FEMS Microbiology Letters, 369, 2022, fnac064

https://doi.org/10.1093/femsle/fnac064
Advance Access Publication Date: 22 July 2022
Research Letter

RESEARCH LETTER – Pathogens & Pathogenicity

Examination of phase-variable haemoglobin–haptoglobin binding proteins in non-typeable *Haemophilus influenzae* reveals a diverse distribution of multiple variants

Zachary N. Phillips[1], Amy V. Jennison[2], Paul W. Whitby[3], Terrence L. Stull[3], Megan Staples[2] and John M. Atack[1,4,*]

[1]Institute for Glycomics, Griffith University, Gold Coast, Queensland, Australia, [2]Queensland Department of Health, Public Health Microbiology, Forensic and Scientific Services, Brisbane, Queensland, Australia, [3]BacVax, Inc., Phoenix, AZ, United States and [4]School of Environment and Science, Griffith University, Gold Coast, Queensland, Australia

*Corresponding author. Institute for Glycomics, Griffith University, Gold Coast, QLD 4215, Australia, Tel: +61 (7) 555 80580; E-mail: j.atack@griffith.edu.au
John M. Atack, http://orcid.org/http://orcid.org/0000-0002-7994-6995

One sentence summary: Iron binding proteins (Hgps) in NTHi are highly diverse, but contain conserved regions, with certain proteins present in all strains, and expressed during invasive disease.

Editor: Klaus Hantke.

ABSTRACT

Non-typeable *Haemophilus influenzae* (NTHi) is a major human pathogen for which there is no globally licensed vaccine. NTHi has a strict growth requirement for iron and encodes several systems to scavenge elemental iron and heme from the host. An effective NTHi vaccine would target conserved, essential surface factors, such as those involved in iron acquisition. Haemoglobin–haptoglobin binding proteins (Hgps) are iron-uptake proteins localized on the outer-membrane of NTHi. If the Hgps are to be included as components of a rationally designed subunit vaccine against NTHi, it is important to understand their prevalence and diversity. Following analysis of all available Hgp sequences, we propose a standardized grouping method for Hgps, and demonstrate increased diversity of these proteins than previously determined. This analysis demonstrated that genes encoding variants HgpB and HgpC are present in all strains examined, and almost 40% of strains had a duplicate, nonidentical *hgpB* gene. Hgps are also phase-variably expressed; the encoding genes contain a $CCAA_{(n)}$ simple DNA sequence repeat tract, resulting in biphasic ON–OFF switching of expression. Examination of the ON–OFF state of *hgpB and hgpC* genes in a collection of invasive NTHi isolates demonstrated that 58% of isolates had at least one of *hgpB* or *hgpC* expressed (ON). Varying expression of a diverse repertoire of *hgp* genes would provide strains a method of evading an immune response while maintaining the ability to acquire iron via heme. Structural analysis of Hgps also revealed high sequence variability at the sites predicted to be surface exposed, demonstrating a further mechanism to

Received: 14 March 2022; **Revised:** 14 June 2022; **Accepted:** 20 July 2022

evade the immune system—through varying the surface, immune-exposed regions of the membrane anchored protein. This information will direct and inform the choice of candidates to include in a vaccine against NTHi.

Keywords: NTHi; invasive disease; iron acquisition; phase variation; Hgp

Introduction

Non-typeable *Haemophilus influenzae* (NTHi) causes significant global morbidity and mortality. NTHi is a human-specific opportunistic pathogen and can colonize the nasopharynx of the human host asymptomatically. Migration of bacteria from this site to other niches within the respiratory tract results in acute and chronic infections, such as middle ear disease (otitis media; OM), exacerbations in chronic obstructive pulmonary disease (COPD), pneumonia, and sinusitis (Van Eldere et al. 2014). NTHi is also a major cause of invasive bacterial infections such as meningitis and septicaemia. The proportion of invasive NTHi disease has been increasing since the introduction of a vaccine against *H. influenzae* serogroup b (Hib) in the mid-1980s (Whitby et al. 2009), with NTHi now the major cause of invasive infections caused by *Haemophilus* species. NTHi invasive infections are a particular problem for children with significant comorbidities. Even without complicating factors, invasive NTHi infections are fatal in up to 17% of children under 1, and in ∼10% of children aged 2–4 years old (Ladhani et al. 2010). There is currently no globally licensed vaccine to prevent NTHi mediated disease despite > 20 years of research. This is due to a high level of inter- and intrastrain diversity of homologous proteins in NTHi isolates. As such, understanding diversity and regulation of conserved and/or essential proteins will provide key information towards development of a vaccine that can target all NTHi strains.

Ideal vaccines and therapies target conserved features to ensure broad effectiveness. NTHi has an absolute growth requirement for iron, making surface factors involved in iron uptake logical vaccine targets. Progress towards targeting NTHi iron-uptake systems has been hindered by factors including significant intra- and interstrain diversity of the encoding genes, and the functional redundancy of many of the proteins. Several core NTHi genes produce iron acquisition proteins that are surface located, such as *hup, hemR, hxuC, hxuB, tbp1, tbp2, hgpB*, and *hgpC* (Whitby et al. 2015). Vaccines require stable expression of antigens to be effective, so candidates that vary expression are not ideal. For example, *hxuC* undergoes repression/derepression (Whitby et al. 2009), and, in addition, *hgpB* and *hgpC* are phase-variable; they undergo rapid and reversible ON–OFF switching of expression (Ren et al. 1999). Even if a vaccine were to target these core iron-uptake genes, accessory genes involved in iron uptake are abundant in NTHi, providing alternate means of iron homeostasis if the core genes were targeted. These accessory genes are also frequently exchanged between strains; e.g. the *speF–potE* operon has been observed to swap with *hgpA* (Whitby et al. 2013). As targeting only core iron-uptake genes is unlikely to be an effective vaccine strategy, and accessory genes are abundant and transient, an approach to generate a rationally designed vaccine containing multiple key antigens, including core iron acquisition factors, may prove to be the key in successfully formulating a vaccine targeting all NTHi strains.

