Infection, Genetics and Evolution 67 (2019) 126–135

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

# Insights into the population structure and pan-genome of *Haemophilus influenzae*

M. Pinto[a], A. González-Díaz[b,c], M.P. Machado[d], S. Duarte[e], L. Vieira[e,f], J.A. Carriço[d], S. Marti[b,c], M.P. Bajanca-Lavado[g], J.P. Gomes[a,*]

[a] *Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health, Lisbon, Portugal*
[b] *Microbiology Department, Hospital Universitari Bellvitge, IDIBELL, University of Barcelona, Barcelona, Spain*
[c] *Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Madrid, Spain*
[d] *Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal*
[e] *Technology and Innovation Unit, Department of Human Genetics, National Institute of Health, Lisbon, Portugal*
[f] *Centre for Toxicogenomics and Human Health (ToxOmics), Genetics, Oncology and Human Toxicology, Nova Medical School, New University of Lisbon, Portugal*
[g] *Haemophilus influenzae Reference Laboratory, Department of Infectious Diseases, National Institute of Health, Lisbon, Portugal*

ARTICLE INFO

*Keywords:*
*Haemophilus influenzae*
Non-typeable
Pan-genome
Pathogenesis
Whole-genome sequencing

ABSTRACT

The human-restricted bacterium *Haemophilus influenzae* is responsible for respiratory infections in both children and adults. While colonization begins in the upper airways, it can spread throughout the respiratory tract potentially leading to invasive infections. Although the spread of *H. influenzae* serotype b (Hib) has been prevented by vaccination, the emergence of infections by other serotypes as well as by non-typeable isolates (NTHi) have been observed, prompting the need for novel prevention strategies. Here, we aimed to study the population structure of *H. influenzae* and to get some insights into its pan-genome. We studied 305*H. influenzae* strains, enrolling 217 publicly available genomes, as well as 88 newly sequenced *H. influenzae* invasive strains isolated in Portugal, spanning a 24-year period. NTHi isolates presented a core-SNP-based genetic diversity about 10-fold higher than the one observed for Hib. The analysis of key factors involved in pathogenesis, such as lipooligosaccharides, hemagglutinating pili and High Molecular Weight-adhesins, suggests that NTHi shape its virulence repertoire, either by acquisition and loss of genes or by SNP-based diversification, likely towards host immune evasion and persistence. Discreet NTHi subpopulations structures are proposed based on core-genome supported with 17 candidate genetic markers identified in the accessory genome. Additionally, this study provides two bioinformatics tools for in silico rapid identification of *H. influenzae* serotypes and NTHi clades previously proposed, obviating laboratory-based demanding procedures. The present study constitutes an important genomic framework that could lay way for future studies on the genetic determinants underlying invasiveness and disease and population structure of *H. influenzae*.

## 1. Introduction

The gram-negative human-restricted bacterium *Haemophilus influenzae* usually colonizes the respiratory tract in both children and adults. It can cause community acquired pneumonia, otitis media, conjunctivitis and sinusitis, and has been associated with chronic obstructive pulmonary disease and cystic fibrosis in patients with co-morbidities. Moreover, this pathogen also causes severe invasive disease infections in all age groups, such as epiglottis, septicemia and meningitis (Agrawal and Murphy, 2011; Garmendia et al., 2012; Rao et al., 1999; Duell et al., 2016). *H. influenzae* can be divided based on polysaccharide capsule production into encapsulated and non-encapsulated strains. The latter are usually referred to as non-typeable *H. influenzae* (NTHi), while encapsulated strains can be further categorized into serotypes based on their distinct capsular antigens (serotypes a to f) (Pittman, 1931). *H. influenzae* serotype b (Hib) were responsible for > 95% of invasive disease until the introduction, in the early 1990s, of a conjugate vaccine against this serotype (Agrawal and Murphy, 2011; Peltola, 2000). While the worldwide use of this vaccine led to the near elimination of Hib, it promoted the emergence of NTHi strains and non-b serotypes, as the main cause of *H. influenzae* associated invasive disease cases nowadays (Agrawal and Murphy, 2011; Tsang and Ulanova, 2017; Slack, 2017; Desai et al., 2015; Ulanova and Tsang, 2014; Sadeghi-Aval et al., 2013; Calado et al., 2011; Rubach

[*] Corresponding author at: Department of Infectious Diseases, National Institute of Health, Av. Padre Cruz, 1649-016 Lisbon, Portugal.
*E-mail address:* j.paulo.gomes@insa.min-saude.pt (J.P. Gomes).

https://doi.org/10.1016/j.meegid.2018.10.025
Received 6 July 2018; Received in revised form 20 October 2018; Accepted 30 October 2018
Available online 31 October 2018

et al., 2011; Giufrè et al., 2011; Campos et al., 2003). In Portugal, after the implementation of the Hib vaccination in 2000 on behalf of the National Immunization Program for all children of pre-school age, an increase of invasive NTHi-associated infections (19% to 77%) concomitant with a decrease of Hib-associated infections (81% to 13%) was observed, with more non-b serotypes also being reported since then (Bajanca-Lavado et al., 2014). This overall shift towards an increase of NTHi prevalence (Wan Sai Cheong et al., 2015; Resman et al., 2011) highlights the importance of studying the mechanisms of invasiveness of these non-encapsulated strains (Garmendia et al., 2012; Duell et al., 2016; De Chiara et al., 2014).

On a genomic basis, encapsulated types of *H. influenzae* constitute a limited set of clonal populations, while NTHi strains have revealed enormous genomic diversity (De Chiara et al., 2014; Eutsey et al., 2013; Hogg et al., 2007; Erwin and Smith, 2007; Kress-Bennett et al., 2016), mainly due to their intrinsic transformable nature (Power et al., 2012; Mell et al., 2011; Gilsdorf et al., 2004; Erwin et al., 2005) and to the high rate of recombination events they can undergo (Power et al., 2012; Takahata et al., 2007; Price et al., 2014). As such, research has focused on the understanding of NTHi pathogenesis either by identifying surface immunogenic antigens that can stimulate an immune response during infection or by characterizing virulence genes and NTHi-host interacting proteins. For example, several studies showed that the adherence phase occurring within the human respiratory tract mucosa is critical and may play an important role in the pathogenicity of NTHi (Fox et al., 2014; Baddal et al., 2015). In order to counteract and evade the attack of complement mediated killing system, NTHi strains also seem to have developed sophisticated strategies, such as employing phase variation mechanisms for the ON/OFF switching of surface adhesins to promote immune evasion and persistence [like lipooligosaccharides (LOS), hemagglutinating pili and High-Molecular Weight (HMW) adhesins] (Rao et al., 1999; Duell et al., 2016; Lichtenegger et al., 2017; Clark et al., 2013; Wong and Akerley, 2012).

The growing evolution and availability of whole genome sequencing (WGS) technologies has increased our knowledge on the genomic structure of pathogenic bacteria. Particularly, previous genome-scale studies on *H. influenzae* have provided insights into the genetic diversity and transmission of virulence factors, drug resistance determinants and pathogenic mechanisms that capacitate NTHi for colonization, survival and persistence in susceptible hosts (Eutsey et al., 2013; Price et al., 2014; Wong and Akerley, 2012; Garmendia et al., 2014; Hu et al., 2016). However, only few studies have performed this analysis with a large set of isolates (De Chiara et al., 2014; Power et al., 2012; Price et al., 2014). In the present work, we characterized the genomes of invasive *H. influenzae* strains isolated in Portugal, spanning a 24-year period, together with already public available *H. influenzae* genomes, in order to gain insights into the pan-genome and population structure of this species. By applying this pan-genomic analysis, we also aimed at further determining the specific genomic make-up of NTHi subpopulations, which may increase our knowledge on the virulence and invasiveness of this pathogen. In the frame of the transition to WGS-based surveillance, this study also introduces bioinformatics tools for in silico rapid identification of *H. influenzae* serotypes and previously proposed NTHi clades, obviating future laboratory-based procedures.

## 2. Material and methods

### 2.1. Strains' phenotypic characterization and whole-genome sequencing

*H. influenzae* strains enrolled in the present study were obtained from the collection of the Portuguese National Institute of Health. Eighty-eight invasive disease causing isolates spanning a 24-year period (from 1992 up to 2016) were selected for WGS (Supplementary Table S1) including 57 NTHi and isolates from distinct serotypes (a, b, d, e and f). Isolates were cultured from frozen stocks on Chocolate Agar supplemented with Polyvitex (BioMérieux) at 35 °C for 18–24 h in 5% $CO_2$ atmosphere. Isolate capsular status was identified by PCR amplification of *bexA* gene (involved in capsule transport) and capsular type was determined by amplification of capsule-specific genes (for serotypes a-f) using primers and conditions previously described (Falla et al., 1994). DNA was extracted from each isolate using a silica-based automatic DNA extractor (EasyMag) and subsequently subjected to the Nextera XT library preparation protocol (Illumina) prior to paired-end sequencing (2 × 250 bp) on a MiSeq apparatus (Illumina) according to the manufacturer's instructions.

