Investigating the Population Structure of the Human Bacterial Pathogen, *Haemophilus influenzae*



Made Ananda Krisna

University of Oxford

Nuffield Department of Clinical Medicine

Hertford College

A thesis submitted for the degree of Doctor of Philosophy

Hillary Term 2024/2025

Abstract

Haemophilus influenzae is a human-adapted pathogen that causes both respiratory and invasive diseases. Following the introduction of the *H. influenzae* type b (Hib) conjugate vaccine, non-typeable *H. influenzae* (NTHi) has emerged as the predominant cause of invasive disease, with reported cases continuing to rise. However, understanding of the species population structure remains limited due to its genetic diversity and the lack of high-resolution analytical tools and representative isolates from low- and middle-income countries (LMICs) such as Indonesia. This DPhil thesis addresses these gaps by developing a genomic framework for high-resolution typing, identifying genetic factors linked to invasiveness in NTHi, and characterising circulating strains in an underrepresented region.

A core genome multilocus sequence typing (cgMLST) scheme was developed using 2,297 high-quality genomes, resulting in a stable set of 1,037 core genes. These genes were functionally annotated and used to construct a cgMLST framework that accurately reflects phylogenetic relationships. The scheme was implemented in the PubMLST database to support accessible and standardised population analysis.

Building on this framework, a genomic clustering system based on the Life Identification Number (LIN) code was applied to define consistent, hierarchical groupings within the species. Demonstrated using published data, the cgLIN scheme enables scalable classification of *H. influenzae* lineages and supports public health applications including antimicrobial resistance (AMR) surveillance and outbreak detection.

A genome-wide association study (GWAS) was conducted to investigate the genetic basis of invasiveness in NTHi. By comparing invasive and non-invasive isolates from global

sources, the analysis identified variants in porin genes, TonB-dependent receptors, and regulatory elements, many of which correspond to known virulence factors. However, no single genetic determinant fully explained the phenotypic outcome of invasive disease in NTHi, suggesting that this trait is polygenic and shaped by extensive recombination, as previously observed in vitro and further supported by large-scale in silico genomic analyses in this study

The final part of the thesis assessed the population structure of *H. influenzae* in Indonesia. A total of 113 isolates from both carriage and invasive disease were characterised using cgMLST and cgLIN. The Indonesian isolates were genetically diverse and largely composed of NTHi lineages similar to those found globally. Ampicillin resistance was common, mediated by mobile genetic elements and mutations in the *ftsI* gene. These findings demonstrate the importance of integrating genomic data with local epidemiological context to inform regional AMR surveillance.

Overall, this thesis provides a publicly available framework for the genomic characterisation of *H. influenzae* and applies it to key questions in population structure, virulence, and resistance. The results support improved surveillance, inform vaccine development, and enhance public health response efforts, particularly in regions with limited genomic infrastructure.

Acknowledgement

Firstly, I am deeply grateful to my primary supervisors at Oxford, Martin Maiden and Odile Harrison. From the very outset of my DPhil, they placed great confidence in me. Martin's scientific guidance and vast knowledge of bacterial genomics have been truly inspiring, while Odile's day-to-day support with methodology and technical issues has been invaluable. Both have afforded me the independence to explore new ideas and tackle research problems creatively.

My thanks also go to my secondary supervisor Raph Hamers and the OUCRU Indonesia team, particularly at the start of my DPhil, for their administrative support. Every discussion with Raph has been both fruitful and thought-provoking, shaping the way I view a scientific career and its many nuances.

I extend my gratitude to two Maiden Lab colleagues: Keith Jolley, whose work on the PubMLST software underpins much of my data and analytical tools, and James Bray, rMLST curator and, more importantly, a good friend who has listened patiently to both DPhil-related and personal concerns. I must also thank Anastasia Unitt, a go-to person for countless random questions since the start of my journey. Thanks are also due to all past and present members of the Maiden and Sheppard Labs for their camaraderie and support.

I would to thank both of my parents: my late mom, who had been supportive of my ambitions in science throughout my school years and my dad who has always supported my career decisions, no questions asked, even when they meant leaving medical residency.

Finally, and most importantly, I am forever grateful for my husband, Billy. Thank you for all the love and all you do that make my life so much better. Thank you for every discussion about science, for taking on the housework because I was too tired, and for being my number one supporter in everything I do.

Statement of Contributions

My supervisors, Martin C. J. Maiden, Odile B. Harrison and Raph L. Hamers, provided conceptual frameworks, scientific guidance and manuscript feedback at every stage of these DPhil projects. The work presented here is my own, except for the following contributions by additional collaborators:

In Chapter 2, William Monteith assisted in developing and streamlining the Python code.

In Chapter 5, Dodi Safari supplied all archived isolates for WGS. Korrie Salsabila, Wisiva T. Paramaiswari, and Wisnu Tafroji conducted microbiological identification from samples; and helped with RT-PCR confirmation, serotyping and sequencing.

Table of Contents

Abstract ..... **i**
Acknowledgement ..... **iii**
Statement of Contributions ..... **iv**
Table of Contents ..... **v**
List of Tables ..... **x**
List of Equation ..... **xi**
Table of Figures ..... **xii**
List of abbreviations ..... **xiii**
Publication in relation to this thesis ..... **xvi**
Chapter 1 ..... **1**
***Haemophilus influenzae*: epidemiology, molecular typing, and genomic insights** **1**
1.1. Pathogenesis of *H. influenzae* 2
1.1.1. Capsule operon encoding *H. influenzae* primary virulence factor, the capsule polysaccharide 2
1.1.2. Other virulence factors contributing to the pathogenesis of *H. influenzae* disease 3
1.2. The changing epidemiology of *H. influenzae* in carriage and disease 5
1.2.1. The impact of Hib vaccination: Successes and challenges in reducing Hib-related disease 5
1.2.2. Emergence of invasive NTHi and non-type b capsulated *H. influenzae* 8
1.3. Typing strategies and the molecular epidemiology of *H. influenzae* 10
1.3.1. Genomic typing methods: MLST, cgMLST, and other genomic classification tools and schemes. 10
1.3.2. Molecular epidemiology of circulating *H. influenzae*: Dominant clonal complexes in capsulated strains and their absence in NTHi. 15
1.4. The rise of resistant *H. influenzae* strains 18
1.4.1. Antimicrobial resistance (AMR) epidemiology in *H. influenzae* 18
1.4.2. Genetic determinants of AMR in *H. influenzae* 20
1.5. Genomic insights: The *H. influenzae* pangenome and the role of recombination in shaping *H. influenzae*'s genetic diversity 5
1.6. Thesis aims 8
1.6.1. Aims and objectives 9
1.6.2. Overview of chapters 10
Chapter 2 ..... **13**
Development and Implementation of a Core Genome Multilocus Sequence Typing (cgMLST) scheme for *Haemophilus influenzae* **13**
Abstract 13
2.1. Introduction 14
2.2. Method 16
2.2.1. Choosing pangenome analysis tools and dataset compilation 16
2.2.2. Core gene identification and curation 20
2.2.3. Validation analyses 22

2.2.4. Phylogenetic analysis 23
2.2.5. Relationships between the cgMLST scheme pairwise allelic mismatch and ML tree branch length 24
2.3. Results 25
2.3.1. *H. influenzae* genomes from PubMLST database used for cgMLST scheme development and validation. 25
2.3.2. One thousand and thirty seven core genes in the validated cgMLST scheme are implicated in important cellular pathways. 27
2.3.3. Clustering groups of *H. influenzae* genomes based on pairwise allelic mismatches of the core genes reflected their phylogenetic relationship. 30
2.4. Discussion 36
2.5. Conclusion 39
Chapter 3 40
Resolving *Haemophilus influenzae* Molecular Epidemiology using Life Identification Number (LIN) Coding 40
Abstract 40
3.1. Introduction 40
3.2. Method 44
3.2.1. Choosing datasets 44
3.2.2. Identification of LIN code partition levels 44
3.2.3. Threshold validation using statistical testing 46
3.2.4. LIN code implementation on PubMLST *H. influenzae* database and the LIN code prefix interpretation 48
3.2.5. Average nucleotide identity 50
3.3. Results 50
3.3.1. Thirteen LIN code partitioning levels were identified to define clusters of *H. influenzae* genomes with 12.25% to 99.9% similarity level 50
3.3.2. Decoding LIN code prefixes: A guide to understanding and converting machine-readable barcode into user-friendly nomenclature 52
3.3.3. Ten major lineages of global collection of *H. influenzae* genomes were identified 55
3.3.4. LIN code clustering bins at clade, subclade, and strain level are useful for epidemiological tracking and investigation. 57
3.4. Discussion 60
3.5. Conclusion 65
Chapter 4 67
Variants in surface-exposed proteins are associated with invasive non-typeable Haemophilus influenzae infection: A genome-wide association study 67
Abstract 67
4.1. Introduction 68
4.2. Methods 71
4.2.2. Capsule type genotyping, genome annotation, and lineage definition. 73

4.2.3. Phylogenetic, recombination, and pangenome analyses 74
4.2.4. Bacterial GWAS pipeline 75
4.2.5. Statistical testing and phenotype prediction using logistic regression 78
4.2.6. Mapping kmers with predictive value 80
4.3. Results 80
4.3.1. Each GWAS dataset consisted of more than 1000 high quality, globally distributed non-typeable *H. influenzae* draft genomes 80
4.3.2. *H. influenzae* genomes from invasive and non-invasive group had a similar pangenome size. 83
4.3.3. Recombination events were detected in most parts of NTHi genomes across different lineages, with comparable effects of recombination between two phenotypic groups. 85
4.3.4. Genetic variants associated with invasive phenotype group in the initial GWAS were mostly detected in the accessory genome of NTHi 89
4.3.5 Validation GWAS confirmed 16 genetic regions associated with invasive NTHi. 92
4.4. Discussion 99
4.5. Conclusion and future directions 108
Chapter 5 109
The population genetics of *Haemophilus influenzae* isolates from Indonesia in carriage and disease 109
Abstract 109
5.1. Introduction 110
5.2. Method 113
5.2.1. Design, setting, and primary dataset 113
5.2.2. *H. influenzae* isolation, identification, and serotyping 114
5.2.3. Antibiotic susceptibility testing (AST) 116
5.2.4. Whole-genome sequencing and post-assembly analyses 117
5.2.5. Global collection of *H. influenzae* genomes 121
5.3. Result 121
5.3.1. High positive rate of *H. influenzae* among carriage and invasive cases were detected at the extremes of age. 121
5.3.2. *H. influenzae* isolates from Indonesia were highly diverse, belonging to 21 different megalineages based on cgLIN code. 123
5.3.3. All phenotypically resistant *H. influenzae* isolates were explained by previously described AMR determinants 127
5.3.4. Known virulence factors-encoding genes were distributed across *H. influenzae* isolates from Indonesia. 132
5.3.5. Everything is everywhere*: H. influenzae* from Indonesia were spread across the phylogeny of global collection of *H. influenzae*. 134
5.4. Discussion 137
5.5. Conclusion and future directions 144
Chapter 6 147