NTHi haemoglobin–haptoglobin binding proteins (Hgps) sequester iron from haemoproteins such as haemoglobin, haemoglobin–haptoglobin and myoglobin–haptoglobin, all primary iron sources in the human host (Morton et al. 2006, Choby and Skaar 2016). The Hgps have been demonstrated to be present in all NTHi strains, and have been identified as virulence determinants (Morton et al. 2004, Xie et al. 2006, Poole et al. 2013, Whitby et al. 2015). Previously, four individual Hgps have been described based on sequence homology—HgpA (Jin et al. 1999), HgpB (Ren et al. 1998), HgpC (Morton et al. 1999), and HgpD (Morton and Stull 1999, Harrison et al. 2005, Whitby et al. 2013). However, these studies did not compare genes or detect duplications. More recent studies found *hgpB* is more prevalent in OM isolates vs throat isolates (Xie et al. 2006), and mutants lacking functional Hgps are less virulent in a rat model (Seale et al. 2006). Genes encoding Hgps are phase-variable; they contain a $CCAA_{(n)}$ simple DNA sequence repeat (SSR) tract in the 5` region of the open-reading frame. Gain or loss of repeats at this SSR tract due to slipped strand mispairing causes the gene to reversibly switch ON (in-frame; expressed) or OFF (out of frame; not expressed). This random expression of Hgps generates population diversity. Phase-variable expression of surface features provides a contingency strategy to allow bacterial pathogens to respond to environmental changes, such as immune pressure. However, as the switching OFF of expression of a vaccine target could lead to vaccine failure, phase-variable candidates are typically not investigated for inclusion in subunit vaccines. Counterintuitively, phase-variable proteins can form part of a rationally designed vaccine if they are required for key stages of disease, or they are highly expressed in particular host niches. This is the case for the current vaccine against serotype B *Neisseria meningitidis*, BexSero, that contains the phase-variable protein NadA, i.e. required for invasive meningococcal infections (Green et al. 2018).

In addition to being phase-variably expressed, *hgpB* and *hgpC* genes have been observed to rapidly acquire point mutations, which are selected for in persistent infections (Garmendia et al. 2014). Phase-variable expression, and a tendency to accumulate amino acid changes in surface exposed regions, suggests that Hgps are under immune pressure. However, their ubiquitous presence, and the essentiality of iron in NTHi growth, means that Hgps could be used as vaccine candidates, though a fundamental knowledge of their diversity is lacking. A rationally designed vaccine containing Hgps would provide targets of essential proteins to protect against all strains. To validate the inclusion of these proteins in a potential NTHi vaccine, we carried out a thorough analysis of both the prevalence and diversity of Hgps, and if selection for particular phase-variants of HgpB and HgpC was occurring in an extensive collection of NTHi isolated from patients presenting with an invasive infection.

Methods

Bacterial isolate collections

Invasive NTHi isolates used for this study were minimally passaged and isolated from patients presenting with *H. influenzae* infections in South East Queensland over a 15-year period (2001–2015; Staples et al. 2017). Information on age, sample site, and geographical location were collected, but information on any comorbidity was not (Staples et al. 2017). The 74 isolates in this

study were selected to represent a random sample of the NTHi strains present in this collection.

DNA preparation and analysis

Bacterial genomic DNA from invasive isolates were prepared as described previously (Phillips et al. 2019). PCRs and analysis were also carried out as previously described (Phillips et al. 2019). *hgp* ON/OFF status was determined from the number of $CCAA_{(n)}$ repeats in the SSR tract present in each gene (based on amplicon peak size) by sizing and quantifying using the GeneScan system (Applied Biosystems International) at the Australian Genome Research Facility (AGRF; Brisbane, Australia), and traces analyzed using PeakScanner software 2.0 (Applied Biosystems International). Primers used within this study are listed in Table S3 (Supporting Information). The results shown in Table 2 indicate whether *hgp*s were ON (> 70% ON; green), OFF (> 70% OFF; red), or mixed ON and OFF (< 70% ON or OFF; orange). The relationship between gene % ON/OFF and expression has been established and used previously (Fox et al. 2014, Atack et al. 2015, Phillips et al. 2019). Ion Torrent PGM genome sequence data for the invasive isolates, which have previously been deposited on the NCBI Sequence Read Archive (PRJEB18702) were searched for sequences that matched individual Hgp groups/branches to ensure the PCR results correlated with actual gene presence. These Ion Torrent genome sequences were also used to determine gene presence of Hgps that were not amplified by PCR/fragment analysis from genomic DNA preps to generate data included in Table 1 and Table S1 (Supporting Information). Data from all strains from NCBI Genbank is presented in Table S2 (Supporting Information).