Genome sequences were assembled using the INNUca v2.6 pipeline (MP Machado et al., *INNUca* GitHub https://github.com/B-UMMI/INNUca), an integrative bioinformatics pipeline for reads quality analysis and de novo genome assembly. Briefly, reads' quality is firstly analyzed and reported with FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) followed by cleaning/quality improvement using Trimmomatic v0.36 (Bolger et al., 2014) (with a sliding window of 5 with a minimum average quality of 20; minimum read length of 55 bp; removal of bases with quality below 3 from the start and end of the read; and a cut of bases, at the start and end of the reads, with biased GC content determined automatically in a sample-specific manner based on FastQC report, typically 15–16 bp at start and 1–2 bp at the end for the sequencing strategy used). Genomes are assembled with SPAdes v3.10 (Bankevich et al., 2012) and subsequently polished using Pilon v1.18 (Walker et al., 2014), with QA/QC statistics (such as depth of coverage and number of contigs) being monitored and reported throughout the analysis. Finally, MLST prediction is performed using the *mlst* v2.4 software (Seemann T, mlst Github https://github.com/tseemann/mlst). Novel identified MLST profiles and gene sequences were submitted the PubMLST website (http://pubmlst.org). Isolates full characterization, including sampling date, sequence type (ST), antibiotic resistance (phenotype and genotype) data, final genome assembly sizes and depth of coverage values are reported in Supplementary Table S1.

To rapidly determine *H. influenzae* serotypes in silico directly using WGS reads, we added a module to the *seq_typing* tool (MP Machado et al., *seq_typing* GitHub https://github.com/B-UMMI/seq_typing). Briefly, *seq_typing* is a software to determine a given sample type using either a read mapping approach or a sequence Blast search (Altschul et al., 1990) against a set of reference sequences. For the read mapping approach (the one used in the present work), the sample's reads are mapped to serotype specific capsular genes sequences using Bowtie2 v2.2.9 (Langmead and Salzberg, 2012), parsed with Samtools v1.3.1 (Li, 2011) and analyzed via ReMatCh v3.3 (MP Machado et al. *ReMatCh* GitHub https://github.com/B-UMMI/ReMatCh). Based on the length of the sequence covered and it's depth of coverage, *seq_typing* returns the type associated with the reference sequence which is more likely to be present. In the case of *H. influenzae*, the defined type is the serotype. The selected sequence will be the one covered to a greater extent and with higher depth of coverage, that passes defined thresholds. For this new feature, we took advantage of the previously published serotype specific capsule biosynthesis loci: *acsABCD* (GenBank accession CP017811) for Hia; *bcsABCD* (FQ312006) for Hib; *ccsABCD* (HQ651151) for Hic; *dcsABCDE* (HM770877) for Hid; *ecsABCDEFGH* (HM770878) for Hie and *fcsABC* (CP005967) for Hif.

### 2.2. Core-genome-based phylogenic and diversity analysis

For the novel *H. influenzae* genome sequences obtained from strains recovered in Portugal (hereinafter designated as “PTHi” genomes), their genetic diversity was first analyzed in the frame of the background of previously published genome sequences. To achieve this, we constructed an assembly-based core-SNP phylogeny using *parsnp* from the Harvest suite (Treangen et al., 2014) with default parameters (with exception of parameter -C, which was adjusted to 2000 in order to maximize the reference coverage). Parsnp performs whole-genome alignment using multiple strategies and tools, and infers an approximate-maximum-likelihood phylogenetic tree using the core-genome SNPs. In this first dataset, we analyzed a total of 305 genomes (Supplementary Table S2), including: the 88 PTHi genomes; 127 closed and draft genome sequences available in the NCBI database; 89 WGS reads dataset available at the European Nucleotide Archive (ENA) (De Chiara et al., 2014) (also assembled here using the INNUca pipeline as previously described); and the Rd KW20 strain closed genome sequence (GenBank accession L42023) (Fleischmann et al., 1995) that was used as a reference. Additionally, a second dataset was analyzed, with the same methodology, but now enrolling only NTHi genome sequence data, that included all 57 Portuguese non-typeable *H. influenzae* (PTNTHi) isolates and 186 NTHi publicly available genomes to classify novel PTNTHi genomes into previously proposed clades that were defined using a discriminant principal component analysis based on core-SNPs (De Chiara et al., 2014).

In order to analyze the genetic diversity specifically within the 88 PTHi isolates (third dataset), an assembly-free core-SNP-based approach was applied, using Snippy v3.2 (https://github.com/tseemann/snippy). Briefly, reads were individually mapped to a selected draft assembled genome (PTHi-5709) and SNP calling was performed on variant sites with the following criteria: a minimum mapping quality of 20; a minimum coverage of 10; and a minimum proportion of variant frequency of 90%. Core-SNPs were extracted using Snippy's core module (*snippy-core*) and a Maximum Likelihood (ML) core-SNP-based phylogenetic tree was reconstructed using RAxML-NG v5.1 (https://github.com/amkozlov/raxml-ng), following the general time-reversible model, a maximum likelihood of substitution rates and nucleotide frequencies and a discrete GAMMA model of rate heterogeneity, with 100 bootstrap replicates. The whole genome alignment was subsequently subjected to the prediction and removal of recombinant regions using Gubbins v2.3.1 software (Croucher et al., 2015) with default parameters, in order to remove the phylogenetic noise produced by these regions (that will contain elevated densities of base substitutions and therefore would reflect distinctive phylogenetic histories from the rest of the genome). Subsequently, a novel core-SNP phylogenetic tree, using RAxML-NG, was reconstructed based only on shared positions after recombination removal. Finally, MEGA7 software (Kumar et al., 2016) was used to determine the overall mean distances and matrices of pairwise comparisons at the nucleotide level. All generated trees were drawn and visualized using FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

### 2.3. Genetic diversity of H. influenzae pathogenesis associated genes

To analyze the presence/absence and genetic diversity of genes known to be involved in *H. influenzae* pathogenesis, a custom database was created using 105 reference genes reported in the Virulence Factor Database (VFDB) (Chen et al., 2005) and in a previous study (Wong and Akerley, 2012) (detailed in Supplementary Table S3). Subsequently, a gene-by-gene analysis using this custom database was performed on the novel PTHi genomes with the chewBBACA suite (Silva et al., 2018) using default parameters. Since the assembly of the HMW adhesin-coding genes using short read technology is problematic due to both their paralogous nature and multiple extensions of simple sequence repeats, as an exception, the relatively conserved regions within the first ~1200 bp, shared by *hmwA1* and *hmwA2* genes (Supplementary Table S3), were used to infer their absence or presence for isolates where these genes could not be called. This was performed both by reads mapping (using Snippy) and Blastn search [from the BLAST+ suite v2.5.0 (Camacho et al., 2008)], against the assembled PTHi draft genomes.

### 2.4. Pan-genomic analysis

PTHi assembled draft genomes were annotated with Prokka v1.12 (Seemann, 2014) prior to the pan-genomic analysis using Roary v3.8.0 software (Page et al., 2015) (with a Blastp minimum percentage of identity set to 70% and without splitting paralogs). Gene clusters found to be present exclusively in encapsulated or non-encapsulated strains were Blastn searched (90% query coverage with 70% identity) against publicly annotated *H. influenzae* closed or draft genomes (Supplementary Table S2) in order to obtain referential locus tags. To further characterize gene cluster functions of the PTHi accessory genome, assemblies were also analyzed using PHASTER (Arndt et al., 2016) in order to firstly determine the presence of intact phages and secondly to relate gene clusters to potential phage-associated genes. Results were visualized and are presented here using Phandango v1.1.0 interactive web application (Hadfield et al., 2017).

Additionally, a specific pan-genomic analysis was performed exclusively on NTHi genomes (second dataset: 57 PTNTHi plus 187 already available genomes) in order to identify accessory genes with the potential to discriminate previously defined NTHi clades (De Chiara et al., 2014). Briefly, after re-annotating all previously published NTHi genomes using Prokka, Roary analysis was performed as previously mentioned, and Scoary v1.6.10 (Brynildsrud et al., 2016) software was used to identify genes potentially discriminating NTHi clades. The presence/absence of all identified genes was subsequently confirmed by Blastn search against all 243 NTHi genome sequences, with a minimum identity of 90% and at least 20% of gene sequence coverage (the query sequence) to prevent false negatives in cases where these genes could have been split into different contigs during the assembly. Finally, using the obtained set of genes from the pan-genome analysis, we include in *patho_typing* tool (MP Machado et al. *patho_typing* GitHub https://github.com/B-UMMI/patho_typing) a module for NTHi clade typing. Shortly, *patho_typing* is a tool for in silico pathogenic typing sample's reads through a read mapping approach using a set of reference sequences and defined rules for sequences presence/absence. Similar to *seq_typing* tool using reads mapping approach, *patho_typing* maps reads to a set of reference sequences using Bowtie2, parse the result with Samtools and analyze it via ReMatCh. Based on the length of the sequence covered, it's depth of coverage and sequence nucleotide identity, *patho_typing* scores those for presence or absence, following defined thresholds. According to the combination of sequences present, a pathotype is returned following a set of rules for sequences presence/absence. In the caso of NTHi, NTHi clade from I to VI is returned according to De Chiara et al., 2014. For NTHi clade typing, the following 17 gene sequences are scored: HIFGL_RS05025, HIFGL_RS03555, HIFGL_RS07710, HIFGL_RS07705, HIFGL_RS09045, HIFGL_RS06855, HIFGL_RS05250 and HIFGL_RS07070 from strain KR494 (RefSeq accession NC_022356), C645_RS00645, C645_RS00650, C645_RS00655 and C645_RS08170 from strain 2019 (NZ_CP008740), HIB_RS06975 and HIB_RS07380 from strain 10,810 (NC_016809), R2846_RS08405 from strain R2846 (NC_017452) and *hmwA* (the first 1269 bp), *hmwB* and *hmwC* from strain NCTC8143 (NZ_LN831035). The genes presence/absence combinations can be found in *patho_typing* GitHub repository. This novel module was found to be accurate, as it was tested against all 44 publicly available WGS reads of *Haemophilus* species other than *influenzae* (Supplementary Table S2) and no single genome could be classified into the described clades, as the combination of present/absent genes presented in the study was never observed. Furthermore, for 36 out of the 44 genomes we found either none of the genes ($n = 27$) or only a single gene ($n = 9$) from the classification panel.

