible. Further support for an HGT event comes from the duplication of region \textit{D} genes within the \textit{N. meningitidis} \textit{cps}, which is attributed to illegitimate recombination in the \textit{galE} gene in a model proposed by Bartley et al. [22]. A satisfactory alternative explanation for the organization of the meningococcal \textit{cps} has not been proposed to date.

It has been posited that the donor of capsule genes to \textit{N. meningitidis} may have been a member of the \textit{Pasteurellaceae}, based on \textit{cps} organization and sequence similarity between regions \textit{B} and \textit{C} of \textit{N. meningitidis} and equivalent genes in \textit{P. multocida} [3]. This is not consistent with the phylogenetic data presented here, which show that regions \textit{B} and \textit{C} of the \textit{N. meningitidis} \textit{cps} more closely resembled homologues of \textit{N. subflava} than any other genus (Fig. 4). Therefore, if the \textit{N. meningitidis} \textit{cps} was acquired by HGT, regions \textit{B} and \textit{C} at least were more likely to have been acquired from another \textit{Neisseria} species. Recombination between closely related species is more probable, due to higher similarity of flanking sequences and the increased potential for a compatible DNA uptake sequence. HGT between \textit{Neisseria} species has been described previously, including cross-species exchanges of \textit{pilE}, another gene with links to virulence [21, 48]. Acquisition of capsule genes in \textit{H. influenzae} has also been proposed to be a result of HGT from a commensal species of the same genus [49]. Interestingly, the suggested organization of the donor ancestral \textit{cps} island described in the model by Bartley et al. [22] matches the organization of the \textit{N. subflava} \textit{cps} described in the present study; however, a potential donor could alternatively be
a close relative that has either not been previously isolated or since become extinct.

The origin of region A genes in *N. meningitidis*, responsible for the differences in capsule serogroups, is less clear, since none of the isolates annotated here possessed a full complement of region A loci with a close resemblance to *N. meningitidis* serogroups. Based on sequence similarity among capsule synthesis genes of *Haemophilus*, *Actinobacillus*, *Mannheimia* and *Neisseria*, a case for horizontal exchange of capsule synthesis genes across genera and the formation of mosaic complements of genes has been made [50, 51]. The lower G+C content of region A (as low as 25–45 %) compared to the rest of the genome (~50 %), a phenomenon also seen in some *E. coli* capsular types, has also been cited as evidence for cross-genus horizontal acquisition of capsule synthesis genes [15, 52]. The exact nature of this potentially complex evolutionary history, and the degree of exchange in recent evolutionary time, remain unclear.

The discovery of capsule genes in NPN highlights the polysaccharide capsule’s role in asymptomatic colonization and transmission, an important stage in meningococcal epidemiology; however, the acquisition of capsule by some genotypes of *N. meningitidis* has had an important impact on their behaviour, increasing their propensity to cause disease. Sequence similarity between NPN capsule genes and *N. meningitidis* sheds some light on the complicated evolutionary processes in these highly transformable organisms. Further sequencing of NPN, as well as other oral and nasopharyngeal commensals, may provide additional insights into the emergence of pathogenic serogroups in this important pathogen.

**Funding information**
The work was funded by the Wellcome Trust (109025/Z/15/A to M. C. J. M.; 104992/Z/14/Z to M. E. A. C.; O. B. H. was supported by the Wellcome Trust Institutional Strategic Support Fund).

**Acknowledgements**
This publication made use of the *Neisseria* Multi Locus Sequence Typing website (https://pubmlst.org/neisseria/), developed by Keith Jolley and Martin Maiden [23] and sited at the University of Oxford, UK. The development of this site has been funded by the Wellcome Trust and the European Union. The authors are grateful to James Bray for assembly of genomes and implementation of rMLST-based speciation classification within pubMLST.

**Conflicts of interest**
The authors declare that there are no conflicts of interest.

**Data bibliography**

---

*[Fig. 4. An unrooted ML phylogeny with 100 bootstraps, generated from an alignment of concatenated amino acid sequences of \( ctrABCDEF \) (or their equivalent homologues) from *Neisseria* species, and other members of the *Neisseriaceae* and *Pasteurellaceae*.]*
15. Irish Meningococcus Genome and. pubMLST id 26870 (2013).

References
12. Weber MV, Claus H, Maiden MC, Frosch M, Vogel U. Genetic mechanisms for loss of encapsulation in polysialyltransferase-


20. 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 2006;7:128.


49. Nielsen SM, de Gier C, Dimopoulou C, Gupta V, Hansen LH et al. The capsule biosynthesis locus of Haemophilus influenzae Rd shows conspicuous similarity to the corresponding locus in Haemophilus sputorum and may have been recruited from this species by horizontal gene transfer. Microbiology 2015;161:1182–1188.