Bioinformatic and structural analysis

Structural modelling was performed by Phyre v2.0 (Kelley et al. 2015), I-TASSER v5.1 (Zheng et al. 2021), Raptor-X web server (Wang et al. 2017), and AlphaFold v2.1.2 (Jumper et al. 2021) in order to generate preliminary 3D structures of Hgps. Models generated by these programmes were considered homologous, with root-mean-square deviations (RMSDs) of < 2 angstroms (Å) typically observed when comparing models from different platforms. Consequently, the model generated by AlphaFold (using default settings) was used for analysis of all surface regions and heme-binding core as it produced a predicted structure that was not influenced by orthologues in other organisms, and would be less likely to provide misleading data when evaluating ligand binding sites. Ligand binding sites were predicted by 3D Ligand Site online service using default parameters (Wass et al. 2010). A total of 75 fully annotated *H. influenzae* genomes from the NCBI GenBank were used for analysis. Gene and protein translation sequences can be found in Data S1 (Supporting Information). Hgp sequences were aligned using CLUSTAL OMEGA v1.2.4 (Madeira et al. 2019) and visualized in default JalView v2.1.1.7 (Waterhouse et al. 2009) using the Overview Window function, visually representing % identity (% ID) between sequences. For Fig. 2(D), the variable domains (VDs) were aligned separately, independently of other sequences (i.e. by examining them without extra 5` or 3`sequences), to detect conserved sequences within these regions. All heterogeneous regions in the alignment were examined for VDs. These VDs were present in the regions predicted to be the functional, surface exposed (and immune accessible) heme-binding domain. Small (≥10 amino acids) heterogeneous regions within the β-barrel were also found, but not examined further as they do not likely interact with the host when the Hgp proteins are in their native state. Conserved sequences were aligned in CLUSTAL OMEGA and then viewed using default view in the Jalview overview feature. This allowed visual representation of % identity (% ID) between sequences and the % ID within that conserved sequence, rather than the % ID across this whole region without grouping into conserved sequences. For example, in Variable Domain 5 (VD5) of Fig 2(D), which had just two conserved sequences, the first conserved sequence had > 80% ID, the second also had > 80% ID, however, there was only ~30% ID between these two conserved sequences.

Results

A diverse range of Hgps are found in *H. influenzae*

Hgps have previously been classified into four groups with approximately 50% sequence identity between them—HgpA, HgpB, HgpC, and HgpD—but with no defined 'cut-offs' for the % identity (% ID) needed to classify individual proteins. Proteins with high identity to each of these sequences have also been named Hgb (Cope et al. 2000), Hhu (Maciver et al. 1996), or not identified further than 'Hgp' (Dixon et al. 2007), confounding study and analysis. We, therefore, sought to rationalize the naming system, and thoroughly characterize the diversity of these proteins present in *H. influenzae* using publicly available genome sequences from the NCBI GenBank. The majority of these strains were classified as NTHi. Investigation showed that the region of DNA containing the *hgp* gene family is often immediately downstream of the *fucI* gene (Fig. 1A). At least two individual *hgp* genes were contained in this region, separated by 30–80 kilobases (kb), but showed considerable variation between strains, and with little homology in terms of either the order of the individual genes present, the sequences encoded between individual *hgp* genes, or the orientation of the *hgp* genes present. Since the number of *hgp* genes in strains vary, and the distance between *hgp* genes also varies, describing the specific synteny of genes is difficult. In a small number of strains an additional genomic region between the *bioA* gene and a *pyk* gene contained an additional, single, *hgp* gene (Fig. 1A).

Through detailed sequence analysis of 75 *H. influenzae* NCBI genomes, we propose a universal, consensus naming scheme, with all examples classified as haemoglobin/haptoglobin binding proteins (Hgp). We used a cut-off value of > 70% identity to group Hgps as individual proteins via whole protein sequence alignment (Fig. 1B). This resulted in groups HgpA–G. Where appropriate, we then further delineated these groups into allelic variants using an > 80% identity cut-off within each group. All Hgp groups are ≥ 70% identical (i.e. all proteins within the HgpA group are ≥ 70% identical to each other). Subgroups (A1 vs. A2, B1 vs. B2, and so on) were branched because they had ≥ 70% but less than ≤ 80% identity to each other within each individual group (i.e. HgpA1 vs. HgpA2 is ≥ 70% identical—they are all HgpA proteins—but are different allelic variants as they are ≤ 80% identical to each other; all HgpA1 proteins are ≥ 80% identical to each other, and so on). Hgps demonstrated high sequence diversity within regions predicted to be surface located (white regions in Fig. 1B), but overall high identity in the backbone regions (blue in Fig. 1B). The HgpA family, can be split into two allelic variants, which we have named HgpA1 and HgpA2. HgpB was the most common Hgp present, with at least one *hgpB* gene found in every genome analyzed. We did not differentiate HgpB

Table 1. (a) The number of *hgp* genes was surveyed in 75 fully annotated genomes from NCBI with total number (No.) and % of the amount screened (%) shown. (b) A collection of invasive NTHi isolates was also surveyed for *hgp* genes. Total number of *hgp* genes detected and their grouping included. Further information of each gene with subgrouping of alleles (e.g. *hgpA1 and hgpA2*) can be found in Table S2 (Supporting Information).

(a)		***hgp* genes in fully annotated NCBI genomes**						
	Genomes	*A*	*B*	*C*	*D*	*E*	*F*	*G*
No.	75	21	107	84	4	6	8	7
%	100	28.0	142.7	112.0	5.3	8.0	10.7	9.3
(b)		***hgp* genes in invasive NTHi isolates**						
	Isolates	***A***	***B***	***C***	***D***	***E***	***F***	***G***
No.	74	20	89	81	15	8	6	2
%	100	27.0	123.6	109.5	20.3	10.8	8.1	2.7