### 2.5. Data availability

A publicly available Bioproject in the ENA database was created to group all reads and assemblies exclusively produced during this study (Project accession number PRJEB26586). All individual run and assembly accession numbers are listed in Supplementary Table S1. The datasets generated during this study are included in the Supplementary Information. All external sequence data used in the present study, retrieved either from NCBI or ENA repositories (last checked December

1st 2017, at the time we started the analysis), are detailed in the Supplementary Files presented throughout the manuscript.

## 3. Results

### 3.1. H. influenzae global phylogeny and genomic diversity

Integration of all novel PTHi genomes into the *H. influenzae* global phylogeny (first dataset) revealed that these 88 isolates reflect the global population diversity of the species, being dispersed along the phylogenetic tree (Supplementary Fig. S1), with isolates of the same serotype segregating together. The assembly-based core-SNP analysis of all 305 genomes predicted that only approximately 47% of the genome is shared, with about 11% of this core genome presenting mutations (i.e., 113,841 polymorphic sites). The in silico determination of *H. influenzae* serotypes for PTHi genomes proved to be in full concordance with results obtained by PCR. Further testing performed for all genomes with publicly available WGS reads (Supplementary Table S2) also showed full concordance with reported serotypes. Moreover, we inferred the classification of all PTNTHi genomes within previously proposed clades (De Chiara et al., 2014) (using the second dataset) and observed that most sequenced NTHi isolates fall within clade V and VI (18 and 23 out of 57, respectively), while still having representative isolates from the remaining four clades (Supplementary Table S1). Of note, according to the metadata provided to the *Haemophilus influenzae* Reference Laboratory, the patients from whom the strains were collected were not epidemiologically linked, suggesting the inexistence of transmission events within the studies population.

A fine-tune assembly free core-SNP-based analysis exclusively enrolling all novel PTHi genomes (third dataset) revealed that these are distinguishable by 153,263 single nucleotide variant (SNV) sites (Supplementary Fig. S2A) with an overall pairwise mean ( ± SE) difference of 30,734 ± 47 between each random pair of isolates. Hib isolates showed an overall mean number of differences of 3214 ± 15, contrasting with the 32,899 ± 46 differences exhibited within the NTHi group, indicating that most of the observed genetic diversity is due to the NTHi (we observed an overall pairwise mean differences of 32,040 ± 67 between these two groups). Due to the high recombination rate known to occur in *H. influenzae* (Connor et al., 2012; Price et al., 2014; Maughan and Redfield, 2009), potential recombinant regions were removed from the core-SNP analysis as many SNPs may have arisen on the course of recombination (~91% of the genome was excluded). This exclusion resulted in only 296 SNVs separating all 88 PTHi isolates (Supplementary Fig. S2B) (with an overall mean number of differences of 40 ± 2), of which 40 fall within non-coding regions, 4 within tRNA, and 252 target a total of 19 genes (Supplementary Table S4). Within the latter, 77 mutations are non-synonymous and 175 are synonymous in relation to the reference genome used (e.g. PTHi-5709). Noteworthy, in both cases (with or without removal of recombinant regions; Supplementary Fig. S2), each specific serotype seems to segregate independently, potentially indicating a conserved genetic backbone of each group. Nevertheless, some less evident cases (e.g. Hia e Hid; Supplementary Fig. S1) require genomic information of additional strains to support this assumption.

### 3.2. Genetic diversity underlying pathogenesis and adaptation

We evaluated the presence/absence, as well as the allelic diversity, of 105 genes (Wong and Akerley, 2012; Chen et al., 2005) known to be associated with *H. influenzae* pathogenesis and adaptation (Fig. 1). These include genes involved in adherence, competence, cell surface, LOS, outer membrane proteins, iron transport, heme biosynthesis and heme-binding complexes. Results showed that several genes involved in adherence are more prone to be acquired or lost, rather than being targeted by genetic alterations, suggesting that *H. influenzae* may acquire a particular set of genes in order to adapt to the context of infection. In fact, we observed that the haemogglutinating pili locus (*hifABCDE* genes), associated with biofilm formation (Garmendia et al., 2012; Rao et al., 1999; St Geme III, 2002) were only found completely present in 7 out of the 88 isolates, and partially (at least one gene) in seven other isolates. Only 30 out of the 57 PTNTHi (53%) isolates presented the HMW adhesin-coding genes. Previous studies have suggested that approximately 75% of NTHi possess the HMW adhesin-coding genes (Rao et al., 1999). Moreover, within this gene category, *hsf* was found in almost all Hib isolates (except one: PTHi-1751), suggesting that it mediates a conserved mechanism of adherence in Hib, while *hia*, genetically close related to *hsf* (Rao et al., 1999; St Geme III et al., 1996), was found to be present in only 6 out of 57 (11%) of PTNTHi strains. Of note, we observed that the simultaneous presence of some adherence associated genes rarely occurred (Fig. 1), namely *hmw*, *hap* and *hia*. In contrast, the *hap* gene was found in most isolates (87 out of 88; 99%) presenting a high degree of genetic variability, with a total of 53 different allelic profiles, particularly for PTNTHi isolates (42 different alleles in 57 isolates). It has been suggested that this gene may have implications in low-level adherence but also serve to facilitate evasion of the immune response (Rao et al., 1999), which could explain its high polymorphic character. A high genetic diversity was likewise observed for *H. influenzae* outer membrane proteins, also believed to be involved in adherence (i.e. *oapA*, *ompP1*, *ompP2*, *ompP5* and *ompP6*) (Garmendia et al., 2012; Osman et al., 2018), which is expected as they maintain contact with the host environment. As previously observed between these genes, *ompP6* presents a rather lower genetic diversity than all the other analyzed outer membrane proteins (Garmendia et al., 2012; Rao et al., 1999).

As *H. influenzae* is known to acquire genetic material, the genes involved in competence were expected to be present in all isolates. In fact, the type IV pilus genes (*pilABCD*), which are necessary for competence, colonization and biofilm formation (Jurcisek et al., 2007), and the *com* genes (*comABCDEF*), homologs of the type II secretion systems and suggested to be involved in the biogenesis of the type IV pilus (Bakaletz et al., 2005), were found in all isolates (at least partially for the *pil* genes), presenting a moderate degree of genetic variability with ~30 different alleles *per* locus (Fig. 1).

As a means to promote persistence during infection, *H. influenzae* possesses mechanisms to acquire iron and heme complex (Rao et al., 1999). As such, we examined the iron transport locus (*hitABC*), heme biosynthesis genes (*hemHMNRXY*) and the genes involved in the heme-binding complex (*huxABC*). Notably, the genetic diversity of these genes was found to be high in PTNTHi, particularly in the heme-binding complex genes, while being more conserved within Hib isolates (Fig. 1). For example, while Hib isolates present up to three different alleles for the *hit* and *hem* genes and five for the *hux* genes, PTNTHi isolates present up to 26, 35 and 47 distinct alleles for the *hit*, *hem* and *hux* genes, respectively. Thus, although in a pure speculation basis, NTHi seem to be more prone to undergo diversification of genetic features likely associated with persistence.

We also examined genes associated with *H. influenzae* cell surface (*vacJ* and *yrbBCDEF*) that have been implicated in serum resistance by rearranging surface LOS, which promotes their survival in the lung. In fact, it has been reported that mutations in these genes can lead to increased antibody binding to this structure with loss of serum resistance (Garmendia et al., 2012; Wong and Akerley, 2012; Nakamura et al., 2011). While these genes seem to be mostly conserved within Hib, for PTNTHi some degree of genetic diversification was unexpectedly found (Fig. 1), which may be disadvantageous for survival in the serum, according to the above-cited literature.

Concerning the LOS associated proteins, long regarded as important virulence factors in *H. influenzae* (Garmendia et al., 2012; Rao et al., 1999; Garmendia et al., 2014), our results show that the corresponding genes are shaped both by high genetic diversification, mainly within PTNTHi, but also by differential presence/absence (Fig. 1). Twenty-four out of the 55 analyzed genes are absent in at least one isolate, most of