General Discussion and Conclusion 147
6.1. Key findings 147
6.1.1. Establishing core genome LIN code nomenclature as the standard in *H. influenzae* population genetics. 148
6.1.2. Complex genetic architecture of invasive NTHi: Variants in core virulence genes and noncoding regions. 150
6.1.3. High prevalence of NTHi carriage and beta-lactam resistance in Indonesia: Genomic insights via cgLIN scheme 152
6.2. Limitations 153
6.3. Implications for public health and research and future directions 154
6.4. Conclusion 155
References 156
Appendices 180
Appendix 2 180
Appendix 2.2. Scatter plot of the number of alleles and length variation for each core gene in the cgMLST scheme, coloured by PHI permutation statistics. 180
Appendix 2.3.1. The maximum-likelihood tree from core genome alignment of 1,376 *H. influenzae* genomes in the validation dataset. 181
Appendix 2.3.2. A minimum-spanning tree based on core genome profile, showing different CGC500, CGC200, and CGC50 groups clustered together. 182
Appendix 2.4. Variation of genetic relatedness among NTHi isolates within the same pathotype clade. 184
Appendix 2.5. The maximum-likelihood tree from core genome alignment of 1,376 *H. influenzae* genomes in the validation dataset 185
Appendix 3 186
Appendix 3.1. Distribution curves of core genome allelic mismatches in H. influenzae subpopulation based on serotype and clonal complex 186
Appendix 3.2. Comparison between clustering results using LIN method at different threshold and pre-existing methods in the literature. 188
Appendix 3.3. Variation of average nucleotide identity (ANI) values within clusters formed at different LINcode thresholds. 189
Appendix 3.4. Heat map of Normalised Mutual Information (NMI) calculated for pairwise clustering result comparison for each possible threshold. 190
Appendix 3.5. A neighbor-joining tree of H. influenzae isolates of clinical and public health importance, as described by Collins et al (2019), was annotated with LIN code prefixes and corresponding nicknames. 191
Appendix 4 192
Appendix 4.1. Q-Q plot of randomly chosen GWAS experiments. 192
Appendix 4.2. Recombination parameters distribution between invasive and non invasive group among NTHi population as histogram (A) and violin plot (B) 193
Appendix 4.3. The phylogenetic tree of NTHi genomes in (A) dataset 1 and (B) dataset 2, annotated by the logistic regression model result. 194

Appendix 4.4. Variants associated with invasive NTHi infection in non-protein coding and intergenic regions as aligned to *H. influenzae* reference genome (GCF_000931575.1) 195

Appendix 4.5. Multiple sequence alignments of (a) tRNA-Methionine (tRNA-Met) and (b) tRNA-Glutamate (tRNA-Glu) genes present in the NTHi reference genome (GCF_000931575.1). 196

Appendix 5 197

Appendix 5.1. Probe and primers for *hpd* gene detection and serotyping of *Haemophilus influenzae.* 197

Appendix 5.2. Genome length and ribosomal MLST (rMLST) designation for four *H. influenzae* isolates from Indonesia excluded from downstream analyses. 199

Appendix 5.3. *ftsI* alleles assigned to non-sensitive *H. influenzae* isolates from Indonesia and the point mutations associated with with each allele. 200

Appendix 5.4. List of point mutations in the *gyrA*, *parC*, and *parE* in 2 *H. influenzae* isolates which are phenotypically resistant to ciprofloxacin and moxifloxacin. 201

List of Tables

Table 1.1. Infections caused by *H. influenzae* 1
Table 1.2. The seven locus MLST scheme of *H. influenzae*. 13
T**able 1.4.** Genetic determinants of AMR in *H. influenzae* 2
Table 2.1. Methods and features of four different pangenome and core genome reconstruction tools 18
Table 2.2. Parameters for assessing quality of *H. influenzae* draft genome assemblies. 19
Table 4.1. Summary of genome assembly quality parameters of the two NTHi datasets. 83
Table 4.2 Protein-coding genes with the most significant kmers associated with NTHi invasive phenotype 91
Table 4.3 Valid kmers in the final logistic regression model and their location in the NTHi genome. 96
Table 5.1. *H. influenzae* surveillance projects from which isolates in this study originated 113
Table 5.2. Molecular characteristics of *H. influenzae* isolates non-sensitive to beta-lactam antibiotics. 129

List of Equation

Equation 4.1. Logistic regression equation 79

Table of Figures

Figure 2.1. Allocation of high-quality H. influenzae draft genome assemblies into development and validation dataset.. 20
Figure 2.2. The workflow of cgMLST scheme development and validation. 21
Figure 2.3. Characteristics of the datasets and genomes employed for developing the *Haemophilus influenzae* cgMLST scheme. 27
Figure 2.4. Functional classification, recombination, and allele variability analysis of *Haemophilus influenzae* core genes in the cgMLST scheme. 28
Figure 2.5. Population structure of 1,376 *H. influenzae* genomes from the validation dataset 31
Figure 2.6. MST of the validation dataset (N = 1376 genomes) constructed from the core genome allelic profile. 33
Figure 2.7. Comparison of pairwise allelic mismatch of cgMLST core genes with the log10 branch length values from the ML tree, implemented in the validation dataset 35
Figure 3.1. Distribution of pairwise allelic mismatches (PAM) for 2,297 high-quality H. influenzae draft genomes and statistical evaluation for clustering at multiple allelic mismatch thresholds. 51
Figure 3.2. Illustration of LIN code grouping and nickname nomenclature implementation using partitioning thresholds defined for *H. influenzae*. 54
Figure 3.3. A Maximum-Likelihood (ML) phylogeny of 2,297 *H. influenzae* genomes annotated by LIN code clustering at lineage and clonal group level. 56
Figure 3.4. Implementation of LIN code clustering at clade, subclade, and strain level to the ML phylogeny of 123 *H. influenzae* genomes. 58
Figure 4.1 The bacterial GWAS pipeline for identifying genetic variants associated with phenotypic outcome of invasive disease in the NTHi population 77
Figure 4.2. Frequency of phenotype and STs, geographical distribution, and year of isolation of two NTHi datasets for GWAS experiment and model evaluation. 82
Figure 4.3. The pangenome size of NTHi in dataset 1 based on their phenotypic group **Error! Bookmark not defined.**
Figure 4.4. (A) The Maximum-Likelihood phylogeny of NTHi in dataset 1 using the core genome alignment. 87
Figure 4.5. Annotation of significant kmers associated with NTHi invasiveness to a reference and high-quality draft genomes. 90
Figure 4.6. Logistic regression 93
Figure 4.7. Regions in the coded proteins affected by nucleotide sequence variants associated with NTHi invasive phenotype identified through the GWAS pipeline.. 98
Figure 5.1 Description and results at each stage of the study workflow from sample selection, *H. influenzae* identification, serotyping, and whole-genome sequencing.... 122
Figure 5.2. The maximum-likelihood phylogeny of *H. influenzae* isolates from Indonesia based on the core genome alignment using the *H. influenzae* cgMLST scheme 125
Figure 5.3. ML phylogeny of *H. influenzae* isolates from invasive cases. 127
Figure 5.4. Virulence factors (VFs)-related genes. 133
Figure 5.5. Minimum-spanning tree of 2,869 *H. influenzae* isolates from the global collection and Indonesia isolates 136

List of abbreviations

AI/AN	American Indian and Alaska Native
AIC	Akaike information criterion
AMC	Amoxicillin-clavulanic acid
AMP	Ampicillin
AMR	Antimicrobial resistance
ANI	Average nucleotide identity
AUC	Area under the curve
BAL	Broncho-alveolar lavage
BI	Bayesian inference
BLAST	Basic Local Alignment Search Tool
BLNAR	Beta-lactamase negative ampicillin resistant
BLPAR	Beta-lactamase positive ampicillin resistant
BSR	BLAST score ratio
CARD	Comprehensive antibiotic resistance database
CC	Clonal complex
CDC	Centers for Disease Control and Prevention
CDS	Coding sequences
CF	Cystic fibrosis
CG	Clonal group
CGC	Core genome cluster
cgLIN	Core genome-based life identification number
cgMLST	Core genome multilocus sequence typing
CHL	Chloramphenicol
CI	Confidence interval
CIP	Ciprofloxacin
CLSI	Clinical and Laboratory Standards Institute
CNS	Central nervous system
COG	Cluster of Orthologous Genes
COPD	Chronic obstructive pulmonary disease
CSF	Cerebrospinal fluid
CTA	Cefotaxime
CTR	Ceftriaxone
DNA	Deoxyribonucleic acid
ECDC	European Centre for Disease Prevention and Control
ECM	Extracellular matrix

ESCMID	European Society of Clinical Microbiology and Infectious Diseases
GC	Genome comparator
GLASS	Global Antimicrobial Resistance Surveillance System
GWAS	Genome wide association study
HGT	Horizontal gene transfer
Hia	*Haemophilus influenzae* type a
Hib	*Haemophilus influenzae* type b
Hic	*Haemophilus influenzae* type c
Hid	*Haemophilus influenzae* type d
Hie	*Haemophilus influenzae* type e
Hif	*Haemophilus influenzae* type f
HTM	*Haemophilus* Test Medium
ICE	Integrative conjugative element
KEGG	Kyoto Encyclopedia of Genes and Genomes
LIN	Life identification number
LMIC	Lower- and middle income countries
LOS	lipopolysaccharides
MCL	Markov cluster
MDR	Multi-drug resistance
MIC	Minimum inhibitory concentration
ML	Maximum likelihood
MLST	Multilocus sequence typing
MOX	Moxifloxacin
MP	Maximum parsimony
MST	Minimum-spanning tree
NCBI	National Center for Biotechnology Information
NS	Non susceptible
NTHi	Non-typeable *Haemophilus influenzae*
ODC	Ornithine decarboxylase
OG	Orthologous group
OLS	Ordinary least squares
ORF	Open reading frame
PAM	Pairwise allelic mismatch
PBP	Penicillin-binding protein
PC	Paralog count
PCR	Polymerase chain reaction
PCV13	Pneumococcus conjugate vaccine 13 valent

PFGE	Pulsed-field gel electrophoresis
PHI	Pairwise homoplasy index
QA/QC	Quality assurance and quality control
QRDR	Quinolone resistance-determining region
RNA	Ribonucleic acid
ROC	Receiver operating curve
RT-PCR	Real-time PCR
SEARO	WHO Southeast Asia Region
SNP	Single nucleotide polymorphism
SRA	Sequence read archive
ST	Sequence type
STGG	Skim milk-tryptone-glucose-glycerol
TBDR	TonB dependent receptor
VF	Virulence factor
VNTR	Variable-number tandem repeat
WGS	Whole genome sequencing
WHO	World Health Organization
WT	Wild type

Publication in relation to this thesis

Chapter 2

Krisna MA, Jolley KA, Monteith W, Boubour A, Hamers RL, Brueggemann AB, Harrison OB, Maiden MCJ: Development and implementation of a core genome multilocus sequence typing scheme for Haemophilus influenzae. *Microb Genom* 2024, 10.

Chapter 5

Krisna, M.A., Alimsardjono, L., Salsabila, K. et al. Whole-genome sequencing of non-typeable Haemophilus influenzae isolated from a tertiary care hospital in Surabaya, Indonesia. *BMC Infect Dis* 24, 1097 (2024).

Putri ND, Salsabila K, Prayitno A, Aprianti SC, Paramaiswari WT, **Krisna MA**, Safari D: Epidemiology of Haemophilus influenzae in children on Lombok Island, Indonesia. *Access Microbiol* 2023, 5.

Chapter 1

Haemophilus influenzae: epidemiology, molecular typing, and genomic insights

Haemophilus influenzae is a commensal bacterium which resides in the upper respiratory tract and has the potential to become an opportunistic pathogen in humans, its only known natural host. Clinical infection of this Gram-negative coccobacillus can manifest as either invasive or non-invasive infection (Table 1.1). Its major virulence factor (VF) of the bacterium is the polysaccharide capsule. Based on the presence or absence of this capsule, *H. influenzae* is classified into encapsulated/typeable and unencapsulated/non-typeable (NTHi) [1, 2]. The encapsulated strains are further divided into six serotypes (a–f) based on distinct capsular polysaccharide structures. Historically, serotype b (Hib) accounted for most invasive *H. influenzae* disease, including meningitis and septicaemia, whereas NTHi predominates in carriage and non-invasive infections [3, 4]. However, since introduction of the Hib vaccine, NTHi has emerged as the leading cause of invasive *H. influenzae* disease [5, 6]. These shifts in serotype prevalence and disease presentation therefore underscore the need to synthesise current knowledge and identify remaining gaps.