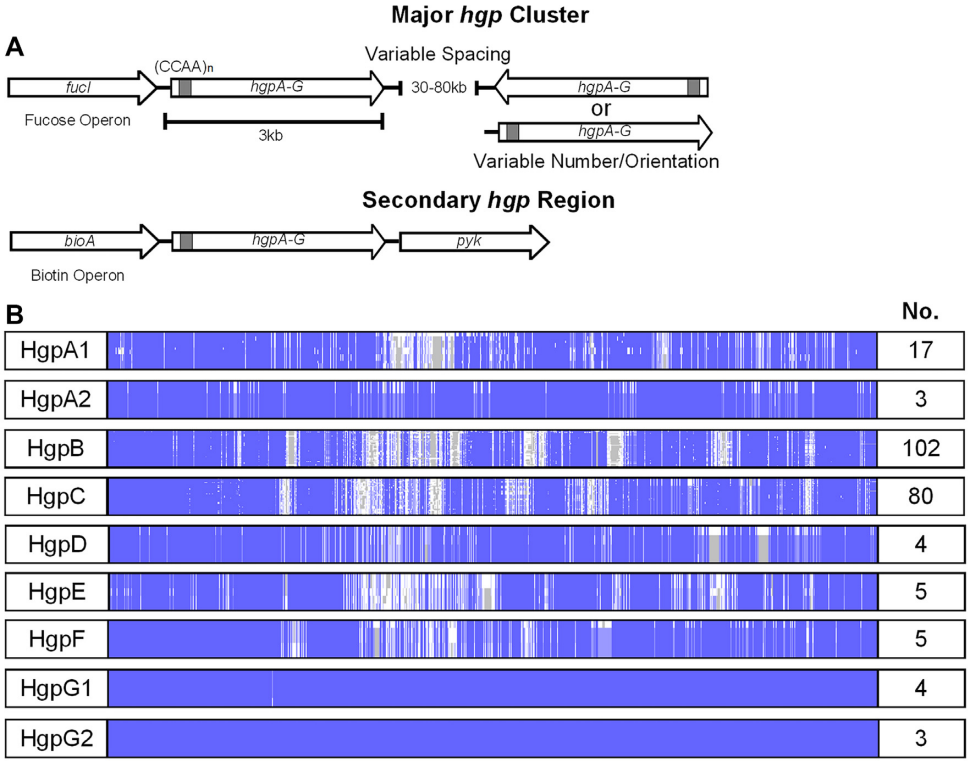

Figure 1. (A) The primary NTHi *hgp* gene cluster is located immediately downstream of the *fucI* gene (encoding a fucose isomerase), with variable distance (30–80 kb) between multiple *hgp* genes located in this region. Our analysis demonstrated that there were at least two *hgp* genes within this primary cluster, but the number of *hgps* varies in number and orientation in individual strains. Additionally, a secondary *hgp* gene can be located between the *bioA* (encoding adenosylmethionine-8-amino-7-oxononanoate aminotransferase) and the a *pyk* gene (encoding pyruvate kinase). This secondary site contains only a single *hgp* gene, and is not present in all strains. (B) Alignment of Hgp amino acid sequences in *H. influenzae* NCBI fully annotated genomes. Protein sequences were aligned by CLUSTAL OMEGA (v1.2.4) and viewed using default JalView (v2.1.1.7) settings, visually representing % identity (% ID) between sequences. The number of sequences aligned is under the 'No.' column. Amino acids are coloured according to the percentage in each column that agree with the consensus sequence, with % identity shown as blue, ranging from > 80% to > 40% identity. Grey areas represent gaps, and white areas indicate < 40% identity with the consensus sequence. We have categorized the previously broad Hgp groups (HgpA–D) using a > 70% identity cut-off to separate Hgps into groups (HgpA–G) and 80% identity to separate alleles (e.g HgpA1 vs. A2).

into allelic groups due to extremely high conservation of the β-barrel backbone in sequences (> 80% identity). HgpC, the second most abundant Hgp, had a similarly high conservation of the β-barrel backbone and was also not divided into alleles (> 80% identity). HgpD was found in just 5.3% of publicly available annotated genomes (4/75), with all examples having > 80% identity and considered a single gene. Our analysis also demonstrated new, previously undescribed Hgps in several *H. influenzae* genomes and in invasive NTHi isolates, which we propose to name HgpE, HgpF, and HgpG. HgpE/F have low identity (~50%) to other Hgps (Fig. 1A; Figure S1a, Supporting Information). HgpG was divided into two allelic variants, which we have named HgpG1 and HgpG2. HgpG had highest identity to HgpC (~60%), but did not meet the threshold of > 70% identity to classified as part of the HgpC group. HgpE and HgpG appear exclusive to *H. influenzae* as no orthologue could be found in another organism (via BLAST analysis). HgpF may have been acquired via horizontal gene transfer from *Pasteurella multocida*, as orthologues