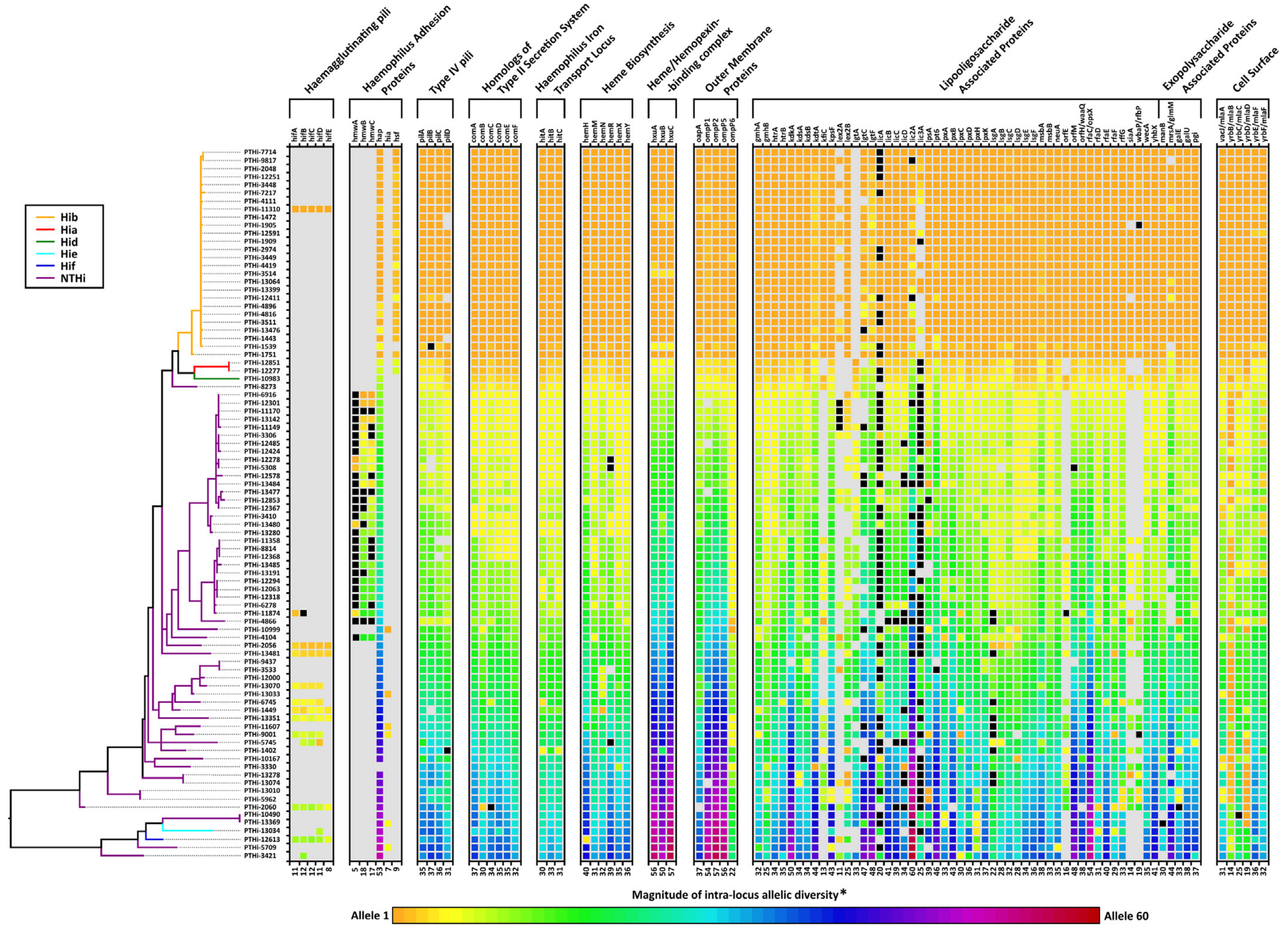

**Fig. 1.** Allelic diversity of genes involved in *H. influenzae* pathogenesis. The 105 genes presented in the figure are the ones described in the literature as being involved in pathogenesis and were retrieved from the Virulence Factor Database (VFDB) (Chen et al., 2005) and from Wong and Akerley, 2012. Reference sequences used for allele calling [using chewBBACA (Silva et al., 2018)] are described in Supplementary Table S3. Each line refers to a specific isolate and each column to a specific gene. Gene are clustered according to their function. Black squares represent genes likely to be present but for which the allele calling could not be performed using the applied methodology. For example, due to high homology, discrepancies in annotations and the presence of short repeats, the first 1269 bp of *hmwA* (shared by *hmw1A* and *hmw2A*) of reference strain NCTC8143 (LN831035) were used to additionally infer the presence of *hmwA* gene. Numbers at the bottom of each gene column refer to the maximum number of different alleles found for each gene. ML phylogenetic tree branches (produced through the assembly-free core-SNP approach without recombined regions) are highlighted by serotype. *For each gene, different alleles found are colour coded, so genes with wider magnitude of colours display higher allelic diversity. Similar to traditional MLST codes, proximal colours (here solely representing dissimilar allele identifiers) do not reflect genetic proximity. For simplification purposes, colour codes were attributed sequentially for each allelic identifier throughout each gene, from the first isolate at the top of the phylogenetic tree until the last isolate at the bottom. Grey areas represent loci not found for specific isolates. Figure was adapted after drawing it using Phandango (Hadfield et al., 2017).

them without any relationship with serotypes or NTHi clades. Nevertheless, four genes, namely *kfiC*, *orfE*, *siaA* and *wbap/rfbP*, were found to be absent in Hia, Hie, Hif and all PTNTHi isolates from clade I and V, and a subset of clade VI (i.e., PTHi-9437, PTHi-3533, PTHi-12,000, PTHi-13,070, PTHi-13,033, PTHi-1449, PTHi-5745 and PTHi-10,167). Although this could suggest a potential shared genomic structure between these isolates, this is not supported by their evolutionary history (Fig. 1). Additionally, one of the main challenges in performing WGS with short reads is the difficulty to assemble genes largely targeted by phase variation. This is the case for some of the *lic* and *lex* genes (Garmendia et al., 2012; Rao et al., 1999), due to the extension of simple sequence repeats (Power et al., 2009), for which allele calling could not be performed for most strains, impairing an in depth analysis of these genes. Nevertheless, with the exception of *lic3A* which is absent in almost half of the Hib isolates, the *lic* genes were found to be present in most strains contrasting with *lex2A* and *lex2B*.

### 3.3. Pan-genomic analysis

Pan-genomic analysis of novel PTHi isolates identified a total of 3424 genes, of which 1308 are shared by all isolates (core-genome) and 2116 compose the accessory genome (Fig. 2A). The obtained accessory genome was larger than any single genome (mean of 1767 coding sequences *per* genome), as previously observed (De Chiara et al., 2014; Hogg et al., 2007), suggesting that *H. influenzae* could have an open pan-genome, due to its competence in acquiring genomic material throughout the infection process. Within the accessory genome, 601 genes, distributed among the 88 genomes, were found to be exclusive of a single isolate (Fig. 2A). Of note, some values presented here for the pan-genome are likely underestimated. In fact, since we opted for a conservative approach, by clustering paralogs and genes with close homology, some genes targeted by phase variation for which the assembly was impaired were not consider.

In this dataset, we identified several genes exclusively present

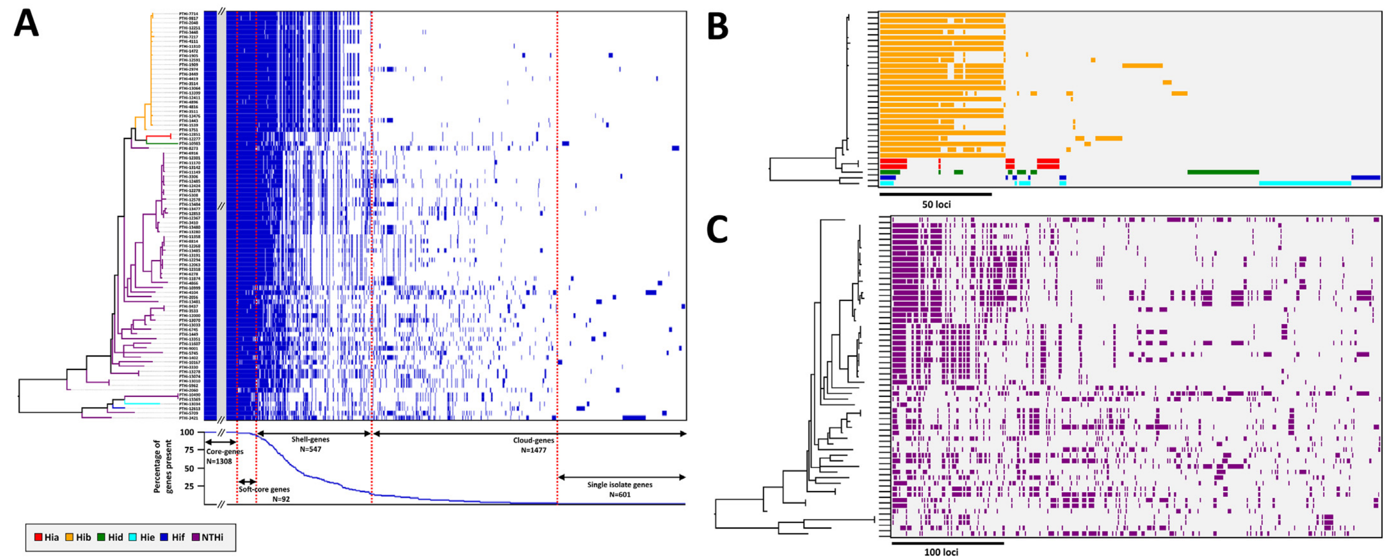

**Fig. 2.** Pan-genomic analysis of novel Portuguese *H. influenzae* isolates. A – Distribution of all genes identified using Roary (Page et al., 2015). ML phylogenetic tree branches (produced through the assembly-free core-SNP approach without recombined regions) are coloured by serotype. Genes were classified by their presence in the 88 analyzed PTHi isolates as: Core when in ≥99%; Soft-core when between 95 and 99%; Shell when between 15 and 95% and Cloud when < 15%. B – Distribution of genes (*N* = 223) found to be exclusively present within capsulated strains. Data on each gene, along with respective locus names, are described in Supplementary Table S5A. C – Distribution of genes found to be exclusively present in NTHi isolates, present in > 3 isolates (*N* = 433 out of 1185). Data on each gene, along with respective locus names, are described in Supplementary Table S5B. Figures were adapted after drawing them using Phandango (Hadfield et al., 2017), with graphs colours in panel B and C also highlighting each serotype for visualization purposes.

within capsulated strains (*N* = 223) (Fig. 2B). Data for each gene are described in Supplementary Table S5A. As expected, several exclusive genes relate to the capsule biosynthesis (28), but most are phage-associated proteins (74) or hypothetical proteins (89). Regarding PTNTHi, no single gene was found to be exclusively present and shared by all these isolates. Nevertheless, we found 433 genes (out of 1185 exclusive to PTNTHi) that were shared by a minimum of four and a maximum of 49 out of the 57 PTNTHi isolates (Fig. 2C and Supplementary Table S5B). The vast majority of these genes are phage-associated (*N* = 218) or hypothetical proteins (*N* = 100). This is agress with previous studies that also showed several phage-associated genes (Eutsey et al., 2013) and high heterogeneity of genes within NTHi accessory genome (De Chiara et al., 2014; Eutsey et al., 2013; Price et al., 2014). Our results showed that most *H. influenzae* isolates present distinct intact phages integrated in their genome, particularly in the case of PTNTHi. Curiously, several genes were found to be shared by NTHi clades, such as the HMW adhesin-coding genes, which were identified in 29 strains. Additionally, the *lsr* genes, a set of 9 genes (*lsrABCDEFGKR*) associated with quorum sensing (Li et al., 2007) with potential involvement in biofilm formation were found to be shared by 28% of the PTNTHi strains, and were also identified in 17% of previously published NTHi *H. influenzae* genomes while all encapsulated strains lack them.