Table 1.1. Infections caused by *H. influenzae* [1, 7, 8].

Anatomical site	Disease
Upper respiratory tract	Sinusitis, laryngitis, tonsilitis
Eye and ear	Conjunctivitis, acute otitis media, mastoiditis
Lower respiratory tract	Bronchitis, pneumoniae, acute exacerbations of COPD, chronic infection in CF
Central nervous system	Meningitis
Other internal organs	Epiglottitis, pleural effusion, endocarditis, arthritis, osteomyelitis.
Systemic infection	Bacteraemia, septicaemia, purpura fulminans (including the Brazilian purpuric fever)

COPD: chronic obstructive pulmonary disease, CF: cystic fibrosis

Accordingly, this chapter reviews the literature on *H. influenzae* with a focus on three core themes relevant to this DPhil project: (1) mechanisms of pathogenesis and associated VFs (addressed in a single subchapter), (2) the public-health significance of this pathogen, which spans three subchapters because of its breadth, and (3) the genomic features that shape its biology and explain epidemiological findings (also addressed in a single subchapter). For each theme, critical challenges and gaps in understanding are highlighted.

1.1. Pathogenesis of *H. influenzae*

1.1.1. Capsule operon encoding *H. influenzae* primary virulence factor, the capsule polysaccharide

The *H. influenzae* capsule is encoded by the *cap* locus, which contains functionally unique regions: regions I, II, and III. Genes in region I (*bexABCDE*) encode proteins involved in the capsule export machinery which require ATP and genes in region III (*hcsAB*) encode proteins responsible for capsule post-polymerisation step. Region II genes of the *cap* locus *are* unique for different serotypes [3, 4]. Although their open reading frames (ORF) *are* defined and the sequence variations extensively studied, only a few genes in the Region II *have* known products and have been fully characterised [9].

Potts *et al* evaluated the genetic diversity of each gene in each region of the *H. influenzae cap* locus. Genes from capsule region I or III (*bexA, bexB, bexC, bexD, hcsA, hcsB*), showed nucleotide percentage identities ranging from approximately 81% to near 100%. Notably, *hcsA* and *hcsB* exhibited greater allelic diversity and slightly lower overall identity (~81% to 96%), indicating higher variability compared to *bex* genes. In contrast, region II genes (*acs, bcs, ccs,*

dcs, ecs, and *fcs*), demonstrated relatively high sequence identity among alleles belonging to specific serotypes (>97%). Genes such as *acs1, acs2, acs3,* and *ecs1* shared nucleotide identities ranging roughly from 97.5% to nearly 100%, with *acs* and *bcs* genes showing the greatest variability within a serotype. Meanwhile, *ccs, dcs,* and *fcs* genes had a nucleotide similarity between 99%-100%, suggesting these genes are more conserved across isolates [3].

Serotype determination for *H. influenzae* was primarily performed using classical serological methods, such as slide agglutination with type-specific antisera. Although rapid and cost-effective, serology had limitations, including subjective interpretation, cross-reactivity, and the inability to serotype strains that express low levels of capsule. Consequently, molecular techniques have gained preference for precise identification and classification. Molecular methods, particularly polymerase chain reaction (PCR)-based assays targeting genes within the *cap* locus (e.g. *bcs2* to detect Hib), offer increased sensitivity, specificity, and reproducibility compared to serological approaches. PCR-based techniques have thus become a standard practice, especially for epidemiological studies and surveillance programs aimed at monitoring the distribution and prevalence of specific *H. influenzae* serotypes following vaccine implementation [10].

1.1.2. Other virulence factors contributing to the pathogenesis of *H. influenzae* disease

Since *H. influenzae* exclusively inhabits the human nasopharynx, it must possess the necessary genetic determinants to survive and persist in this niche, ensuring sustained colonisation. As an opportunistic pathogen, it is also equipped with various factors that can promote tissue damage, trigger inflammatory responses, and, in some cases, facilitate invasion through the mucosal barrier into the bloodstream. Collectively, components that enable *H.*

influenzae to establish infection and cause disease are referred to as virulence factors (VFs) [11, 12]. Key VFs involved in each stage of *H. influenzae* molecular pathogenesis, ranging from colonisation to invasion and immune evasion, are summarised.

H. influenzae initiates infection by adhering to epithelial cells in the human respiratory tract. This is primarily mediated by various adhesins such as pili, Hap, Hia, and HMW1/HMW2 proteins, which facilitate attachment to host mucosal surfaces [11, 12]. The host proteins targeted by these adhesins are mainly extracellular matrix (ECM) proteins like laminin and vitronectin, which become exposed due to inflammation or viral infections [13-15]. Protein E and Protein F are other examples of proteins expressed by *H. influenzae* which mediate binding to ECM [14]. Multiple mechanisms contribute to bacterial evasion of mucociliary clearance, including both direct and indirect effects. Directly, *H. influenzae* lipopolysaccharides (LOS) can reduce ciliary movement by activating host protein kinase C epsilon [16]. Indirectly, the formation of biofilms enhances bacterial persistence through disruptions of mucociliary transport [12, 13, 17].

After successful colonisation, *H. influenzae* adapts to the nutrient-limited environment of the nasopharynx through metabolic flexibility and scavenging systems, including the ability to acquire iron and heme from the host [17]. Additionally, the bacterium LOS triggers inflammation while resists host antimicrobial peptides by modifying its LOS composition through incorporation of host-derived sialic acid, which helps camouflage the bacterium and reduce immune recognition [17]. *H. influenzae* also produces IgA1 proteases cleave host immunoglobulins, weakening mucosal immunity and facilitating sustained bacterial presence [12].

Once established, *H. influenzae* can penetrate deeper tissues by disrupting tight junctions between epithelial cells and invading host cells. This is facilitated by Hap and HMW proteins, which promote bacterial aggregation and internalisation [14, 15]. Furthermore, *H. influenzae,* specifically NTHi, have an active approach to promote internalisation to macrophage through expression of macrophage survival factor proteins, encoded by genes *msfA1-A4* [15, 18]. Through this route, while still directly avoiding immune detection and killing, *H. influenzae* can cross the endothelial layer and enter the bloodstream.

Persistence in the bloodstream relies on two factors: first, the capacity to continuously evade immune detection and killing, and second, the capability to proliferate in an iron-limited environment. For encapsulated *H. influenzae*, the polysaccharide capsule resists phagocytosis [17], while the NTHi mostly relies on modulating expression of its surface-exposed antigens, primarily through phase-variation [13, 17]. Another important mechanism of immune evasion is through factor H binding, a complement regulatory protein. *H. influenzae* possess multiple, highly efficient iron and heme acquisition systems, such as HgpABC and HhuA for haemoglobin and haemoglobin-haptoglobin utilisation; HxuABC for heme-hemopexin utilisation; TbpAB for transferrin and lactoferrin binding; HitABC for ferric-iron transport system [12].

1.2. The changing epidemiology of *H. influenzae* in carriage and disease

1.2.1. The impact of Hib vaccination: Successes and challenges in reducing Hib-related disease.

In the pre-Hib vaccine era, epidemiologic studies reported that incidence rates of Hib meningitis in children aged 0–4 years ranged from approximately 19 to 69 per 100,000 in the United States. This translates to approximately 1 in 200 children developed an invasive Hib

infection before the age of 5, amounting to around 12,000 cases of Hib meningitis annually [19, 20]. The rates were even higher, between 150 and 450 per 100,000, in indigenous populations in regions such as Alaska, Northern Canada, and parts of Australia [20]. Case fatality rate also varied significantly by region, from 3-6% in developed nations in the U.S. and Europe, to as high as 28% in Africa [19, 20].

Following the introduction of Hib polysaccharide-conjugate vaccines in the late 1980s, the epidemiology of Hib infections shifted dramatically. In high-income countries, routine vaccination led to reductions of over 90–95% in the incidence of invasive Hib disease. For example, surveillance data from the United States showed a decline in Hib meningitis incidence in children under 5 by more than 95% by the mid-1990s. In England and Wales, the incidence in this age group dropped from 35.5 per 100,000 before immunization to just 0.06 per 100,000 post-vaccine introduction, and even lower to 0.014 per 100,000 during the COVID-19 pandemic [20, 21]. Out of 5852 serotyped *H. influenzae* isolates collected from 2012/13 to 2022/23, only 118 (2%) were Hib and 84% of these cases occurred in adults (median age 51 years old). Additionally, the most common clinical diagnosis in Hib cases was bacteraemic pneumonia (56%) [21]. Reports from Ontario, Canada from 2014 to 2018 revealed that out of 1,273 serotyped *H. influenzae* from invasive cases, only 2.3% were due to Hib, with the incidence rate of 0.04 per 100,000 [22]. In Denmark, out of 638 *H. influenzae* isolates from invasive disease cases, 62 were Hib [23]. Hib invasive diseases are now more common in adults over 65 years old and infants less than 1 year-old [21, 22].

Similar dramatic declines have been documented in resource-limited countries; in The Gambia, the annual incidence of Hib meningitis fell from 60 per 100,000 to virtually 0 after vaccine introduction [19]. However, in Argentina, 41.1% of all *H. influenzae* isolates

originating from invasive disease, were found to be Hib. This number is much higher compared to findings in other countries and was attributed to factors such as irregular vaccine coverage rates, delayed vaccination schedules, and non-compliance with booster doses (national average coverage: 79.1%) generating insufficient herd immunity [24]. Even though not as high, Hib was responsible for 21.5% of invasive *H. influenzae* cases in Brazil between 2009 and 2021. This finding was thought to be caused by declining Hib vaccination coverage in Brazil, which dropped to 71.5% in 2021 from the ideal >95% [25]. National health data from Thailand between 2015 and 2019 found an overall incidence of *H. influenzae* disease of 1.5-1.9 per 100,000. However, serotype data were not available [26].

Despite widespread implementation of Hib vaccination programs, Hib infections remain a concern due to waning long-term immunity and the persistence of disease in partially vaccinated or unvaccinated populations. Hefele *et al.* investigated Hib seroprevalence in Lao People's Democratic Republic between unvaccinated adolescents and vaccinated children (3 doses at 6, 10, and 14 weeks of age). In both groups, nearly all had anti-Hib antibody titre for short-term protection while long-term protection was only detected in 45.6% of unvaccinated and 58.9% vaccinated group [27]. Hong *et al.* supported this finding: in France, seroprevalence failure only occurred in children immunised according to the 2+1 schedule, not the 3+1 schedule [28]. In Portugal, among 41 Hib cases in a twelve-year period, 26 were classified as vaccine failures, which were defined as Hib invasive disease occurring ≥2 weeks after 1 Hib dose given after the first birthday, or ≥1 week after ≥2 doses given before 1 year of age. Two children died from epiglottitis and one child developed sensorineural hearing loss [29]. Even though inclusion of Hib vaccine in the national immunisation program has been evaluated to be highly cost-effective, China has yet to implement this [30].

1.2.2. Emergence of invasive NTHi and non-type b capsulated *H. influenzae*

Because current vaccination strategies target only Hib, public health authorities have reported that the majority of invasive diseases caused by *H. influenzae* are now attributed to NTHi. In addition, other serotypes have become increasingly prevalent. For instance, in England, invasive infections were predominantly caused by NTHi (83%), followed by serotype f (Hif, 10%), serotype e (Hie, 3%), and serotype a (Hia, 1%) [21]. Similarly, most studies confirm that NTHi has emerged as the leading cause of invasive *H. influenzae* disease [20-22]; however, there are regional variations, such as in Argentina and Brazil, where NTHi accounted for only 44.5% and 51.4% of cases, respectively [24, 25].