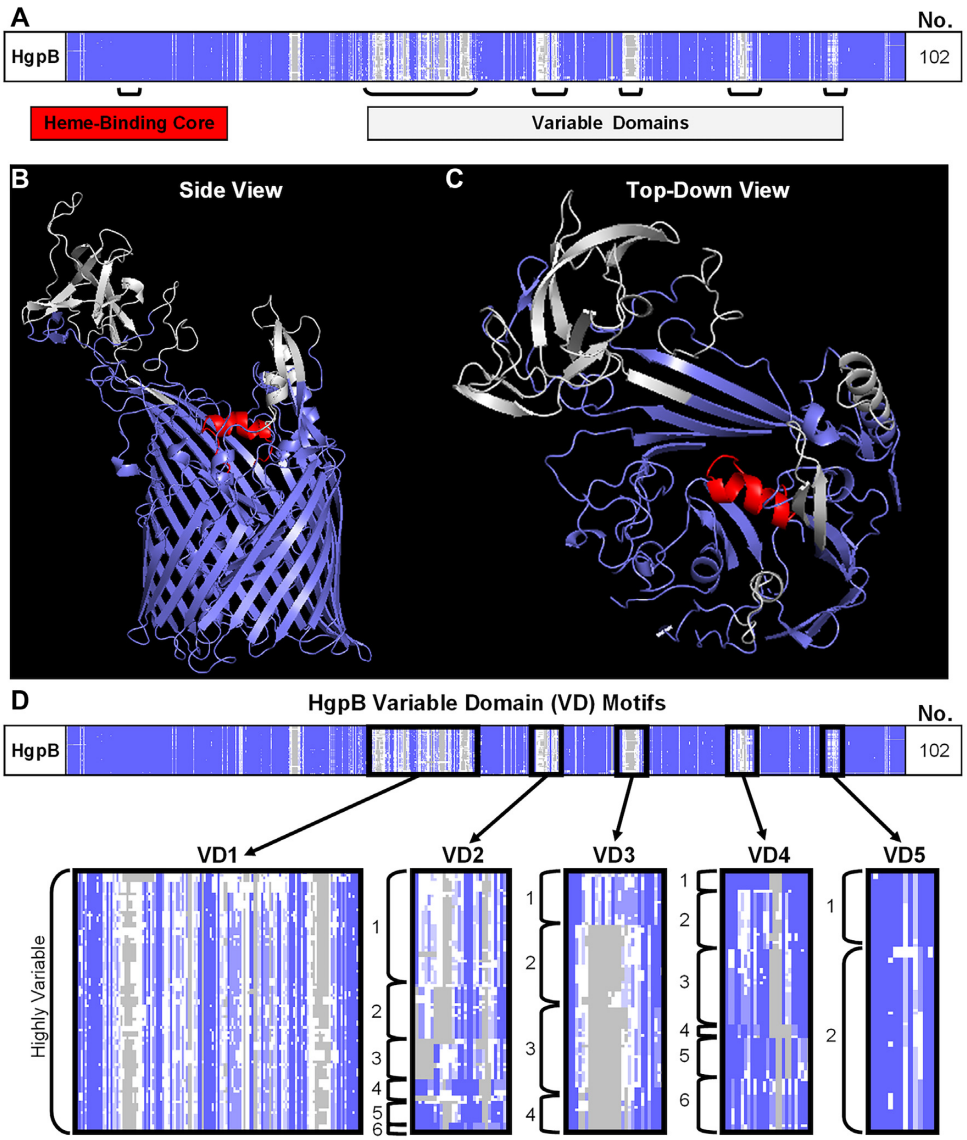

Figure 2. (A) The location of the surface domains and heme-binding core within aligned HgpB protein sequences. The structure of HgpB (from strain NCTC13377) was predicted using AlphaFold (v2.1.2), with (B) side and (C) top-down view provided. The VDs of the Hgps are located in surface-exposed areas (white). The β-barrel structure was highly conserved within Hgp groups (blue). The heme-binding core was surface accessible and highly conserved between Hgp groups (red). (D) Variable (surface) Domains (VD1–5) of HgpB contain highly variable sequences. Individual sequences were identified by aligning all the sequences present from each of the VDs (separate from the whole sequence) in CLUSTAL OMEGA (v1.2.4) and viewed using default settings in JalView overview (v2.1.1.7). A total of 102 HgpB protein sequences were included in the alignments. Amino acids are coloured according to the percentage in each column that agree with the consensus sequence, with % identity shown as blue, ranging from > 80% to > 40% identity. Grey areas represent gaps, and white areas indicate < 40% identity with the consensus sequence. VD1–the largest surface domain—had the highest sequence variability, and was not separated into individual conserved sequences. VD2–VD5 had a lower amount of diversity than VD1, and as such we have been able to individually identify the number of variants within each of these VDs (numbered 1–6) indicated on the left-hand side of each individual VD alignment.

were abundant within this organism but found infrequently in *H. influenzae* (Figure S2, Supporting Information). HgpE/F were both found in ~9% of genomes, while HgpG was only present in ~5% of genomes.

Hgp amino acid sequences vary at surface exposed regions

Surface exposed regions of proteins are typically highly immunogenic, and as such prone to high sequence variation. Variation is caused by accumulation and selection of point mutations, which has been observed to occur in Hgps (Garmendia et al. 2014). Analysis of Hgp sequences revealed high variation between and within strains at Hgp surface exposed regions. i.e. if a genome had two or more copies of *hgpB*, each copy produces a distinct variant of that protein. Within Hgp groups (i.e. within HgpB alone) we observed high sequence variation at sites predicted to be surface exposed. The exception to this variability was the surface accessible heme-binding core (red in Fig. 2), which retained high sequence identity in groups and also across all Hgps (HgpA–G; Figures S3 and S4, Supporting Information). The heme-binding core was identified through

submission of the AlphaFold model to 3DLigandSite online services. Further analysis of these variable surface domains (VDs) in HgpB showed conserved sequences were present in these regions (Fig. 2D) that could be split into different allelic variants based on sequence identity. However, even sequences that we classified as the same were not identical, likely due to accumulation of mutation/polymorphisms, so we used a cut-off of > 80% to classify these sequences as the same allelic variant or not within each VD. For example, the smallest variable domain, VD5, (Fig. 2D), could be classified as two distinct sequences. There was > 80% identity within sequences we classify as the same, but only ∼30% identity between the two different sequences present at this VD. The major variable domain, VD1, in HgpB (Fig. 2D) was highly diverse, with over 25 different sequence variants present.

Individual *H. influenzae* strains can encode multiple, duplicated *hgpB* and *hgpC* genes

Following our systematic analysis of Hgp sequences to classify Hgps consistently, we examined the number of *hgp* genes in both the publicly available fully annotated *H. influenzae* genomes present in NCBI Genbank ($n = 75$), and an invasive NTHi isolate collection ($n = 74$). Genbank contains a variety of both carriage and disease isolates. Invasive NTHi isolates used for this study were isolated from patients suffering from *H. influenzae* infections in SE Queensland over a 15-year period (2001–2015; Staples et al. 2017). Information on age, sample site, and geographical location were collected, but not on comorbidities (Staples et al. 2017). The prevalence of each of the proposed groups is presented in Table 1(a/b), with the number of genes encoded per strain, and the diversity of each of the *hgp* genes present broadly consistent between strains with publicly available genomes, and our invasive isolate collection (Table S1d, Supporting Information). For example, *hgpA* was found in 28% of NCBI genomes and 27% of invasive isolates, and *hgpB* in all genomes and all invasive isolates, with duplicates, i.e. many strains encoded multiple copies of both *hgpB* and *hgpC*. Multiple functional *hgpB* genes were present in ∼36% of strains, and ∼13% of strains encoded multiple *hgpC* genes (Table S1e, Supporting Information). Examining the sequences of these multiple genes from individual strains demonstrated that these were typically different allelic variants of the same *hgp* gene (Table S1d, Supporting Information). Our analysis also demonstrated that *hgpD* was more prevalent in invasive isolates vs. publicly available genomes (∼20% vs. ∼5%).