Following this, a brief pan-genomic analysis was performed on a total of 243 NTHi genomes (second dataset) in order to identify potential populations structures within NTHi. After classifying NTHi isolates into previously proposed clades (De Chiara et al., 2014), 17 genes were identified that could be associated with each clade, constituting clade-specific signatures regarding the exclusive presence/absence of genes (Fig. 3). For example, clade I isolates exclusively possess and lack three and two genes, respectively. Interestingly, all isolates from clade I and VI (with three exceptions) lack the HMW adhesin-coding genes (Fig. 3). This suggests that the previously observed discrete NTHi population structure, based on core-SNPs (De Chiara et al., 2014) might be corroborated by the presence/absence of genes in the accessory genome. As such, we suggest that the differential identification of the described genes in Fig. 3 could constitute a valuable scheme to rapidly place NTHi within their respective subpopulations, which in turn might aid in establishing relationships with distinct pathogenesis and adaptation mechanisms, as well as epidemiological data.

## 4. Discussion

The present work aims to get insights into the pan-genome and population structure of *H. influenzae*. It enrols 305 strains including 88 invasive *H. influenzae* strains isolated in Portugal, spanning a 24-year period, together with 217 publicly available genomes. Since the Portuguese Hi genomes sequenced in the present study were congruently incorporated into the *H. influenzae* global phylogeny, it is reasonable to consider that the conclusions derived from the PTHi isolates dataset can be generalized to the *H. influenzae* species. The worldwide decrease in Hib related disease due to the effective implementation of a targeted vaccine has led to the emergence of NTHi and non-b serotypes as being responsible for the majority of invasive disease cases (Agrawal and Murphy, 2011; Tsang and Ulanova, 2017; Slack, 2017; Desai et al., 2015; Ulanova and Tsang, 2014; Sadeghi-Aval et al., 2013; Calado et al., 2011; Rubach et al., 2011; Giufrè et al., 2011; Campos et al., 2003). While it has been reported that capsulated strains belong to more clonal lineages (Gilsdorf et al., 2004; Meats et al., 2003; Musser et al., 1995), the identification of defined population structure within NTHi has remained challenging, mainly due to their competence for transformation and a high rate of homologous recombination during infection (Garmendia et al., 2014; Power et al., 2012; Connor et al., 2012; Takahata et al., 2007; Price et al., 2014; Poje and Redfield, 2003). In agreement with previous studies (Rao et al., 1999; De Chiara et al., 2014; Power et al., 2012; Musser et al., 1995; Staples et al., 2017), our results show that NTHi isolates are largely responsible for the genetic diversity observed within *H. influenzae* species. Furthermore, as expected, we observed that Hib isolates are closely related and present a strikingly lower level of intra-group diversity (about 10-fold lower) when compared with NTHi. This relies both on the analysis of the core-genome phylogeny (Supplementary Fig. S1A) and on the evaluation of genes associated with pathogenesis (Fig. 1). All Hib isolates continued to segregate together before and after removal of

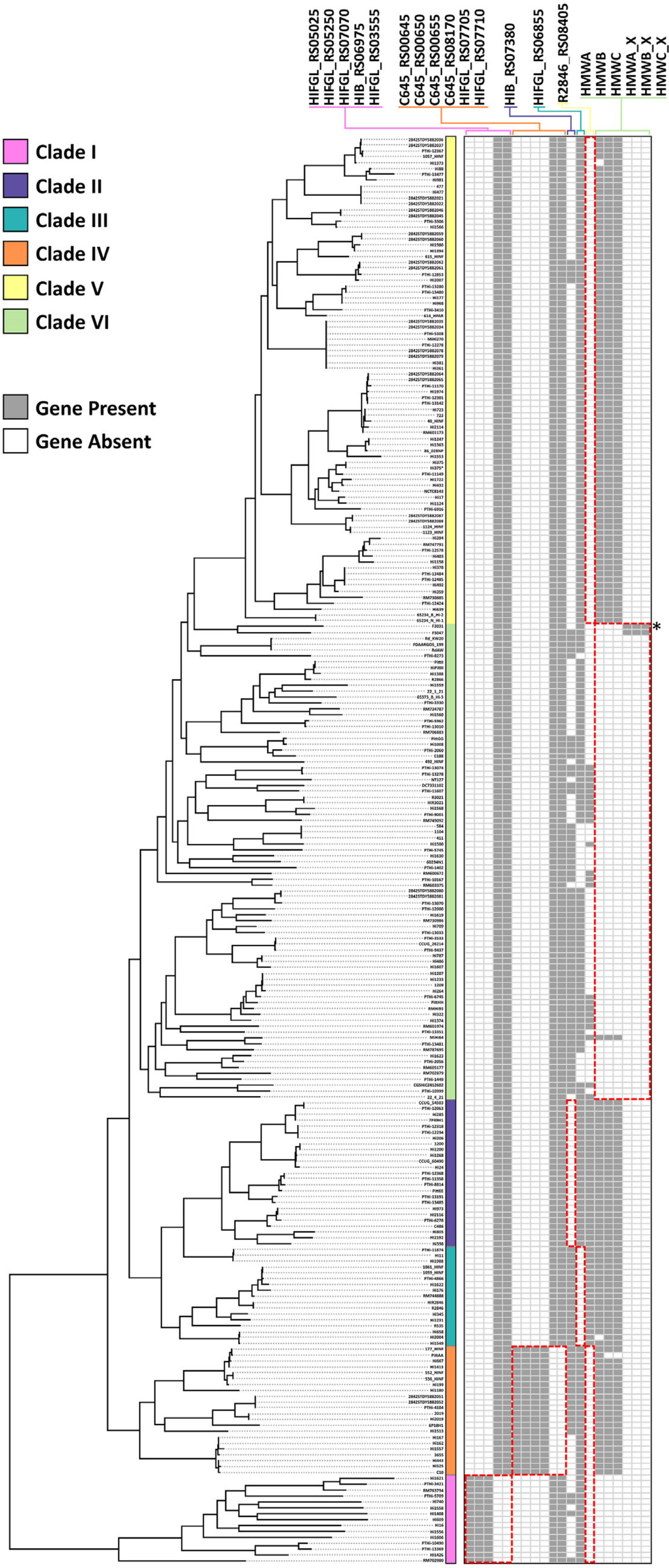

**Fig. 3.** Assembly-based core-SNP phylogenetic tree of non-typeable *H. influenzae*. The clade distribution is according to De Chiara et al., 2014. Tree was constructed using *parsnp* from the Harvest suite, including 243 NTHi genomes and represent a core genome of ~51%. Presence and absence of accessory genome genes representing a potential relationship with NTHi clades are also displayed. Clade-specific signatures regarding the exclusive presence/absence of genes are highlighted for each clade (red dashed boxes). Due to high homology, discrepancies in annotations and the presence of short repeats, the first 1269 bp of *hmwA* (shared by *hmw1A* and *hmw2A*) of reference strain NCTC8143 (LN831035) were used to infer its presence. *Refers to the *hmw* genes (*hmwABC*) of reference strains F3031 (GenBank accession FQ670178) and F3047 (FQ670204), where for *hmwA* the first 1344 bp were used to infer its presence. These are annotated as the *hmw* genes in the genome of these strains, although they present a very distinct gene sequence that is not found in any *H. influenzae* genome sequenced to date. As such, these were labelled in the present Figure as *hmw_X* (i.e., *hmwA_X*, *hmwB_X* and *hmwC_X*), as to differentiate them from the most commonly found sequences. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

potential recombination regions (Supplementary Fig. S1A), sustaining their strongly homogeneous genomic population structure. The overall maintenance of the genomic structure can be explained by the random dispersion of recombination events throughout the genome, which is known to be facilitated due to the tremendous competence for transformation of this species (Power et al., 2002; Mell et al., 2011; Gilsdorf et al., 2004; Erwin et al., 2005). Remarkably, when looking at the whole PTHi phylogeny after recombination removal, only 296 (out of

153,263) potential vertically inherited SNPs remain, which could potentially constitute good markers of the genetic backbone of each *H. influenzae* serotype group.