Weinberg *et al.* reported the first invasive NTHi outbreak in a Detroit elementary school in 2023 involving four cases among children aged 5–6 years. Outbreak investigation was triggered after the death of a child due to invasive *H. influenzae* disease followed by three additional cases within 7 days in the same school and grade. The three children were hospitalised. All patients were non-Hispanic Black or African American boys with symptoms including fever, myalgia, lethargy, and headache [31].

McTaggart *et al.* observed a slight increase in trend of overall invasive *H. influenzae* disease between the year 2014 and 2018 in Ontario, Canada, from 1.67 to 2.06 cases per 100,000 population, representing a 5.6% annual percent change. NTHi accounted for 74.2% of infections. Nevertheless, in contrast to findings from England, infections caused by Hia represented a higher proportion in their study, constituting 8.9% of all cases [22]. Although Hia infections have historically been rare in England, recent years have shown a steady increase in Hia cases across all age groups. In the latest period (2021/22), 19 Hia cases were

reported, compared to an average of 8 Hia cases/year in the five epidemiological years prior [32].

Brown *et al.* reported that invasive *H. influenzae* disease disproportionately affects specific populations, highlighting that the incidence among American Indian and Alaska Native (AI/AN) patients was significantly higher (3.0 cases per 100,000 population) compared to the overall national rate of 1.8 per 100,000. This disparity becomes even more evident when analysing serotype distribution: 37.8% of cases among AI/AN patients were attributed to serotype a (Hia), compared to only 5.8% in the general population [33]. Although NTHi is commonly implicated in invasive disease, a multicenter study conducted across ten Canadian hospitals found meningitis was still notably associated with typeable strains, especially Hia, which accounted for 39% of paediatric meningitis cases. In contrast to other reports, NTHi represented only 36% of 118 cases, while Hia accounted for the majority (54%), followed by Hif at 26% [34].

Another encapsulated serotype of *H. influenzae* currently gaining prominence is Hif. A systematic review by Reilly *et al.* evaluated the epidemiological trends of invasive Hif infection and found clear evidence of an increasing incidence, particularly following widespread implementation of Hib vaccination programs. The review reported a pooled incidence of invasive Hif infections 0.15 (0.05-0.40) per 100,000 population per year. Many studies included in the review compared this incidence rate before and after Hib vaccine implementation and the median increase in Hif incidence reached over 260% [35].

There are still limited data on the prevalence of *H. influenzae* from lower- and even upper-middle income countries (LMIC), specifically in the WHO South-East Asia Region (SEARO). This region consists of countries that face multiple health challenges with high

disease burden and low government spending on healthcare, compared to the global average [36]. The last data on *H. influenzae* invasive disease from Bangladesh was from the year 1999-2003, 6 years before the implementation of Hib vaccine into the country's national immunisation programme. As expected, 98% of *H. influenzae* isolates were type b [37]. Similarly, only one report from Thailand on the burden of *H. influenzae* disease among children which required hospitalisation prior to the implementation of Hib vaccine. However, serotype data were not unavailable [26].

Until 2024, no data on *H. influenzae* invasive disease from Indonesia were available although several studies concerning carriage rates had been undertaken with studies examining *H. influenzae* as the aetiology of non-invasive disease. Overall *H. influenzae* carriage prevalence among healthy children aged 12-24 months old was 27.5% [38] and 69.7% of acute otitis media among school children were caused by this organism, 95.3% of which non-typeable [39] Non type-b *H. influenzae* was also the most common causative pathogen for hospitalised community-acquired cases in children aged 2-5 (n = 73/188, 38.%). This study was conducted just after the inclusion of Hib vaccine in the national immunisation program [40].

1.3. Typing strategies and the molecular epidemiology of *H. influenzae*

1.3.1. Genomic typing methods: MLST, cgMLST, and other genomic classification tools and schemes.

Molecular typing of bacterial pathogens has become an indispensable tool in epidemiology, infection control, and microbial research. Conventional typing methods rely on phenotypic characteristics such as serotyping, usually based on serological reactions

between antisera and capsular antigens, or "biotyping", which are mostly based on enzymatic reactions. The "biotyping" scheme for *H. influenzae* classified isolates based on the expression of three metabolic enzymes: ornithine decarboxylase (ODC), urease, and tryptophanase. While there was a strong correlation between genotype and phenotype in this biotyping method, it did not provide insights into genetic relatedness or clinical disease outcomes, limiting its practical utility [23, 41].

In addition, phenotypic typing methods often lack the resolution required to differentiate between closely related variants. In contrast, molecular typing techniques, which target the genetic material of bacteria, offer a higher degree of precision and reproducibility. Methods like pulsed-field gel electrophoresis (PFGE), multilocus sequence typing (MLST), and variable-number tandem repeat (VNTR) analysis have been widely adopted for their ability to trace outbreaks, monitor pathogen evolution, and inform public health interventions. PFGE, which had been the standard for outbreak investigations in the past, relies on the separation of large DNA fragments to generate distinct banding patterns for comparative analysis. MLST provides a sequence-based approach by indexing the allelic variation in multiple housekeeping genes, making it a robust tool for long-term and global epidemiological studies [42, 43]. Each sequence type (ST) is composed of a unique allelic variation and similar STs can be clustered into a clonal complex (CC), typically using the globally-optimised BURST method [44].

Over the past decade, an increasing number of bacterial typing methods have emerged, largely driven by the growing accessibility of whole-genome sequencing (WGS). However, the European Society of Clinical Microbiology and Infectious Diseases (ESCMID) Study Group for Epidemiological Markers indicated in 2013 that WGS-based analyses were still not rapid

enough for efficient data extraction, synthesis, and interpretation, primarily due to the need for extensive bioinformatics support [45]. This concern remained relevant six years later, with experts continuing to highlight the challenges posed by a lack of bioinformatics expertise and the complexity of sequencing data interpretation [46].

In outbreak investigations and epidemiological tracking, the primary objective is to determine whether bacterial isolates are genetically identical or distinct. To achieve this, an approach with the highest discriminatory power is usually preferred when high-resolution genomic data are available. With the availability of genomic data, often obtained through WGS, a reference-free gene-by-gene analysis, as exemplified by MLST, can be expanded to include tens or even hundreds of full-length genes within a given species. When a broader set of highly conserved core genes, defined as those present in nearly all isolates of a bacterial species, is incorporated into this gene-by-gene approach, it is referred to as core genome multilocus sequence typing (cgMLST). Like MLST, which characterises bacterial isolates based on the allelic variation of a limited number of housekeeping genes, cgMLST employs a similar process of allele designation. Each unique sequence variant of a core gene is assigned a distinct allele number, and the combination of allele numbers across the core gene set is used to define a unique core gene sequence type (cgST) [47, 48]. The PubMLST database was developed to store allelic variants and ST definitions for gene-by-gene typing schemes of important bacterial pathogens. It allows users to upload their own genomic data and annotate it with relevant typing information. In addition to being regularly updated and publicly available, the database features a web-based, user-friendly interface. It is also equipped with comparative genomics plugins, enabling users to perform analyses in a seamless, plug-and-play manner [49, 50].

In 2003, Meats *et al.* first published the MLST scheme for *H. influenzae* which was implemented to 131 *H. influenzae* isolates to characterise the bacterial population structure [51]. The seven housekeeping genes included in the scheme were chosen based on their function and location in the *H. influenzae* Rd genome sequence, the first *H. influenzae* whole-genome sequence [51, 52]. The selection criteria for these genes were as follows: (1) involved in general metabolic processes, (2) located in genomic regions predominantly containing genes with conserved functions, and (3) not subject to diversifying selective pressure [51]. The final loci included in the *H. influenzae* MLST scheme are summarised in **Table 1.2**.

Table 1.2. The seven locus MLST scheme of *H. influenzae*.

Locus	Gene product	Length of sequenced fragment for MLST (bp)
adk	Adenylate kinase	477
atpG	ATP synthase F1 subunit gamma	447
frdB	Fumarate reductase iron-sulfur protein	489
fucK	Fuculokinase	345*
Mdh	Malate dehydrogenase	405
Pgi	Glucose-6-phosphate isomerase	468
recA	RecA protein	426

*Fragment length for *fucK* can vary between 344-346 as stated and recorded on the PubMLST *H. influenzae* typing database https://pubmlst.org/bigsdb?db=pubmlst_hinfluenzae_seqdef&l=1&page=downloadAlleles&tree=1 (accessed 21st January 2025)

Previous studies on the *H. influenzae* genome have defined both its pan-genome and core-genome sizes. They have also demonstrated that a cgMLST scheme can deliver high-resolution characterisation of the species population structure. However, most of these studies analysed fewer than 500 draft genomes, which is considered the minimum threshold for reliable detection of paralogous loci, an essential step in core genome analyses [53-55]. Paralogs, which are duplicated genes originating from a single ancestral gene, can complicate genetic comparisons and obscure true evolutionary relationships among genomes [56, 57].

While some studies have applied cgMLST for *H. influenzae*, the schemes used were defined by proprietary software, with typing definitions not publicly accessible [58, 59]. Consequently, none of the existing cgMLST schemes have been validated or implemented on publicly available platforms.

Beyond MLST-based typing, one attempt was made to establish alternative typing schemes for *H. influenzae*, known as the pathotype concept, which was specifically developed for the NTHi population. This method classified strains based on phylogenetic relationships inferred from core genome single nucleotide polymorphism (SNPs). Six distinct clades (Clade I–VI) were identified, with classification determined by the presence or absence of 17 accessory genes [54, 60, 61]. While this approach provided some genomic insight into strain diversity, its clinical relevance remains limited, and it is not applicable to encapsulated *H. influenzae* strains.

The most widely used typing method in *H. influenzae* molecular epidemiology studies has been the 7-locus MLST, which defines STs and CCs. However, when higher-resolution population structure analyses were performed using WGS, studies often reconstructed phylogenies without a standardised typing scheme or clustering framework, leading to arbitrary naming and numbering of genomically similar isolates into “clades” or “clusters” based on subjective interpretation or predefined clustering thresholds. For example, in 2019, Potts *et al.* conducted a genomic characterization of 688 *H. influenzae* isolates collected in the United States between 1999 and 2017 and, based on maximum-likelihood phylogenetic analysis, categorized them into three clades: Clade I, II, and III [3]. A prospective study in Germany analysed 215 *H. influenzae* isolates collected between October 2019 and March 2020 and identified six genetic clusters using an arbitrary distance threshold in a minimum-

spanning tree [58]. The inconsistency in clustering methodologies between these studies makes direct comparisons difficult, as the three clades from the U.S. study and the six genetic clusters from the German study cannot be reliably correlated. As a result, determining the genetic similarity or divergence between isolates across different studies remains challenging due to the lack of a standardised framework for WGS-based classification. This finding indicates the need for reproducibility and stability in genomic characterisation of *H. influenzae*.

1.3.2. Molecular epidemiology of circulating *H. influenzae*: Dominant clonal complexes in capsulated strains and their absence in NTHi.

The molecular epidemiology pattern between encapsulated *H. influenzae* and NTHi is very different to each other. Capsulated *H. influenzae* serotypes (Hia, Hib, Hic, Hid, Hie, and Hif) are primarily associated with one or a few dominant STs. NTHi STs did not cluster into tight clonal complexes, reflecting the high level of genetic recombination and diversity within this group. When placed in a phylogenetic context, this genetic distinction suggests an evolutionary split between encapsulated and NTHi lineages, emphasising the greater genomic stability of encapsulated lineages compared to the highly heterogeneous NTHi population [3].