hgpB or *hgpC* genes are phase-varied ON in almost 60% of invasive isolates

At least one of *hgpB* or *hgpC* are present in all NTHi invasive isolates and publicly available genomes (Table S2, Supporting Information), suggesting they play an important role in NTHi survival. To determine if there was a selection for either *hgpB* and *hgpC* phase-variation during invasive infection, we carried out fragment length analysis of the $CCAA_{(n)}$ SSR tract present in the *hgpB* and *hgpC* genes using gene specific primers. This analysis demonstrated that 31.5% of *hgpB* genes were ON (Table 2a-i), i.e. expressed, and that 22.2% of *hgpC* genes were ON (Table 2a-ii). Previous *in vitro* growth studies have shown only one functioning *hgp* gene is needed to retain successful heme utilization from haptoglobin (Morton et al. 1999). Infectivity is also retained by the presence of a single *hgp* gene *in vivo* (in the infant rat model; Seale et al. 2006). Because of these factors, and as there are duplicate *hgpB* and *hgpC* genes in multiple genomes, we also examined how many of the invasive isolates had at least one *hgpB/C* gene ON (Table 2b). A total of 58.3% of invasive isolates have one of either *hgpB* or *hgpC* ON. These results suggest there is no Hgp (A–G) predominantly required for invasive infection, but increased expression of just one of either *hgpB* or *hgpC* does occur in invasive disease.

Discussion

Haemophilus influenzae has an absolute growth requirement for iron and heme, making all genes associated with iron and heme uptake relevant to disease, and potentially vaccine development. We have evaluated the distribution of Hgps in fully annotated *H. influenzae* genomes available in NCBI Genbank, the majority of which were NTHi strains, and in an invasive NTHi isolate collection, and propose a unified nomenclature for categorizing Hgps. The prevalence at which we observed Hgps were similar between the invasive isolate collection and fully annotated publicly available genomes (Table S1d, Supporting Information) with the exception of *hgpD*, which is present in ∼20% of invasive NTHI isolates vs only ∼5% of publicly available genomes (Table 1). Geographical differences between publicly available genomes (world-wide) vs. our invasive isolates (SE QLD, Australia) may have influenced the prevalence of HgpD as these invasive isolates likely represent a subset of strains circulating in the SE QLD region, although an importance for *hgpD* in invasive NTHi disease cannot be ruled out. There was no particularly dominant sequence type (using MLST) in either the invasive isolates (Staples et al. 2017) nor public genomes (Table S2, Supporting Information), and each contained a seemingly random selection of ∼50 different sequence types.

We have identified HgpE, HgpF, and HgpG as separate proteins within the repertoire of Hgps encoded by *H. influenzae* and branched existing groups from HgpA into allelic variants HgpA1 and A2. Of particular interest were *hgpB* and *hgpC*, as one of these genes was present in all strains. *hgpB* was found twice in ∼38% of NTHi strains, and ∼35% of invasive NTHi isolates. A subset of strains also contained multiple *hgpC* genes with 15% of strains and ∼14% of invasive NTHi isolates encoding two HgpC proteins. As no studies have examined the impact of duplicate *hgp* genes, it is unclear if these duplications provide an advantage other than that of simply having an extra variable *hgp* gene. HgpB has been reported to have a higher affinity for haptoglobin than HgpC (Seale et al. 2006), which may explain *hgpB* being more abundant in isolates. However, the same study also reported that HgpA has a higher affinity to haptoglobin than HgpC, and HgpA was only found in ∼28% of genomes whereas HgpC was in 96% of strains examined, so binding affinity alone perhaps does not explain the increased presence of *hgpB*.

As *hgp* genes undergo phase variation, we examined the expression state (ON vs. OFF) of *hgpB* and *hgpC* in an invasive NTHi isolate collection. We found that neither of these genes were primarily ON in this collection. However, ∼58% of isolates had at least one *hgpB* or *hgpC* ON. Importantly, expression of just a single Hgp allows successful growth and colonization (Morton et al. 1999, Seale et al. 2006). As such, we suggest the importance of Hgps is not dependent on one particular type (HgpA–G), but rather the *number* of expressed *hgp* genes. *Haemophilus influenzae* must maintain iron homeostasis to survive, and encoding multiple functional *hgp* genes offers increased contingencies against immune pressure. A correlation between an increase in the number of available Hgps and virulence has been observed

Table 2. (a) The expression state (phase-varied ON or OFF) of *hgpB* (i) *and hgpC* (ii) in an invasive isolate collection was assessed via fragment length analysis. All strains had at least one *hgpB* and 96% had at least one *hgpC* (Figure S3, Supporting Information). (b) A summary of invasive isolates with at least one *hgpB*, *hgpC* and *any* of the *hgpB* or *hgpC* genes in-frame/ON. See Figure S3 (Supporting Information) for all data. We were unable to amplify a PCR product for any *hgp* gene products from two of the invasive isolates, so were not included.

a (i)	*hgpB*			
	OFF	ON	Mixed	Total
No.	58	28	3	89
%	65.2	31.5	3.4	100
	Gene presence in 72 samples = 123.6%			
(ii)	*hgpC*			
	OFF	ON	Mixed	Total
No.	61	18	2	81
%	75.3%	22.2%	2.5%	100%
	Gene presence in 72 samples = 112.5%			

(b)	At least one *hgp* gene ON in genomes		
	hgpB	*hgpC*	*hgpB/C*
No.	32	19	42
%	44.4	26.4	58.3

previously, supporting this conjecture, but more work needs to be carried out to prove this.