While *H. influenzae* pathogenesis begins with the colonization of the upper airways, it can spread from the nasopharynx to the sinuses, middle ear, trachea and lower airways, in some cases leading to the invasion of the central nervous system or bloodstream (Wong and Akerley, 2012), which was generally associated with capsulated strains (Garmendia et al., 2014). Usually restricted to the human respiratory tract, this pathogen can display a diverse repertoire of clinical manifestations, suggesting that the genetic diversification of key factors, particularly for NTHi, likely underlies its persistence and survival. In fact, we observed that about 90 out of the 105 (~95%) analyzed genes associated with *H. influenzae* pathogenesis present a high level of allelic variation in NTHi, while others are more prone to be acquired or lost (Fig. 1). Within the latter, adhesion associated genes *hifABCDE*, *hmwABC* and *hia* (not associated with capsulated isolates) were rarely observed simultaneously in any isolate, as previously observed for example for *hia* and *hmw* (Barenkamp and St Geme, 1996; St Geme III and Grass, 1998), suggesting that the diversification of its genetic repertoire allows NTHi to optimize its ability to adhere to host epithelium (Rao et al., 1999; Duell et al., 2016). It has been observed that piliated Hib presenting the haemagglutating pili genes *hif* have a higher capacity for colonization (Weber et al., 1991), and it was suggested that NTHi isolates would behave similarly (Rao et al., 1999; St Geme III, 2002), although the genes are generally found in only 15% of non-typeable isolates (Geluk et al., 1998; Krasan et al., 1999), as observed in our study. On the other hand, the HMW adhesin-coding genes, found in 60 to 75% of NTHi (Rao et al., 1999; St Geme III et al., 1998; Davis et al., 2014), were observed here in 30 out of the 57 PTNTHi isolates (53%). These are known to be targets of phase variation as mean to promote diversity to evade the host immune system, since they were identified as major targets of antibody response during *H. influenzae* infection (St Geme III, 2002). Moreover, HMW proteins may be able to interact with distinct eukaryotic receptors, functioning at different points during the process of colonization (Duell et al., 2016; St Geme 2002). We found that only NTHi strains from clades I and VI lack the complete set of these genes (Fig. 3), suggesting the existence of NTHi subpopulations with potential decreased adherence ability. Furthermore, a high degree of genetic diversification was observed for outer membrane proteins, probably resulting from the need to evade the host immune response (Garmendia et al., 2012; Osman et al., 2018), with *ompP6*, a recent vaccine candidate, displaying the lowest level of allelic diversity (Fig. 1).

The LOS structures have long been known to be important virulence factors (Garmendia et al., 2012; Rao et al., 1999; Duell et al., 2016; Wong and Akerley, 2012) being involved in *H. influenzae* adaptation and survival in the lung and bloodstream (Gawronski et al., 2009; Hood et al., 1996). The LOS is a heterogeneous surface structure on NTHi, with several genes involved in LOS biosynthesis frequently found to be targeted by phase variation (Garmendia et al., 2012; Rao et al., 1999) and, as we observed for most of the analyzed genes, presenting a high degree of allelic diversity, while also being differentially present/absent among isolates (Fig. 1). This contrasts, for example, with the smaller repertoire of LOS outer core structures and lack of phase variation observed for the closely related *Haemophilus parainfluenzae* species, which is believed to be associated with its reduced disease severity (Young and Hood, 2013). We observed a moderate level of genetic diversification in cell surface genes (*vacJ* and *yrbBCDEF*) associated with rearranging surface LOS. These rearrangements are believed to be associated with antibody binding to this structure, promoting serum resistance for the survival in the lung environment (Garmendia et al., 2012; Wong and Akerley, 2012; Nakamura et al., 2011). Considering this, we have no reasonable explanation for the genetic polymorphism found in these genes in PTNTHi. A high genetic diversity is also observed for genes involved in the uptake of iron and heme. Assuming that the selective pressure associated with both colonization and invasiveness is solely host dependent, it is reasonable to assume that the higher polymorphism observed for NTHi arises from the lack of capsule, making these strains more vulnerable to such host pressures.

NTHi possess a remarkable ability to acquire novel genomic material during infection (Maughan and Redfield, 2009) and it has been predicted that the *H. influenzae* pan-genome could reach up to ~6000 distinct genes (Hogg et al., 2007), with isolates acquiring genetic material during the co-colonization of the respiratory tract with other opportunistic bacteria, such as *S. pneumoniae* or *Moraxella catarrhalis* (Demuri et al., 2017; Xu et al., 2017). In the present study enrolling 88 *H. influenzae* isolates the pan-genome comprised 3424 genes, with a core-genome of 1308 genes (Fig. 2), although these values may be slightly underestimated due to the methodological constrains described above. Similar *H. influenzae* core-genomes sizes have been reported, where discreet differences are likely due to different datasets, sampling sizes and methodological approaches (e.g. 1207 (De Chiara et al., 2014) and 1437 (Hogg et al., 2007) genes, representing 63% and 52% of shared genome regions, respectively). While we observed several genes exclusively present ($n = 223$) within Hib genomes, including the capsule biosynthesis associated genes, the number of exclusive genes within PTNTHi was ~5-fold higher ($n = 1185$). Of note, within PTNTHi only 433 out of these 1185 genes (37%) were found to be shared by more than three isolates. Accordingly, although with a small subset, it has been observed that ~10% of the genes possessed by a clinical isolate are novel and that the overall distribution of these genes for NTHi is non-uniform (Shen et al., 2005), with several genes present in only one isolate (Hogg et al., 2007). Nevertheless, *H. influenzae* isolates maintain their relative genome size (Supplementary Table S1), suggesting that a balance between acquired and lost genes is energetically constrained. We observed that most of accessory genome of *H. influenzae* are putative phage-associated genes, i.e., 33% for Hib and 50% for PTNTHi, contrasting with the ~14% previously observed for NTHi (Hogg et al., 2007).

One key aspect for the increased colonization, serum resistance and chronicity of infection is the ability of *H. influenzae* to form biofilms within the host (Duell et al., 2016; Hogg et al., 2007; Langereis and Hermans, 2013; Webster et al., 2006). In fact, the genomic repertoire involved in adherence (e.g., outer membrane, pili-associated and LOS-associated proteins) has implication in the ability of *H. influenzae* to form these structures, either by the presence/absence of genes or their genetic diversity, as observed in the present study. On the other hand, quorum sensing coordinates activities for the formation of biofilms (Langereis and Hermans, 2013; Swords, 2012). As such, while all encapsulated strains (PTHi or published) lack them, we observed that that 16 PTNTHi isolates, as well as 32 published NTHi isolates, possessed the *lsrABCDEFGKR* genes (Li et al., 2007; Swords, 2012). Although further studies and experimental validation is required, this raises the question as to whether these isolates might have an increased ability to form biofilms, potentially promoting persistence in the host.

The availability to perform large-scale WGS studies has allowed researchers to gain insight into the population structures, genetic diversity and identification of key genomic features of pathogenic bacteria. While some bacteria have had their genomes sequenced by the thousands, for *H. influenzae* these large scale studies are still sparse, delaying our knowledge in understanding genomic diversity of this pathogen, which could ultimately lead to novel strategies in dealing with invasive disease, particularly for NTHi. On this regard, the application of microbial Genome Wide Associations Studies, with datasets containing both invasive and non-invasive strains, linked with complete metadata, might highlight the key genomic features associated with invasiveness, e.g. SNPs, indels or exclusive genes. We believe the present study contributes to the identification of novel genomic features present within the accessory genome of *H. influenzae*, which may be used to identify particular NTHi subpopulations (Fig. 3) in accordance with data reported in previous research based on core-phylogenetic analysis (De Chiara et al., 2014). In fact, although these isolates present high levels of genetic diversity, it was suggested that NTHi populations may be made up of defined lineages. Here, with the analysis of the pangenome from 243 NTHi, we highlight a set of 17 genes that may be used to classify NTHi isolates into specific clades and aid in establishing relationships with NTHi pathogenesis, particularly for the case of invasive NTHi on the onset of bacteremia or meningitis. Still, we believe that more genomic data is paramount in order to further validate these genes as reliable genetic markers, as data also suggests that there may be other less reported NTHi clades. Another contribution relies on providing two in silico prediction tools for the rapid identification of both *H. influenzae* serotypes and NTHi clades (after validating those 17 candidate genetic markers), which could be used to facilitate *H. influenzae* surveillance, obviating the demanding laboratory-based procedures. Ultimately, the data generated in the present study constitutes an important genomic database that could lay way for future studies on the genetic determinants and population structure of *H. influenzae*, underlying invasiveness and disease.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2018.10.025.

## Acknowledgements and funding

This work was supported by GenomePT (ref. POCI-01-0145-FEDER-022184) from Fundação para a Ciência e Tecnologia, Portugal.

## References

Agrawal, A., Murphy, T.F., 2011. *Haemophilus influenzae* infections in the *H. influenzae* type b conjugate vaccine era. J. Clin. Microbiol. 49, 3728–3732.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., Wishart, D.S., 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 44, W16–W21.

Baddal, B., Muzzi, A., Censini, S., Calogero, R.A., Torricelli, G., Guidotti, S., Taddei, A.R., Covacci, A., Pizza, M., Rappuoli, R., Soriani, M., Pezzicoli, A., 2015. Dual RNA-seq of nontypeable *Haemophilus influenzae* and host cell transcriptomes reveals novel insights into host-pathogen cross talk. MBio 6 (e01765–15).

Bajanca-Lavado, M.P., Simões, A.S., Betencourt, C.R., Sá-Leão, R., Portuguese Group for Study of Haemophilus influenzae invasive infection, 2014. Characteristics of *Haemophilus influenzae* invasive isolates from Portugal following routine childhood vaccination against *H. influenzae* serotype b (2002 − 2010). Eur. J. Clin. Microbiol. Infect. Dis. 33, 603–610.

Bakaletz, L.O., Baker, B.D., Jurcisek, J.A., Harrison, A., Novotny, L.A., Bookwalter, J.E., Mungur, R., Munson Jr., R.S., 2005. Demonstration of Type IV pilus expression and a twitching phenotype by *Haemophilus influenzae*. Infect. Immun. 73, 1635–1643.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Mol. Cell. Biol. 19, 455–477.