Genomic analysis revealed that 88% of the sequenced Hib isolates from England (2012/13-2022/23) belonged to the "CC6 lineage" (defined as ST 6 and its single-locus variants) [21]. Although the term "lineage" is used variably depending on phylogenetic context [62], an ST is strictly the unique allelic profile of an isolate across all MLST loci [63]. A similar pattern was observed in *H. influenzae* isolates collected through the U.S. national surveillance program over an 18-year period, where ST 6 was the predominant ST associated with Hib.

Additionally, a smaller ST, ST 222, was also identified as being linked to the capsule type [3]. Zanella *et al* reported that most (68.4%) of Hib isolates from Brazil belonged to ST 6 [25]. A report from Denmark found that all Hib isolates belonged to the ST-6 complex, which comprised of ST 6, ST 59, ST 190, ST 206, ST 709, and ST 1448 [23]. In cases of Hib associated with vaccine failure, all analyzed isolates were also part of the ST-6 complex, with the majority (92%) classified as ST 6 [29].

The vast majority of Hia strains from a prospective national study in England belonged to a few closely related STs (ST 1511, ST 23, ST 56, ST 576), all part of the ST-23 complex, collectively accounting for approximately 78% of the total isolates [32]. Similarly, 79% of USA Hia isolates characterised by Potts *et al* belonged to this clonal complex and 55.5% of Hia isolates from Brazil belonged to ST 23 [25]. A comparative genomic study of multinational invasive Hia isolates consistently found most of these isolates to be part of this CC, while the rest belonged to a distinct ST, the ST 62. This ST was not associated with any clonal complex [64].

For other capsule types, Hic, Hid, and Hif are mainly associated with ST 9, ST 10, and ST 124, respectively, while Hie isolates cluster into five closely related STs (ST 18, ST 66, ST 121, ST 127, ST 386) [3]. Multiple studies also showed that that ST-124 and ST124 complex were associated with Hif [3, 23, 35]. These data, along with a more in depth comparative pangenome analysis of Hif, confirm the low genetic variability of this capsule group, suggesting that the increasing Hif incidence is not driven by capsule switching or major genetic diversification [35, 65]. In fact, this pattern was also observed in other capsule types, indicating that genetic adaptations, other than capsule switching, enhancing virulence and

immune evasion was the underlying mechanism of increasing incidence of non-Hib capsulated *H. influenzae* [3, 66].

A total of 293 NTHi isolates from the USA belonged to 125 different STs [3] with 479 isolates from Denmark associated with 131 different STs, including 40 novel STs which had not been identified elsewhere [23, 67]. This high diversity was also apparent across different geographical regions, including Norway [68], Germany [58], Japan [69], and China [15]. ST 103 is the most common among NTHi in Norway and Denmark [23]. Frank *et al* reported that most NTHi from non-invasive cases across Germany belonged to ST 57, followed by ST 103 [58]. However, ST 12 is the dominant NTHi ST based on studies conducted in Japan and China [15, 69]. The only reports indicating specific high-risk NTHi ST/CC were ST 164 and ST 1714 which were associated with NTHi invasive disease among adults with HIV in Atlanta, Georgia [70]. In addition, a recent NTHi outbreak case in Michigan, identified the isolates as part of the ST 714, which showed close genetic relationships with a previously reported cluster in Georgia [31].

There are few to no published reports on the molecular epidemiology of *H. influenzae* from countries within the WHO Southeast Asia Region (SEARO). This lack of data limits our understanding of the genetic diversity and population structure of *H. influenzae* in this region, which is critical for effective disease surveillance and vaccine policy decisions. Expanding genomic surveillance efforts in SEARO countries would provide valuable insights into regional strain distribution and emerging resistance trends, ultimately aiding in the development of targeted public health strategies.

1.4. The rise of resistant *H. influenzae* strains

1.4.1. Antimicrobial resistance (AMR) epidemiology in *H. influenzae*

H. influenzae is among the twelve WHO Global Antimicrobial Resistance Surveillance System (GLASS) target bacterial pathogens [71]. Ampicillin-resistant *H. influenzae* is designated "medium priority" on the WHO Bacterial Priority Pathogen List because it causes serious infections and ampicillin remains a cornerstone of empirical therapy for community-acquired disease. The emergence of ampicillin resistance in *H. influenzae* therefore represents a significant public-health threat, highlighting an urgent need to address this problem in clinical and public health settings [72]. As this bacterium naturally resides in the human nasopharynx, it is constantly under the selective pressure of excessive exposure to first-line antibiotics, such as ampicillin, [73] due to antibiotic self-administration practice or unnecessary prescription [74].

McTaggart *et al* reported susceptibility of *H. influenzae* isolates in Ontario, Canada within the period 2014-2018. Resistance to ampicillin varied by serotype, ranging from 0.8% in Hif (N = 128) to 27.2% in NTHi (N = 940). The number of antibiotics for which each serotype shows nonsusceptibility (NS > 0%) was also largest in NTHi, with resistance to 8 antibiotics found (ampicillin, ampicillin-sulbactam, cefaclor, cefuroxime, chloramphenicol, clarithromycin, sparfloxacin, trimethoprim-sulfamethoxazole). In contrast, resistance to 5 antibiotics was detected in Hia and Hie, 4 in Hif, and 3 in Hib. Additionally, prevalence of resistance to some was higher in a specific serotype compared to others, such as 54.2% clarithromycin non-susceptibility in Hie and 43.3% cotrimoxazole non-susceptibility in Hib. Cefaclor non-susceptibility was highest among NTHi (4.9%) while resistance to cefepime, a fourth-generation cephalosporin, was only apparent in Hif (1.6%) [22]. Complementary to this

study was a systematic review by Reilly *et al.* which focused only on Hif. Here, it was observed that 11.5% Hif were resistant to ampicillin (n = 22), 11.5% to cotrimoxazole (n = 22), and 1% to clarithromycin (n = 2) [35].

Similarly, the rate of ampicillin resistance among *H. influenzae* isolated from invasive cases in Brazil was 17.1% (N = 235). 60% of ampicillin-resistant isolates were NTHi, amounting to 20% of ampicillin resistance prevalence within NTHi (N = 709). Almost all (233/235) ampicillin-resistant isolates were beta-lactamase producers. No resistance was observed for ceftriaxone or rifampicin. This study also reported an azithromycin resistance rate of 0.1% and ciprofloxacin resistance of 0.2%. In contrast to McTaggart's report which found almost no cases of chloramphenicol resistance, here it was observed in 17.1% of *H. influenzae* isolates [25].

Until January 2025, only a few systematic reviews evaluating the existing evidence of AMR in *H. influenzae* have been published [74-76]. Abavisani *et al.* conducted a systematic review and meta-analysis on multidrug resistant (MDR) *H. influenzae*. Sixteen studies from 9 countries were included in the synthesis, all of them tested antibiotic susceptibility in accordance with the Clinical and Laboratory Standards Institute (CLSI). Among a total of 19,787 *H. influenzae* isolates across the included studies, five antibiotics with the highest pooled prevalence were sulfamethoxazole (45.6%), ampicillin (36%), tetracycline (19.9%), cefuroxime (19.1%), and azithromycin (15.3%). This study also calculated the pooled prevalence of MDR *H. influenzae,* amounting to 23.1%, with slightly higher rate among isolates originating from Asian countries (24.6%) compared to the rest (15.7%) [74]. However, there was no uniform definition on what constituted as MDR across different studies. For example, Su *et al.* defined MDR *H. influenzae* as those non-susceptible to at least one agent in three to

four antibiotic groups, regardless of the group [77]. Zhou *et al.*, on the other hand, did not formally define this but categorised MDR *H. influenzae* in the same way as Su *et al.*, except that every single MDR isolate had to be resistant to a beta-lactam group [78]. Lastly, a report from Japan by Yamada *et al.* only identified one MDR isolate, which was defined as resistant to penicillin, penicillin/beta-lactamase inhibitor, macrolides, tetracyclines, and quinolones [79].

These data are primarily derived from studies between 2020-2024; however, overall, there has been little change in *H. influenzae* resistance rates across multiple antibiotic classes. An older systematic review examined AMR in *H. influenzae* across 15 Latin American countries over a decade (1990–2000). Findings indicated that ampicillin resistance was observed in 17.2% of non-invasive isolates and 21.9% of invasive isolates. Resistance to co-trimoxazole was notably higher, affecting 41.9% of non-invasive strains and 26.9% of invasive strains [75]. These findings suggest that, while resistance patterns have remained relatively stable over time, persistent AMR challenges necessitate continuous surveillance and interventions such as effective antibiotic stewardship programs.

1.4.2. Genetic determinants of AMR in *H. influenzae*

H. influenzae has emerged as a significant public health challenge due to its evolving resistance to multiple classes of antibiotics. In particular, beta-lactam resistance is commonly driven by the production of β-lactamases (e.g., blaTEM-1) and alterations in penicillin-binding proteins, particularly the PBP3, which diminish the efficacy of first-line treatments like amoxicillin. Due to the growing prevalence of ampicillin-resistant *H. influenzae* in various countries, current treatment guidelines now favour using β-lactam/β-lactamase inhibitor

combinations or third-generation cephalosporins. When these options are not suitable because of confirmed resistance or patient intolerance, alternative therapies such as co-trimoxazole, macrolides, or fluoroquinolones may be considered [76]. The increasing resistance to macrolides further complicates treatment, especially among paediatric patients as this antibiotic group has a favourable safety profile for this population [80]. Fluoroquinolone resistance poses an additional threat by limiting alternative therapeutic options [81]. Moreover, the increasing use of co-trimoxazole for multiple clinical indications, especially in developing countries, has been paralleled by the emergence of resistance mechanisms that compromise its utility [82, 83].

Table **1.3** summarises AMR determinants associated with resistance to beta-lactams, macrolides, fluoroquinolones, and co-trimoxazole that had previously been reported in the literature.

Table 1.3. Genetic determinants of AMR in *H. influenzae*

Antibiotic group	Gene	Mutation/Resistance mechanism	Notes	References
Beta-lactam	*ftsI*[a]	**Critical mutations:** Amino acid substitutions in PBP3 at the entrance of the active site pocket (i.e. conserved motifs STVK, SSN, KTG): G490E, A502V, R517H, and N526K **Additional mutations:** near the active site, usually present in combination with critical mutations: S357N, M377I, S385T, and L389F **Other identified mutations:** V329I, D350N, A368T, A388V, P393L, A437S, I449V, I491V, R501L, A502S, A502T, A502V, V511A, I519L, A530S, and T532S, V547I, N569S, E603N	Associated with BLNAR (β-lactamase-nonproducing resistance); accumulation of substitutions increases resistance levels. When combination of substitutions involved R517H + S385T + L389F: associated with highest median ceftriaxone MIC	[84, 85] (Critical and additional mutations) [58, 86] (other identified mutations) [87] (*ftsI* grouping) [88] (ceftriaxone resistance combination)
	blaTEM-1	Acquisition of the TEM-type-β-lactamase gene (class A broad spectrum beta-lactamase). Variants of blaTEM-1: TEM-1B (most common), TEM-234, TEM-104, TEM-1C	Found on plasmids or ICE; linked to BLPAR.	[84, 85, 89] [67]
	blaROB-1	Acquisition of the ROB-1 beta-lactamase	-	[84]
	lpoA	Amino acid substitution: M151I	Encodes lipoprotein A, a cofactor for PBP1a	[58]
Macrolide	*acrR*	Nonsense or frameshift mutations disrupting the negative regulator of the AcrAB efflux pump	Leads to overexpression of the AcrAB pump and clarithromycin resistance.	[85]
	acrB	Amino acid substitution: R327S	Mutation acquired in a stepwise manner contributing to increased resistance.	[85, 90]
	acrR	Premature stop codon Amino acid substitution: R327	A regulatory gene for *acrAB*	[91]
	mefA/E	An exogenous macrolide resistance gene	Often results in high-level resistance; may be horizontally transferred.	[85, 89, 91]