Our analysis demonstrated a large amount of sequence diversity in surface exposed domains of Hgps, particularly the major surface domain. Single nucleotide polymorphisms (SNPs) have been previously observed to be selected for at predicted surface encoding domains of *hgpB* and *hgpC* during persistent infection, suggesting microevolution of Hgps during infection (Garmendia et al. 2014). Microevolution has also been observed for *hgpA* from sequential samples from COPD patients (Pettigrew et al. 2018), and *hgpC* during subsequent rounds of OM (Harrison et al. 2020). Selective pressure has been seen to drive changes in immune accessible regions in proteins, such as Opa and pili, in *Neisseria* spp (Malorny et al. 1998, Rotman et al. 2016, Sadarangani et al. 2016), although our sequence analysis does not provide evidence for the exact mechanism by which the sequence variation of Hgps occurs, and requires significant further study. We did find specific sequences common to VDs, which also appear prone to acquiring SNPs. It is perhaps unsurprising that the major surface domains had the highest sequence variability, as these regions are likely the most immune accessible and, therefore, prone to selective pressures. The expression of Hgps is likely complicated and dynamic; driven by factors such as number of *hgp* genes encoded, iron source availability, activity of other iron-uptake systems, and pressure from the host immune system. To successfully use Hgps as candidates in a rationally designed vaccine against NTHi, their ability to phase-vary needs to be considered, as does the sequence variability at immune accessible surface domains. Interestingly, the heme-binding core of Hgps appeared to be highly conserved and immune accessible, providing a rationale for including this region in any vaccine formulation containing Hgps.

Our analysis provides a rationalized naming scheme to classify the Hgps of *H. influenzae*. We have demonstrated the diversity and prevalence of these iron acquisition factors within this important human pathogen. We also show that a subset of these proteins, HgpB and HgpC, are present in all NTHi isolates, with expression of at least one likely during invasive disease. This expression during a key stage of disease, and the conserved nature of the heme-binding region, means Hgps, through targeting the heme-binding core, could be considered as components of a rationally designed subunit vaccine against NTHi. The inclusion of Hgps, perhaps as a protein fragment containing the heme-binding core, would target a key protein family required for NTHi growth and survival, and ensure the efficacy of an NTHi vaccine to target all strains.

Supplementary data

Supplementary data are available at *FEMSLE* online.

Funding

This work was supported by the Australian Research Council (ARC) Discovery Project grant DP180100976 to J.M.A. We thank Griffith University for providing Z.P. with a PhD scholarship. Publication and research costs of this work were supported by a generous donation from the Bourne Foundation, Melbourne, Australia.

Acknowledgements

We thank the Australian Genome Research Facility (AGRF), Brisbane, for carrying out fragment length separation.

Conflicts of interest. Z.N.P., A.V.J., M.S., and J.M.A.: no conflicts of interest. T.S. and P.W. are employees of BacVax, Inc., United States, which is seeking to develop an NTHi vaccine.

References

Atack JM, Srikhanta YN, Fox KL *et al.* A biphasic epigenetic switch controls immunoevasion, virulence and niche adaptation in non-typeable *Haemophilus influenzae*. *Nat Commun* 2015;**6**:7828.

Choby JE, Skaar EP. Heme synthesis and acquisition in bacterial pathogens. *J Mol Biol* 2016;**428**:3408–28.

Cope LD, Hrkal Z, Hansen EJ. Detection of phase variation in expression of proteins involved in hemoglobin and hemoglobin-haptoglobin binding by nontypeable *Haemophilus influenzae*. *Infect Immun* 2000;**68**:4092–101.

Dixon K, Bayliss CD, Makepeace K *et al.* Identification of the functional initiation codons of a phase-variable gene of *Haemophilus influenzae*, lic2A, with the potential for differential expression. *J Bacteriol* 2007;**189**:511–21.

Fox KL, Atack JM, Srikhanta YN *et al.* Selection for phase variation of LOS biosynthetic genes frequently occurs in progression of non-typeable *Haemophilus influenzae* infection from the nasopharynx to the middle ear of human patients. *PLoS ONE* 2014;**9**:e90505.

Garmendia J, Viadas C, Calatayud L *et al.* Characterization of nontypable *Haemophilus influenzae* isolates recovered from adult patients with underlying chronic lung disease reveals genotypic and phenotypic traits associated with persistent infection. *PLoS ONE* 2014;**9**:e97020.

Green LR, Lucidarme J, Dave N *et al.* Phase variation of NadA in invasive *Neisseria meningitidis* isolates impacts on coverage estimates for 4C-MenB, a MenB vaccine. *J Clin Microbiol* 2018;**56**:e00204–18.

Harrison A, Hardison RL, Fullen AR *et al.* Continuous microevolution accelerates disease progression during sequential episodes of infection. *Cell Rep* 2020;**30**:2978–88.e3.

Harrison A, Dyer DW, Gillaspy A *et al.* Genomic sequence of an otitis media isolate of nontypeable *Haemophilus influenzae*: comparative study with *H. influenzae* serotype d, strain KW20. *J Bacteriol* 2005;**187**:4627–36.

Jin H, Ren Z, Whitby PW *et al.* Characterization of hgpA, a gene encoding a haemoglobin/haemoglobin-haptoglobin-binding protein of *Haemophilus influenzae*. *Microbiology* 1999;**145**:905–14.

Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with alphafold. *Nature* 2021;**596**:583–89.

Kelley LA, Mezulis S, Yates CM *et al.* The phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 2015;**10**:845–58.

Ladhani S, Slack MP, Heath PT *et al.* Invasive *Haemophilus influenzae* disease, Europe, 1996–2006. *Emerg Infect Dis* 2010;**16**:455–63.

Maciver I, Latimer JL, Liem HH *et al.* Identification of an outer membrane protein involved in utilization of hemoglobin-haptoglobin complexes by nontypeable *Haemophilus influenzae*. *Infect Immun* 1996;**64**:3703–12.

Madeira F, Park YM, Lee J *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* 2019;**47**:W636–W41.

Malorny B, Morelli G, Kusecek B *et al.* Sequence diversity, predicted two-dimensional protein structure, and epitope mapping of Neisserial opa proteins. *J Bacteriol* 1998;**180**:1323–30.

Morton DJ, Stull TL. Distribution of a family of *Haemophilus influenzae* genes containing CCAA nucleotide repeating units. *FEMS Microbiol Lett* 1999;**174**:303–9.

Morton DJ, Whitby PW, Jin H *et al.* Effect of multiple mutations in the hemoglobin- and hemoglobin-haptoglobin-binding proteins, HgpA, HgpB, and HgpC, of *Haemophilus influenzae* type b. *Infect Immun* 1999;**67**:2729–39.

Morton DJ, Van Wagoner TM, Seale TW *et al.* Utilization of myoglobin as a heme source by *Haemophilus influenzae* requires binding of myoglobin to haptoglobin. *FEMS Microbiol Lett* 2006;**258**:235–40.

Morton DJ, Bakaletz LO, Jurcisek JA *et al.* Reduced severity of middle ear infection caused by nontypeable *Haemophilus influenzae* lacking the hemoglobin/hemoglobin–haptoglobin binding proteins (Hgp) in a chinchilla model of otitis media. *Microb Pathog* 2004;**36**:25–33.

Pettigrew MM, Ahearn CP, Gent JF *et al.* *Haemophilus influenzae* genome evolution during persistence in the human airways in chronic obstructive pulmonary disease. *Proc Natl Acad Sci* 2018;**115**:E3256–e65.

Phillips ZN, Brizuela C, Jennison AV *et al.* Analysis of invasive nontypeable *Haemophilus influenzae* isolates reveals selection for the expression state of particular phase-variable lipooligosaccharide biosynthetic genes. *Infect Immun* 2019;**87**:e00093–19.

Poole J, Foster E, Chaloner K *et al.* Analysis of nontypeable *Haemophilus influenzae* phase-variable genes during experimental human nasopharyngeal colonization. *J Infect Dis* 2013;**208**:720–27.

Ren Z, Jin H, Morton DJ *et al.* hgpB, a gene encoding a second *Haemophilus influenzae* hemoglobin- and hemoglobin-haptoglobin-binding protein. *Infect Immun* 1998;**66**:4733–41.

Ren Z, Jin H, Whitby PW *et al.* Role of CCAA nucleotide repeats in regulation of hemoglobin and hemoglobin-haptoglobin binding protein genes of *Haemophilus influenzae*. *J Bacteriol* 1999;**181**:5865–70.

Rotman E, Webber DM, Seifert HS. Analyzing *Neisseria gonorrhoeae* pilin antigenic variation using 454 sequencing technology. *J Bacteriol* 2016;**198**:2470–82.

Sadarangani M, Hoe CJ, Makepeace K *et al.* Phase variation of opa proteins of *Neisseria meningitidis* and the effects of bacterial transformation. *J Biosci* 2016;**41**:13–9.

Seale TW, Morton DJ, Whitby PW *et al.* Complex role of hemoglobin and hemoglobin-haptoglobin binding proteins in *Haemophilus influenzae* virulence in the infant rat model of invasive infection. *Infect Immun* 2006;**74**:6213–25.

Staples M, Graham RMA, Jennison AV. Characterisation of invasive clinical *Haemophilus influenzae* isolates in Queensland, Australia using whole-genome sequencing. *Epidemiol Infect* 2017;**145**:1727–36.

Van Eldere J, Slack MP, Ladhani S *et al.* Non-typeable *Haemophilus influenzae*, an under-recognised pathogen. *Lancet Infect Dis* 2014;**14**:1281–92.

Wang S, Sun S, Li Z *et al.* Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol* 2017;**13**:e1005324–e24.

Wass MN, Kelley LA, Sternberg MJ. 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res* 2010;**38**:W469–73.

Waterhouse AM, Procter JB, Martin DMA *et al.* Jalview version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 2009;**25**:1189–91.

Whitby PW, Seale TW, Morton DJ *et al.* Antisera against certain conserved surface-exposed peptides of nontypeable *Haemophilus influenzae* are protective. *PLoS ONE* 2015;**10**:e0136867–e67.

Whitby PW, Seale TW, VanWagoner TM *et al.* The iron/heme regulated genes of *Haemophilus influenzae*: comparative transcriptional profiling as a tool to define the species core modulon. *BMC Genomics* 2009;**10**:6.

Whitby PW, VanWagoner TM, Seale TW *et al.* Comparison of transcription of the *Haemophilus influenzae* iron/heme modulon genes in vitro and in vivo in the chinchilla middle ear. *BMC Genomics* 2013;**14**:925.

Xie J, Juliao PC, Gilsdorf JR *et al.* Identification of new genetic regions more prevalent in nontypeable *Haemophilus influenzae* otitis media strains than in throat strains. *J Clin Microbiol* 2006;**44**:4316–25.

Zheng W, Zhang C, Li Y *et al.* Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Reports Methods* 2021;**1**:100014.