Barenkamp, S.J., St Geme, J.W.I.I.I., 1996. Identification of a second family of high-molecular-weight adhesion proteins expressed by non-typable *Haemophilus influenzae*. Mol. Microbiol. 19, 1215–1223.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Brynildsrud, O., Bohlin, J., Scheffer, L., Eldholm, V., 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. 17, 238.

Calado, R., Betencourt, C., Gonçalves, H., Cristino, N., Calhau, P., Lavado, P.B., 2011. Complicated meningitis caused by a rare serotype of *Haemophilus influenzae* in Portugal. Diagn. Microbiol. Infect. Dis. 69, 111–113.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2008. BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Campos, J., Román, F., Pérez-Vázquez, M., Oteo, J., Aracil, B., Cercenado, E., Spanish Study Group for Haemophilus influenzae Type E, 2003. Infections due to *Haemophilus influenzae* serotype E: microbiological, clinical, and epidemiological features. Clin. Infect. Dis. 37, 841–845.

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., Jin, Q., 2005. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 33, D325–D328.

Clark, S.E., Eichelberger, K.R., Weiser, J.N., 2013. Evasion of killing by human antibody and complement through multiple variations in the surface oligosaccharide of *Haemophilus influenzae*. Mol. Microbiol. 88, 603–618.

Connor, T.R., Corander, J., Hanage, W.P., 2012. Population subdivision and the detection of recombination in non-typable *Haemophilus influenzae*. Microbiology 158, 2958–2964.

Croucher, N.J., Page, A.J., Connor, T.R., Delaney, A.J., Keane, J.A., Bentley, S.D., Parkhill, J., Harris, S.R., 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 43, e15.

Dabernat, H., Delmas, C., Seguy, M., Pelissier, R., Faucon, G., Bennamani, S., Pasquier, C., 2002. Diversity of beta-lactam resistance-conferring amino acid substitutions in penicillin-binding protein 3 of Haemophilus influenzae. Antimicrob. Agents Chemother. 46, 2208–2218.

Davis, G.S., Patel, M., Hammond, J., Zhang, L., Dawid, S., Marrs, C.F., Gilsdorf, J.R., 2014. Prevalence, distribution, and sequence diversity of *hmwA* among commensal and otitis media non-typeable *Haemophilus influenzae*. Infect. Genet. Evol. 28, 223–232.

De Chiara, M., Hood, D., Muzzi, A., Pickard, D.J., Perkins, T., Pizza, M., Dougan, G., Rappuoli, R., Moxon, E.R., Soriani, M., Donati, C., 2014. Genome sequencing of disease and carriage isolates of nontypeable *Haemophilus influenzae* identifies discrete population structure. Proc. Natl. Acad. Sci. U. S. A. 111, 5439–5444.

Demuri, G.P., Gern, J.E., Eickhoff, J.C., Lynch, S.V., Wald, E.R., 2017. Dynamics of bacterial colonization with *Streptococcus pneumoniae*, *Haemophilus influenzae* and *Moraxella catarrhalis* during symptomatic and asymptomatic viral upper respiratory infection. Clin. Infect. Dis. 66, 1045–1053.

Desai, S., Jamieson, F.B., Patel, S.N., Seo, C.Y., Dang, V., Fediurek, J., Navaranjan, D., Deeks, S.L., 2015. The Epidemiology of Invasive *Haemophilus influenzae* Non-Serotype B Disease in Ontario, Canada from 2004 to 2013. PLoS ONE 10, e0142179.

Duell, B.L., Su, Y.C., Riesbeck, K., 2016. Host-pathogen interactions of nontypeable *Haemophilus influenzae*: from commensal to pathogen. FEBS Lett. 590, 3840–3853.

Erwin, A.L., Smith, A.L., 2007. Nontypeable *Haemophilus influenzae*: understanding virulence and commensal behavior. Trends Microbiol. 15, 355–362.

Erwin, A.L., Nelson, K.L., Mhlanga-Mutangadura, T., Bonthuis, P.J., Geelhood, J.L., Morlin, G., Unrath, W.C., Campos, J., Crook, D.W., Farley, M.M., Henderson, F.W., Jacobs, R.F., Mühlemann, K., Satola, S.W., van Alphen, L., Golomb, M., Smith, A.L., 2005. Characterization of genetic and phenotypic diversity of invasive nontypeable *Haemophilus influenzae*. Infect. Immun. 73, 5853–5863.

Eutsey, R.A., Hiller, N.L., Earl, J.P., Janto, B.A., Dahlgren, M.E., Ahmed, A., Powell, E., Schultz, M.P., Gilsdorf, J.R., Zhang, L., Smith, A., Murphy, T.F., Sethi, S., Shen, K., Post, J.C., Hu, F.Z., Ehrlich, G.D., 2013. Design and validation of a supragenome array for determination of the genomic content of *Haemophilus influenzae* isolates. BMC Genomics 14, 484.

Falla, T.J., Crook, D.W., Brophy, L.N., Maskell, D., Kroll, J.S., Moxon, E.R., 1994. PCR for capsular typing of *Haemophilus influenzae*. J. Clin. Microbiol. 32, 2382–2386.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., McKenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocyne, J.D., Scott, J., Shirley, R., Liu, L.I., Glodek, A., Kelley, J.M., Weidman, J.F., Phillips, C.A., Spriggs, T., Hedblom, E., Cotton, M.D., Utterback, T.R., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Fine, L.D., Fritchman, J.L., Fuhrmann, J.L., Geoghagen, N.S.M., Gnehm, C.L., McDonald, L.A., Small, K.V., Fraser, C.M., Smith, H.O., Venter, J.C., 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496–512.

Fox, K.L., Atack, J.M., Srikhanta, Y.N., Eckert, A., Novotny, L.A., Bakaletz, L.O., Jennings, M.P., 2014. 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. PLoSOne 9, e90505.

Garmendia, J., Martí-Lliteras, P., Moleres, J., Puig, C., Bengoechea, J.A., 2012. Genotypic and phenotypic diversity of the noncapsulated *Haemophilus influenzae*: adaptation and pathogenesis in the human airways. Int. Microbiol. (15), 159–172.

Garmendia, J., Viadas, C., Calatayud, L., Mell, J.C., Martí-Lliteras, P., Euba, B., Llobet, E., Gil, C., Bengoechea, J.A., Redfield, R.J., Liñares, J., 2014. 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 9, e97020.

Gawronski, J.D., Wong, S.M., Giannoukos, G., Ward, D.V., Akerley, B.J., 2009. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. Proc. Natl. Acad. Sci. U. S. A. 106, 16422–16427.

Geluk, F., Eijk, P.P., van Ham, M.S., Jansen, H.M., van Alphen, L., 1998. The fimbria gene cluster of nonencapsulated *Haemophilus influenzae*. Infect. Immun. 66, 406–417.

Gilsdorf, J.R., Marrs, C.F., Foxman, B., 2004. *Haemophilus influenzae*: genetic variability and natural selection to identify virulence factors. Infect. Immun. 72, 2457–2461.

Giufrè, M., Cardines, R., Caporali, M.G., Accogli, M., D'Ancona, F., Cerquetti, M., 2011. Ten years of Hib vaccination in Italy: prevalence of non-encapsulated *Haemophilus influenzae* among invasive isolates and the possible impact on antibiotic resistance. Vaccine 29, 3857–3862.

Hadfield, J., Croucher, N.J., Goater, R.J., Abudahab, K., Aanensen, D.M., Harris, S.R., 2017. Phandango: an interactive viewer for bacterial population genomics. Bioinformatics 34, 292–293.

Hogg, J.S., Hu, F.Z., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, J.C., Ehrlich, G.D., 2007. Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. Genome Biol. 8, R103.

Hood, D.W., Deadman, M.E., Allen, T., Masoud, H., Martin, A., Brisson, J.R., Fleischmann, R., Venter, J.C., Richards, J.C., Moxon, E.R., 1996. Use of the complete genome sequence information of *Haemophilus influenzae* strain Rd to investigate lipopolysaccharide biosynthesis. Mol. Microbiol. 22, 951–965.

Hu, F., Rishishwar, L., Sivadas, A., Mitchell, G.J., Jordan, I.K., Murphy, T.F., Gilsdorf, J.R., Mayer, L.W., Wang, X., 2016. Comparative genomic analysis of *Haemophilus haemolyticus* and nontypeable *Haemophilus influenzae* and a new testing scheme for their discrimination. J. Clin. Microbiol. 54, 3010–3017.

Jurcisek, J.A., Bookwalter, J.E., Baker, B.D., Fernandez, S., Novotny, L.A., Munson Jr., R.S., Bakaletz, L.O., 2007. The PilA protein of non-typeable *Haemophilus influenzae* plays a role in biofilm formation, adherence to epithelial cells and colonization of the mammalian upper respiratory tract. Mol. Microbiol. 65, 1288–1299.

Krasan, G.P., Cutter, D., Block, S.L., St Geme, J.W.I.I.I., 1999. Adhesin expression in matched nasopharyngeal and middle ear isolates of nontypeable *Haemophilus influenzae* from children with acute otitis media. Infect. Immun. 67, 449–454.

Kress-Bennett, J.M., Hiller, N.L., Eutsey, R.A., Powell, E., Longwell, M.J., Hillman, T., Blackwell, T., Byers, B., Mell, J.C., Post, J.C., Hu, F.Z., Ehrlich, G.D., Janto, B.A., 2016. Identification and characterization of msf, a novel virulence factor in *Haemophilus influenzae*. PLoS ONE 11, e0149891.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874.

Langereis, J.D., Hermans, P.W., 2013. Novel concepts in nontypeable *Haemophilus influenzae* biofilm formation. FEMS Microbiol. Lett. 346, 81–89.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987–2993.