			Less common than mutational mechanisms.	
	msrD	An exogenous macrolide resistance gene Works together with *mefE* as a part of a macrolide efflux system.	Less common than mutational mechanisms.	[91]
	ermB	Acquisition of the *ermB* gene encoding a 23S rRNA methylase that modifies the macrolide target site	Cherkaoui et al: All fluoroquinolone-resistant isolates possess *ermB* gene	[81, 91]
	Ribosomal protein L4 gene	Amino acid substitution: T64K/T64R, T66R/T66K, T121S	-	[81, 91]
	Ribosomal protein L22 gene	Amino acid substitution: G91D Insertion/deletion in amino acid location 86-96		[91]
	23s rRNA	Point mutation: A2058G, A2059C, A2611G, and C2610T	Mutations in 23S rRNA correlate with high azithromycin MICs; increased number of altered copies yields higher resistance levels.	[91]
Fluoroquinolone	*gyrA*	Amino acid substitution[b]: S84L, D88G	Initiates reduced susceptibility; a primary event in the development of resistance. Multiple substitutions can exist at one position.	[58, 81, 85]
	parC	Amino acid substitution[b]: S84I	When combined with *gyrA* mutations, leads to higher MIC values and enhanced resistance.	[58, 81, 85]
	parE	Amino acid substitution: D420N	-	[81]
Co-trimoxazole	*folA*	Amino acid substitution: I95L, F154S/V	Encodes dihydropteroate reductase	[58]
	folP	Amino acid substitution and insertion: P64Ins(SFLYN), N65D and G189C	Encodes dihydropteroate synthase	[58]
	sul2	Acquisition of the *sul2* gene which encodes for a dihydropteroate synthase	-	[58, 92]

dfrA1	Acquisition of the *dfrA1* gene which encodes for dihydropteroate reductase that is less susceptible to trimethoprim	-	[92]

[a)] Based on the variations/polymorphisms in the resistance-related region (i.e. active site pocket) in PBP3 protein, *ftsI* gene is categorised into four groups. **Group I** is considered as wild-type sequence, although there may be some polymorphism outside resistance-related region. Isolates with *ftsI* variant belonging in this group remain susceptible to beta-lactam group antibiotics. **Group II** variants possess polymorphisms within the resistance-related region, but without critical mutation. Phenotypically, isolates are still susceptible to beta-lactams. **Group III** had critical mutations in the active site and **group IV** had additional mutations near the active site, as detailed in the **Error! Reference source not found.**. Isolates with group III mutations are typically resistant to amoxicillin and amoxicillin/clavulanic acid but susceptible to third-generation cephalosporins [84, 87]. This classification was developed based on the construction of *ftsI* phylogenetic tree.

[b)] Typically in the quinolone resistance-determining region (QRDR)

PBP: penicillin-binding protein, ICE: integrative conjugative element, BLPAR: β-lactamase-producing resistance, MIC: minimum inhibitory concentration

1.5. Genomic insights: The *H. influenzae* pangenome and the role of recombination in shaping *H. influenzae*'s genetic diversity

In 2007, Hogg *et al* reported the first characterisation of *H. influenzae* pangenome, specifically for the nontypeable strains. After conducting gene clustering and pairwise comparison, it was concluded that there were around 1,450 core gene clusters shared among 13 NTHi genomes. However, there was a total of 2,786 gene clusters and each strain had a set of unique genes not discovered in other strains [93]. In an attempt to develop a pangenome array for *H. influenzae*, Eutsey *et al* conducted a similar evaluation for 24 NTHi strains, which covered the 13 strains originally assessed by Hogg *et al*. Applying the same approach, more gene clusters were identified reaching a total of 4,547 [94]. The two studies proved that *H. influenzae* displayed an extensive diversity based on the number of shared (i.e. core) genes. Subsequently, Power *et al* further investigated species diversity at SNP level, utilising 18 Hib strains and identified regions of high SNP density, which was proven to represent recombination events [95], possibly promoted by the bacterium's natural competence to undergo transformation [96].

This natural competence in *H. influenzae* can impact genome-wide sequence similarity, far more than the point mutations passed down through clonal inheritance, which refers to the propagation of genetic material by **vertical transmission** from a parent cell to its progeny via binary fission [97]. This could be problematic in phylogenetic reconstruction because most methods, including Maximum Parsimony (MP), Maximum Likelihood (ML), and Bayesian Inference (BI), assume that all sites in a sequence evolve under a common ancestry without lateral genetic exchange [98]. Therefore, accounting for the effects of recombination

and removing recombinant regions during phylogenetic reconstruction is crucial for accurately depicting the genetic relatedness and population structure of *H. influenzae*.

In general, there are two approaches used to remove recombination effects. The first approach relies on finding sequence exchange based on the evidence of unexpected similarity between genetically distinct isolates. When this evidence is present, it reflects either: 1) movement of sequence from donor to recipient, or 2) the unrelated isolates acquired the sequence independently from a common donor. Through the comparison of sequence alignments of different regions in the genome, recombination breakpoints can be identified. For example, within a sequence alignment, two regions are evaluated. Putative recombination events can be defined when one region aligns well with one evolutionary lineage but the other aligns better with a different lineage and the location between the two regions is the recombination breakpoint. Algorithms such as RDP3, cBrother, and GARD use this principle [99].

The second approach searches for evidence of variable sequences among closely related isolates. Therefore, putative recombinant regions are defined as regions containing a high density of polymorphisms, relative to the background level, in a concept known as 'clonal frame'. This approach is most useful to study recombination within a subset of bacterial species (i.e. isolates belonging to the same ST or lineage in the species), especially if sequence imports originate from other lineages [99, 100]. The ClonalFrame model used is implemented by Bayesian inference. Although typically recombination between isolates within the same lineage (i.e. internal import) is best modelled by a different model, the coalescent with gene conversion, ClonalFrame model still has a role in this scenario. However, detecting recombination in this case is more difficult because the source of recombination originates

from within the population and hence less substitutions are introduced. As a result, the relative recombination rate (R/theta) is consistently underestimated and more pronounced when the recombination events involve only a short tract of sequence [100].

Several studies investigating recombination events in *H. influenzae* examined specific genes or regions in the genome, with an emphasis on AMR determinants. For example, Michel *et al* evaluated three MDR *H. influenzae* isolates from Belgium and found evidence of interspecies *murF, murE,* and *ftsI* recombinant in all three. In each of these genes, there were regions with high base substitution density with a high percentage of similarity of genes found in other *Haemophilus* species, including *H. haemolyticus* and *H. parainfluenzae* [101]. An *in vitro* study confirmed that mutated *gyrA* and *parC* from a quinolone-resistant strain was transformed to recipient strains and followed by an increase MIC [102]. However, there were two previous reports evaluating recombination rate across different bacterial species, including *H. influenzae*. First, Vos and Didelot implemented the ClonalFrame model to predict relative recombination rate (r/m) of the species based on the 7 MLST -loci. The dataset used in this experiment was comprised of 104 and 27 capsulated and non-capsulated *H. influenzae*, respectively. The predicted r/m for *H. influenzae* was 3.7 with CI 95% 2.6-5.4. As comparison, another commensal/opportunistic human pathogen *Staphylococcus aureus* was predicted to have r/m of 0.1 (CI 95% 0-0.6) while the highly recombining species *Helicobacter pylori* had an r/m of 13.6 (CI 95% 12.2-15.5) [103]. The second report by Torrance *et al* evaluated the same recombination parameter by employing approximate Bayesian computation (ABC) method using a core genome of bacterial species through mapping to a specific reference strain. This method integrates multiple recombination signals, including homoplasies, linkage disequilibrium, and polymorphism pattern and the predicted r/m for *H. influenzae* was 19.26 (CI 95% 11.9-20.86). When compared to prediction results from other species, there is a

potential inaccuracy for *H. influenzae*, necessitating a more representative sampling of bacterial genomes [104].

Both in vitro biological evidence and computational simulations strongly indicate that *H. influenzae* is a highly recombining microorganism. This extensive recombination likely enables *H. influenzae* to acquire genetic variations in both its core and accessory genomes, facilitating the emergence of traits such as increased virulence and antibiotic resistance.

1.6. Thesis aims

Several critical knowledge gaps remain in understanding the molecular epidemiology of *H. influenzae*. One major limitation is the lack of a standardised, publicly accessible, and universally adopted typing scheme for high-resolution genomic surveillance. While MLST has long been the standard for population structure analysis, it lacks the discriminatory power required to resolve fine-scale evolutionary relationships. WGS-based approaches, such as cgMLST, have been proposed as a solution, yet most existing schemes are either proprietary or lack validation on public platforms. Additionally, clustering methods for genomic classification remain inconsistent across studies, with different thresholds and frameworks leading to the arbitrary naming of clades and genetic clusters. This lack of standardisation hampers direct comparisons between studies and complicates efforts to track the emergence and spread of virulent and antimicrobial-resistant *H. influenzae* strains. Establishing a widely accepted cgMLST scheme, coupled with a standardised and stable clustering approach, is essential for improving surveillance, guiding vaccine strategies, and identifying high-risk lineages.

Another key area of uncertainty is the genetic basis underlying the increasing incidence of invasive NTHi. While epidemiological studies have documented a rise in invasive NTHi cases, there has been no comprehensive evaluation of the genetic determinants responsible for this phenotype. In addition, genomic data on *H. influenzae* carriage and invasive disease, particularly from WHO SEARO, remains scarce. Since asymptomatic carriage serves as a major reservoir for transmission and potential virulence evolution, understanding the genetic diversity and population structure of *H. influenzae* in diverse settings is crucial. Expanding genomic surveillance could provide valuable insights into strain dynamics, antimicrobial resistance evolution, and factors influencing invasiveness. Addressing these gaps would not only enhance our understanding of *H. influenzae* epidemiology but also inform future vaccine development and public health interventions by identifying novel targets for immunisation and tracking the emergence of strains with increased pathogenic potential.

1.6.1. Aims and objectives

Based on these conclusions, this DPhil thesis aims to address critical gaps in the molecular epidemiology and genomic diversity of *H. influenzae* by developing a publicly accessible high-resolution typing framework, investigating the genetic determinants underlying the increasing incidence of invasive NTHi, and characterising circulating *H. influenzae* in the underrepresented region, WHO SEARO, particularly Indonesia. This research will provide valuable insights into population structure, strain evolution, and factors influencing virulence in *H. influenzae*, ultimately contributing to improved public health strategies and vaccine development.

The following are the objectives of this DPhil thesis:

1. Develop and validate a cgMLST scheme for *H. influenzae.*
2. Identify genetic discontinuity thresholds for hierarchal clustering using core genome life identification number code (cgLIN) as an extension of the cgMLST scheme to define a stable, high-resolution method for characterising the population structure of *H. influenzae.*
3. Identify the genetic variants associated with the phenotypic outcome of invasive disease in the NTHi.
4. Investigate the population structure of circulating *H. influenzae* in Indonesia from carriage and disease by implementing the cgMLST scheme and cgLIN code.

1.6.2. Overview of chapters

Chapter 1: Introduction to *Haemophilus influenzae*: Epidemiology, molecular typing, and genomic insights.

This chapter provides a comprehensive overview of *H. influenzae*, covering its microbiology, epidemiology, molecular typing methods, genomic diversity, and AMR. It highlights the impact of Hib vaccination, which has reduced invasive Hib disease but led to a rise in NTHi and other serotypes, with limited epidemiological data from LMICs. A major challenge identified is the lack of a standardised, publicly accessible cgMLST scheme, hindering high-resolution comparative genomic studies. In addition, while *H. influenzae* is highly recombinogenic, the genetic determinants driving NTHi invasiveness remain unexplored. This chapter also outlines known VFs and AMR mechanisms in the species.

Chapter 2: Development and implementation of a core genome multilocus sequence typing scheme for *Haemophilus influenzae.*

In Chapter 2, the development and validation of the cgMLST scheme for *H. influenzae* is described. The dataset, consisting of publicly available complete reference genomes and high-quality draft genomes, was divided into development and validation sets to ensure the robustness of the scheme. Various pangenome analysis tools were employed to identify a comprehensive set of core genes, which were then functionally classified. Phylogenetic analyses confirmed that the cgMLST scheme reliably represents genetic relatedness among isolates

Chapter 3: Resolving *Haemophilus influenzae* molecular epidemiology using life identification number (LIN) coding.

Chapter 3 presents the implementation of the LIN method for *H. influenzae* to establish a stable genomic nomenclature for public health and research. This approach, based on core genome profiles, was integrated into the PubMLST *H. influenzae* database to facilitate standardised clustering within the species. Thirteen LIN code thresholds were defined, ranging from broad evolutionary groupings ('megalineage') to fine-scale partitions suitable for outbreak investigations. The LIN method provides each *H. influenzae* genome with a unique barcode, offering insights into genetic relationships.

Chapter 4: Variants in surface-exposed proteins are associated with invasive non-typeable *Haemophilus influenzae* infection: A genome-wide association study

Chapter 4 explores the genetic factors associated with the invasive potential of NTHi, which has become the predominant cause of invasive *H. influenzae* infections in the post-Hib vaccine era. Using a genome-wide association (GWAS) approach, the study analysed publicly available genomes to identify genetic variants linked to invasiveness. A k-mer-based method was applied, allowing the detection of both core and accessory gene variations without the need

for a reference genome. Furthermore, a subsequent validation GWAS using an independent dataset confirmed findings, followed by the development of a logistic regression model to predict the invasive outcome of NTHi infection.

Chapter 5: The population genetics of *Haemophilus influenzae* isolates from Indonesia in carriage and disease

In Chapter 5, *H. influenzae* isolates from Indonesia were characterised, focusing on serotype distribution, genomic diversity, and AMR. WGS were used to define the population structure of these isolates, utilising cgMLST scheme and LIN code defined in Chapters 2 and 3. Assessment of serotype classification was conducted using qPCR method and consolidated by analysing capsule operons. For isolates with available antibiotic susceptibility results, genotypic AMR determinants was evaluated. Presence of VFs for NTHi from invasive cases was determined, as well as the presence of genetic variants identified in Chapter 4. Finally, all *H. influenzae* isolates were compared to the curated global isolates from PubMLST.

Chapter 6: General discussion and conclusion.

Lastly, Chapter 6 presents a summary of key findings and an overall discussion of the preceding chapters (Chapters 2 to 5). This chapter concludes my DPhil thesis, outlining the main takeaways and their implications, followed by proposed future research directions to address existing knowledge gaps and tackle clinical and public health challenges related to *H. influenzae* disease.

Chapter 2

Development and Implementation of a Core Genome Multilocus Sequence Typing (cgMLST) scheme for *Haemophilus influenzae*

Abstract

Haemophilus influenzae, a frequent nasopharyngeal commensal and occasional invasive pathogen, exhibits extensive genomic diversity that demands high-resolution typing methods. Here, we report the development of a core-genome multilocus sequence-typing (cgMLST) scheme for *H. influenzae*, constructed via pangenome analysis of 14 complete reference genomes and 2,297 high-quality draft assemblies. The draft genome was divided into a development cohort (n = 921), used to identify candidate core loci, and a validation cohort (n = 1 376), used to refine the final list of loci to ensure the reliability of the proposed cgMLST scheme. Functional annotation classified the resulting core genes, and congruence between allelic-profile and nucleotide-alignment phylogenies was confirmed by Spearman correlation and ordinary least-squares (OLS) regression ($R^2 = 0.945$). Initial screening yielded 1,067 core genes, which were finalised at 1,037 after validation; over 70% encode proteins involved in metabolism or genetic information processing. This cgMLST scheme provides a robust, high-resolution framework for elucidating the population structure of this clinically important species.

2.1. Introduction

Haemophilus influenzae is a fastidious Gram-negative commensal coccobacillus that can act as an opportunistic human pathogen. Classification is determined by capsular polysaccharide expression, producing six serotypes (a–f) and unencapsulated nontypeable strains (NTHi) lacking a capsule. This species is confined to humans, predominantly colonising the upper respiratory tract as part of the normal microbiota, yet it can cause both non-invasive and invasive infections [1, 2]. Before introduction of the Hib polysaccharide-conjugate vaccine, serotype b accounted for most invasive H. influenzae disease worldwide; at the time of writing (February 2024), over 70 percent of cases are attributed to NTHi [5, 6], and incidence has risen across all age groups in recent years [5, 22, 105]. Invasive NTHi isolates have been reported to demonstrate increasing resistance to β-lactams (ampicillin and cephalosporins), fluoroquinolones and macrolides [79, 106, 107]. On the other hand, no vaccines targeting non-Hib strains have been developed in the past five years [108-111]. Although surveillance programmes incorporating whole-genome sequencing (WGS) exist in many high-income countries [112], standardised high-resolution typing schemes and nomenclature for key lineages remain lacking.

The molecular classification of human pathogens is an integral component of microbiological diagnostics and surveillance. A prevalent approach is multilocus sequence typing (MLST), which examines variation in six to eight housekeeping gene fragments to assign isolates to sequence types (STs). STs can be grouped into clonal complexes (CCs) according to their allelic profiles [47, 48]. Core genome MLST (cgMLST) expands this scheme to include hundreds of core genes conserved across a species, enabling finer-scale resolution of genomic variation [113, 114]. As with standard MLST, cgMLST considers each allelic

difference as a single event to reduce recombination-driven branch length inflation [63]. While MLST remains widely applied in the absence of WGS data, cgMLST offers enhanced lineage delineation and mitigates recombination effects [115]. Consequently, it has been recommended for highly recombinogenic bacteria, including H. influenzae [116, 117]. Indeed, NTHi demonstrates extensive genomic diversity largely driven by horizontal gene transfer (HGT) and recombination [95].

Earlier analyses of *H. influenzae*'s core genome has informed the creation of cgMLST schemes, yet these efforts have typically relied on fewer than 500 draft genomes, a sample size necessary for robust detection of paralogous loci and accurate core-genome analysis [53-55]. Paralogues, which derive from gene duplication, can obscure true genetic relationships among isolates [56, 57]. Furthermore, none of the existing cgMLST schemes has been both rigorously validated and made publicly accessible. Previous work has identified profile completeness as the most critical in silico parameter influencing cgMLST accuracy [118]. The BIGSdb software that hosts the PubMLST website routinely scans for new alleles within defined schemes (including cgMLST), deposits them in the sequence-definition (seqdef) database with assigned allele numbers, and thus defines complete cgMLST profiles [50]. To date, PubMLST hosts cgMLST schemes for numerous pathogens of public-health importance including Neisseria meningitidis [119], Neisseria gonorrhoeae [120], Campylobacter sp [121], Streptococcus agalactiae, S. pneumoniae [122], S. uberis [123], Vibrio cholera [124], V. parahaemolyticus [125], Bacillus anthracis [126], Bacillus cereus [127], Burkholderia mallei [128], Acinetobacter baumannii [129], and Clostridium perfringens [130]. This underscored the value of high-resolution schemes in microbial surveillance.

A cgMLST scheme for *H. influenzae* was developed and validated in this chapter and has since been implemented in the PubMLST database. This publicly accessible system can be utilised by public-health authorities worldwide to characterise *H. influenzae* genomes in a standardised, high-resolution manner. By reducing the bias introduced by frequent recombination events, it enhances understanding of the species population biology, particularly for NTHi. In turn, it may support vaccine development and contextualise the spread of antimicrobial resistance.

2.2. Method

2.2.1. Choosing pangenome analysis tools and dataset compilation

Reference genomes were used to develop a computational pipeline combining several pangenome analysis software packages

Several open-source pangenome analysis software packages were available at the time of analyses (May 2023), including Roary [131], PIRATE [132], PanX [133], PGAP [134], PPanGGOLiN [135], MetaPGN [136], PEPPAN [137], chewBBACA [55], and Panaroo [138]. Software not updated within the last five years was excluded from this analysis. Roary and PPanGGOLiN were omitted because a recent report found that over-splitting of paralogues in these programmes inflated accessory genome estimates and underestimated the number of core genes [138]. Therefore, four packages, PIRATE, PEPPAN, chewBBACA, and Panaroo were employed. A 'two-step' strategy was implemented to identify the optimal pipeline for *H. influenzae* core-gene detection. First, the methods and unique features of each tool were evaluated (Table 2.4); second, pangenome analyses were conducted on a reference genome set using each package, and the outputs were compared. The complete workflow and decision-

making that followed for each step are published in dx.doi.org/10.17504/protocols.io.4r3l22o64l1y/v1.

Table 2.4. Methods and features of four different pangenome and core genome reconstruction tools.

Tools	Description	Input files	Paralogs detection	Pseudogene detection	References
PIRATE	Rapid investigation of pangenome by classifying orthologous gene families with varied sequence similarity threshold.	Annotation files of draft genomes (.gff) from Prokka	Sequence clustering with CD-HIT and iterative MCL where clusters containing > 1 sequence per genome will go through paralogs classification step, which differentiate fission vs duplicated genes.	None	[132]
PEPPAN	Pipeline for pangenome construction which is based on similarity prediction, phylogenetic tree, and synteny.	Fasta and annotation files (.gff, from any source) of draft genomes	Combining tree-based (single BLAST hit, neighbour-joining, or maximum-likelihood algorithm) and synteny-based approach to detect paralogs.	Yes[a]	[137]
chewBBACA	A pipeline with multiple functions to support wg/cgMLST creation and validation which relies on the BSR-based allele calling for validating allele variation.	• Fasta files of draft genomes. • Prodigal training file for bacterial species.	During BSR-based allele calling, each loci is given paralog count (PC) number, which equals to how many times a matching CDS to that loci also matches with other loci.	None	[55]
Panaroo	A graph-based clustering tool which corrects annotation error by taking into account information provided by each genome.	Annotation files of draft genomes (.gff)	Sequence clustering at high threshold (98%) with CD-HIT. Clusters with >1 occurrence will be temporarily classified as putative paralogs. The graphical representation of each paralog cluster are assessed using the global context of the graph.	None	[138]

[a]*Pseudogene is defined as a CDS in genome with a significantly shorter length than other orthologous genes (default setting 0.8, equal or more to this threshold the gene is considered intact).*

[b]*Pseudogene identification can be done by conducting a manual curation.*

Abbreviation: wgMLST: whole-genome multilocus sequence typing; cgMLST: core genome multilocus sequence typing; MCL: Markov Cluster Algorithm; CDS: coding sequence; BSR: BLAST score ratio

Development and validation datasets of high-quality draft genomes

Draft genomes were retrieved from the PubMLST database on 24 September 2022 to assemble the study dataset. All publicly available *H. influenzae* entries were subjected to quality assessment using criteria outlined in Table 2.5 [139]. Genomes failing to meet these standards were individually reviewed for potential inclusion. Following this evaluation, 2,397 isolates were selected and their provenance data recorded.

Table 2.5. Parameters for assessing quality of *H. influenzae* draft genome assemblies.