Li, J., Attila, C., Wang, L., Wood, T.K., Valdes, J.J., Bentley, W.E., 2007. Quorum sensing in *Escherichia coli* is signaled by AI-2/LsrR: effects on small RNA and biofilm architecture. J. Bacteriol. 189, 6011–6020.

Lichtenegger, S., Bina, I., Durakovic, S., Glaser, P., Tutz, S., Schild, S., Reidl, J., 2017. Serum resistance and phase variation of a nasopharyngeal non-typeable *Haemophilus influenzae* isolate. Int. J. Med. Microbiol. 307, 139–146.

Maughan, H., Redfield, R.J., 2009. Tracing the evolution of competence in *Haemophilus influenzae*. PLoS ONE 4, e5854.

Meats, E., Feil, E.J., Stringer, S., Cody, A.J., Goldstein, R., Kroll, J.S., Popovic, T., Spratt, B.G., 2003. Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by multilocus sequence typing. J. Clin. Microbiol. 41, 1623–1636.

Mell, J.C., Shumilina, S., Hall, I.M., Redfield, R.J., 2011. Transformation of natural genetic variation into *Haemophilus influenzae* genomes. PLoS Pathog. 7, e1002151.

Musser, J.M., Granoff, D.M., Pattison, P.E., Selander, R.K., 1995. A population genetic framework for the study of invasive diseases caused by serotype b strains of *Haemophilus influenzae*. Proc. Natl. Acad. Sci. U. S. A. 82, 5078–5082.

Nakamura, S., Shchepetov, M., Dalia, A.B., Clark, S.E., Murphy, T.F., Sethi, S., Gilsdorf, J.R., Smith, A.L., Weiser, J.N., 2011. Molecular basis of increased serum resistance among pulmonary isolates of non-typeable *Haemophilus influenzae*. PLoS Pathog. 7, e1001247.

Osman, K.L., Jefferies, J.M., Woelk, C.H., Cleary, D.W., Clarke, S.C., 2018. The adhesins of non-typeable *Haemophilus influenzae*. Expert Rev. Anti-Infect. Ther. 16, 187–196.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T., Fookes, M., Falush, D., Keane, J.A., Parkhill, J., 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31, 3691–3693.

Peltola, H., 2000. Worldwide *Haemophilus influenzae* type b disease at the beginning of the 21st century: global analysis of the disease burden 25 years after the use of the polysaccharide vaccine and a decade after the advent of conjugates. Clin. Microbiol. Rev. 13, 302–317.

Pittman, M., 1931. Variation and type specificity in the bacterial species *Haemophilus influenzae*. J. Exp. Med. 53, 471–492.

Poje, G., Redfield, R.J., 2003. Transformation of *Haemophilus influenzae*. Methods Mol. Med. 71, 57–70.

Power, P.M., Sweetman, W.A., Gallacher, N.J., Woodhall, M.R., Kumar, G.A., Moxon, E.R., Hood, D.W., 2009. Simple sequence repeats in *Haemophilus influenzae*. Infect. Genet. Evol. 9, 216–228.

Power, P.M., Bentley, S.D., Parkhill, J., Moxon, E.R., Hood, D.W., 2012. Investigations into genome diversity of *Haemophilus influenzae* using whole genome sequencing of clinical isolates and laboratory transformants. BMC Microbiol. 12, 273.

Price, E.P., Sarovich, D.S., Nosworthy, E., Beissbarth, J., Marsh, R.L., Pickering, J., Kirkham, L.A., Keil, A.D., Chang, A.B., Smith-Vaughan, H.C., 2014. *Haemophilus influenzae*: using comparative genomics to accurately identify a highly recombinogenic human pathogen. BMC Genomics 16, 641.

Rao, V.K., Krasan, G.P., Hendrixson, D.R., Dawid, S., St Geme 3rd, J.W., 1999. Molecular determinants of the pathogenesis of disease due to non-typable *Haemophilus influenzae*. FEMS Microbiol. Rev. 23, 99–129.

Resman, F., Ristovski, M., Ahl, J., Forsgren, A., Gilsdorf, J.R., Jasir, A., Kaijser, B., Kronvall, G., Riesbeck, K., 2011. Invasive disease caused by *Haemophilus influenzae* in Sweden 1997-2009; evidence of increasing incidence and clinical burden of non-type b strains. Clin. Microbiol. Infect. 17, 1638–1645.

Rubach, M.P., Bender, J.M., Mottice, S., Hanson, K., Weng, H.Y., Korgenski, K., Daly, J.A., Pavia, A.T., 2011. Increasing incidence of invasive *Haemophilus influenzae* disease in adults, Utah, USA. Emerg. Infect. Dis. 17, 1645–1650.

Sadeghi-Aval, P., Tsang, R.S., Jamieson, F.B., Ulanova, M., 2013. Emergence of non-serotype b encapsulated *Haemophilus influenzae* as a cause of pediatric meningitis in northwestern Ontario. Can. J. Infect. Dis. Med. Microbiol. 24, 13–16.

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069.

Shen, K., Antalis, P., Gladitz, J., Sayeed, S., Ahmed, A., Yu, S., Hayes, J., Johnson, S., Dice, B., Dopico, R., Keefe, R., Janto, B., Chong, W., Goodwin, J., Wadowsky, R.M., Erdos, G., Post, J.C., Ehrlich, G.D., Hu, F.Z., 2005. Identification, distribution, and expression of novel genes in 10 clinical isolates of nontypeable *Haemophilus influenzae*. Infect. Immun. 73, 3479–3491.

Silva, M., Machado, M.P., Silva, D.N., Rossi, M., Moran-Gilad, J., Santos, S., Ramirez, M., Carriço, J.A., 2018. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. Microb. Genom. 4, e000166.

Slack, M.P.E., 2017. The evidence for non-typeable *Haemophilus influenzae* as a causative agent of childhood pneumonia. Pneumonia 9, 9.

St Geme III, J.W., 2002. Molecular and cellular determinants of non-typeable *Haemophilus influenzae* adherence and invasion. Cell. Microbiol. 4, 191–200.

St Geme III, J.W., Grass, S., 1998. Secretion of the *Haemophilus influenzae* HMW1 and HMW2 adhesins involves a periplasmic intermediate and requires the HMWB and HMWC proteins. Mol. Microbiol. 27, 617–630.

St Geme III, J.W., Cutter, D., Barenkamp, S.J., 1996. Characterization of the genetic locus encoding *Haemophilus infuenzae* type b surface fibrils. J. Bacteriol. 178, 6281–6287.

St Geme III, J.W., Kumar, V.V., Cutter, D., Barenkamp, S.J., 1998. Prevalence and distribution of the hmw and hia genes and the HMW and Hia adhesins among genetically diverse strains of nontypeable *Haemophilus influenzae*. Infect. Immun. 66, 364–368.

Staples, M., Graham, R.M.A., Jennison, A.V., 2017. Characterization of invasive clinical *Haemophilus influenzae* isolates in Queensland, Australia using whole-genome sequencing. Epidemiol. Infect. 145, 1727–1736.

Swords, W.E., 2012. Quorum signaling and sensing by nontypeable *Haemophilus influenzae*. Front. Cell. Infect. Microbiol. 2, 100.

Takahata, S., Ida, T., Senju, N., Sanbongi, Y., Miyata, A., Maebashi, K., Hoshiko, S., 2007. Horizontal gene transfer of *ftsI*, encoding penicillin-binding protein 3, in *Haemophilus influenzae*. Antimicrob. Agents Chemother. 51, 1589–1595.

Treangen, T.J., Ondov, B.D., Koren, S., Phillippy, A.M., 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. 15, 524.

Tsang, R.S.W., Ulanova, M., 2017. The changing epidemiology of invasive *Haemophilus influenzae* disease: Emergence and global presence of serotype a strains that may require a new vaccine for control. Vaccine 35, 4270–4275.

Ulanova, M., Tsang, R.S.W., 2014. *Haemophilus influenzae* serotype a as a cause of serious invasive infections. Lancet Infect. Dis. 14, 70–82.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., Earl, A.M., 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9, e112963.

Wan Sai Cheong, J., Smith, H., Heney, C., Robson, J., Schlebusch, S., Fu, J., Nourse, C., 2015. Trends in the epidemiology of invasive *Haemophilus influenzae* disease in Queensland, Australia from 2000 to 2013: what is the impact of an increase in invasive non-typable *H. influenzae* (NTHi)? Epidemiol. Infect. 143, 2993–3000.

Weber, A., Harris, K., Lohrke, S., Forney, L., Smith, A.L., 1991. Inability to express fimbriae results in impaired ability of *Haemophilus influenzae* b to colonize the nasopharynx. Infect. Immun. 59, 4724–4728.

Webster, P., Wu, S., Gomez, G., Apicella, M., Plaut, A.G., St Geme III, J.W., 2006. Distribution of bacterial proteins in biofilms formed by non-typeable *Haemophilus influenzae*. J. Histochem. Cytochem. 54, 829–842.

Wong, S.M., Akerley, B.J., 2012. Genome-scale approaches to identify genes essential for *Haemophilus influenzae* pathogenesis. Front. Cell. Infect. Microbiol. 2, 23.

Xu, Q., Wischmeyer, J., Gonzalez, E., Pichichero, M.E., 2017. Nasopharyngeal polymicrobial colonization during health, viral upper respiratory infection and upper respiratory bacterial infection. J. Inf. Secur. 75, 26–34.

Young, R.E., Hood, D.W., 2013. *Haemophilus parainfluenzae* has a limited core lipopolysaccharide repertoire with no phase variation. Glycoconj. J. 30, 561–576.