Parameters	Threshold	Rationale for threshold	Reference(s)
rMLST score	≥ 85%	A low score for rMLST means uncertainty in species detection.	[140]
MLST and rMLST allele designation	One allele per loci	MLST and rMLST loci with multiple alleles suggest impure sequenced isolates and/or poor-quality data.	[140]
Genome length	1.4 Mb – 2.4 Mb	Range of Hinf genome length. A value outside this range may indicate contamination or incorrect species.	[95, 141, 142]
GC content	37 – 41mol%	Whole-genome GC content is constant for the same species hence deviation from this range denotes possible contamination.	[143]
Number of contigs	≤ 500	Lower number of contigs reflect a good-quality sequencing process and genome assembly.	[144]

Genomes were divided into a development dataset (N = 986) and a validation dataset (N = 1,411). The development dataset comprised isolates with well-documented provenance (capsule genotype, source, disease and region; Figure 2.1), ensuring reliable automated annotation in PubMLST, while the validation set included records with less complete metadata for workflow testing. In Step 1, genomes were classified by source and disease into seven groups (blood/bacteraemia, CSF/meningitis, blood/other invasive, any/pneumonia,

any/carriage, any/non-invasive, missing), and half of each group (excluding those with missing data) were randomly selected (n = 536). Step 2 addressed the type b and NTHi bias by proportionally adding non-b, non-NTHi capsulated genomes not already chosen, raising the total to 733 and revealing a European geographical skew. In Step 3, further genomes from non-European regions were added to approximate the overall dataset's continental distribution. The process yielded 986 genomes in the final development dataset.

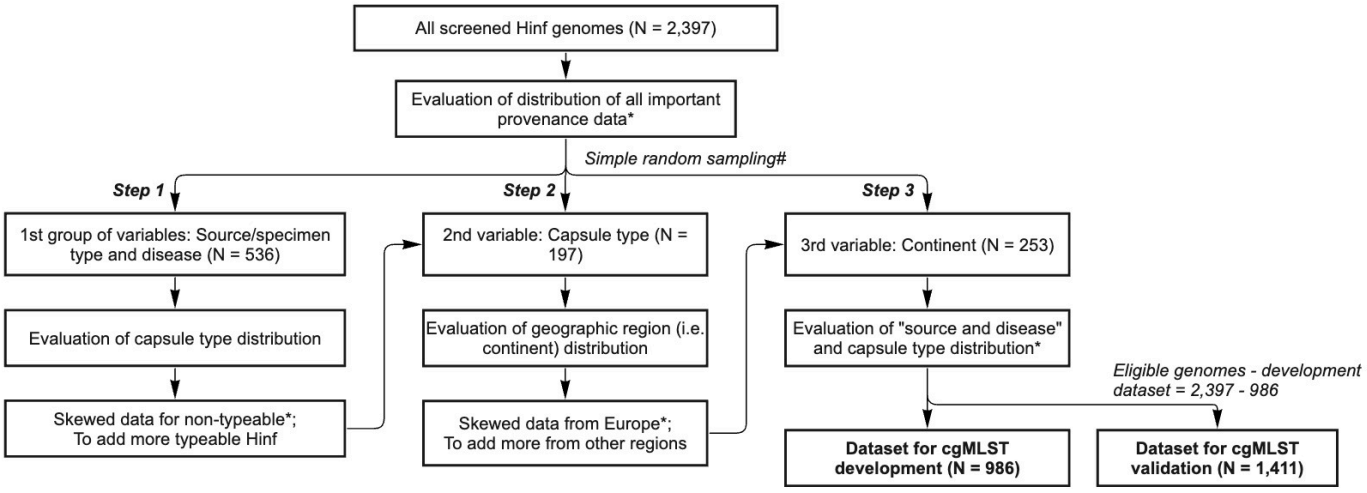


Figure 2.1. Allocation of high-quality H. influenzae draft genome assemblies into development and validation dataset. Note: *Importance provenance data are source/specimen type, disease, serotype based on genotype (i.e. capsule type), and geographic location (i.e. continent). #Steps for selection were done consecutively. For each step, samples were selected through a simple random sampling method from the pool of 2,397 Hinf genomes while considering that any selected sample from the previous step would not be included.

2.2.2. Core gene identification and curation

Draft genome assemblies in both datasets were annotated for protein-coding genes using Prokka [145]. For the scheme generation, core genes were defined as those present in at least 95% of isolates, to accommodate the draft status of the genomes and the occasional absence of genes in the original isolates [113, 115, 119]. PIRATE was used to identify core

genes by running BLASTP at amino-acid identity thresholds of 50, 60, 70, 80, 90, 95 and 98%. It groups genes into families and retains those present in a single copy in at least 95% of isolates, thereby excluding paralogues. Subsequently, the following steps were employed: (1) a confirmation that these provisional core genes were also classified as core by Panaroo, which includes steps to correct annotation errors, and (2) evaluation whether PEPPAN or Panaroo flagged any of them as paralogues, to ensure their removal from the final cgMLST scheme (Figure 2.2). This stringent exclusion of paralogues supports fully automated annotation of H. influenzae core genomes in PubMLST. The outputs of all three tools were combined using an in-house Python script available at https://github.com/artmisk13/cgmlst_hinf.

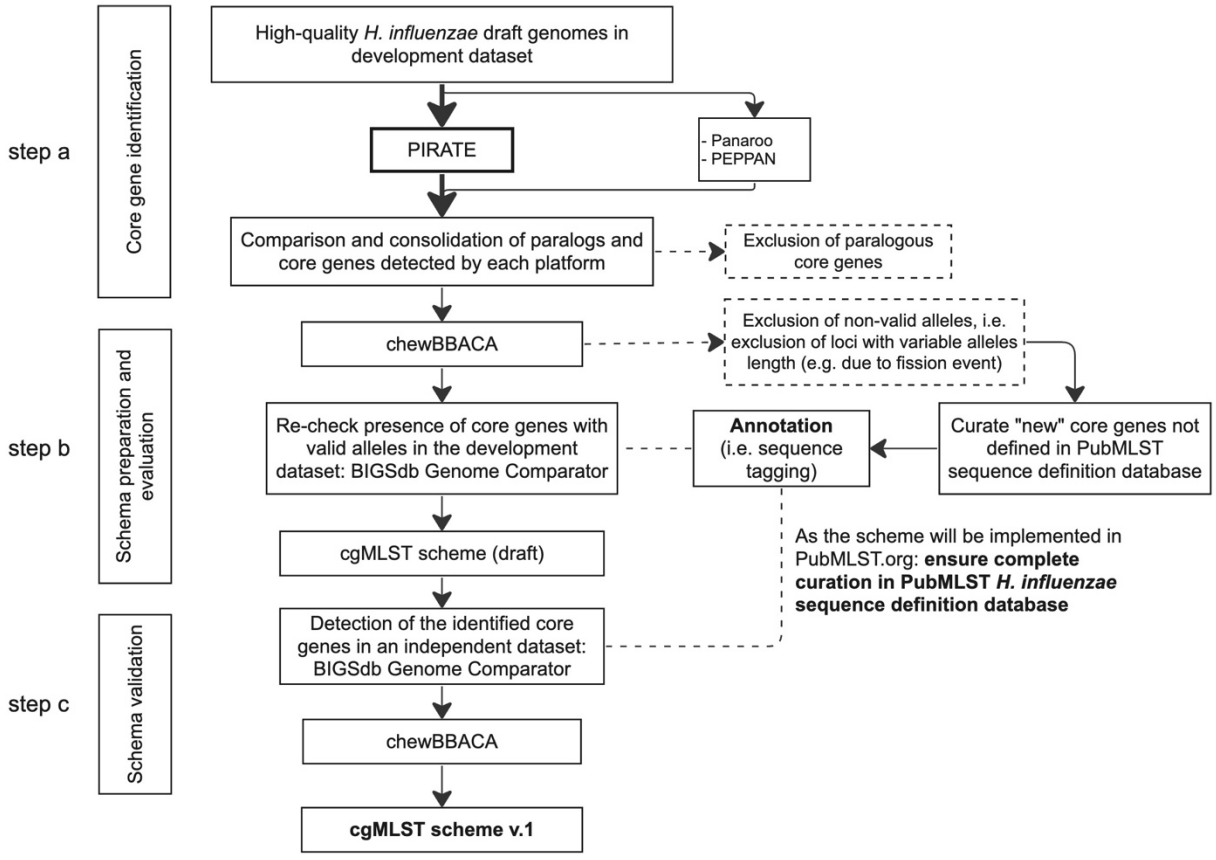


Figure 2.2. The workflow of cgMLST scheme development and validation.

Each core gene from step a (Figure 2.2) was screened for invalid alleles with chewBBACA using default settings (Figure 2.2, step b). An allele was deemed invalid if it contained ambiguous characters; had a length not divisible by three; included in-frame stop codons; or lacked start or stop codons [55]. Genes confirmed as valid were then compared against the PubMLST database to identify any loci not yet defined, allowing their curation. To detect novel alleles automatically, BIGSdb's sequence-tagging function was applied (minimum 90% identity, 70% coverage, BLASTN word size 20) [49]. In PubMLST, each H. influenzae locus receives a unique 'HAEM' prefix plus a numeric identifier. For example, the capsule-transport gene bexA is catalogued as HAEM1156.

2.2.3. Validation analyses

The preliminary cgMLST scheme was tested against the validation dataset to verify that the core genes remained core across an independent data set. This verification was carried out using the BIGSdb Genome Comparator tool in PubMLST with default settings. Genes that continued to meet the core definition following this process formed the final *H. influenzae* cgMLST scheme. Subsequent analyses on these loci included assessment of allelic variation, functional categorisation and intragenic recombination.

Allelic variability of the core genes was assessed by calculating the total allele count and allele lengths using an in-house Python script (https://github.com/artmisk13/cgmlst_hinf). Functional annotation was performed with eggnog-mapper v2.1.11, using Diamond in blastx mode and HMMER under default settings [146]. Each gene received a Cluster of Orthologous Genes (COG) designation [147], which was subsequently grouped according to the KEGG BRITE functional hierarchy[148]. Intragenic recombination was evaluated by the pairwise

homoplasy index (PHI) method implemented in PhiPack. PHI measures recombination via incompatibility between each alignment site and its downstream sites, and the significance is determined by comparing the observed PHI statistic to a normal distribution expected mean and variance to generate a p-value [149, 150].

2.2.4. Phylogenetic analysis

A maximum-likelihood (ML) tree was generated based on the core genome nucleotide alignment of 1376 genomes in the validation dataset using RAxML (version 8) and ClonalFrameML, which accounts for recombination events [100, 151]. The tree was then annotated with the set of metadata detailed below, along with the method(s) employed for their retrieval:

1. **Core genome cluster (CGC) at 500, 200 and 50 allelic mismatches**. Each genome was assigned a core genome sequence type (cgST) based on its allelic profile across the 1 037 core genes. To accommodate draft genomes, up to 25 missing loci were permitted for cgST assignment. Genomes with defined cgSTs were then organised into core genome clusters (CGCs) using single-linkage clustering at thresholds of 500, 200 and 50 allelic mismatches, ensuring that each isolate in a cluster differs by no more than the specified number of alleles from at least one other member.
2. **Capsule type based on genome sequences** (i.e. capsule genotype) as assigned by the Hicap suite software [152].
3. **CCs** were defined using the globally optimized eBURST (goeBURST) algorithm implemented in PHYLOViZ 2.0.

4. **Pathotype clade classification system for NTHi**. Earlier analyses of NTHi population structure delineated six clades (I–VI) based on the presence or absence of 17 accessory loci (Supplementary Table 2.1), a grouping that mirrored core-genome single nucleotide polymorphisms (SNPs) phylogeny [54, 60, 61]. These loci were identified by BLASTN searches against the genome dataset, using thresholds of ≥90% identity, ≥70% alignment and a word size of 20.
5. **Biotype** based on the presence or absence of genes (Supplementary Table 2.2) encoding three metabolic enzymes: ornithine decarboxylase (ODC), urease, and tryptophanase [23, 41].

Additionally, a minimum-spanning tree (MST) based on the core genome allelic profile was constructed, utilizing the GrapeTree plugin on the PubMLST isolate database [49, 153].

2.2.5. Relationships between the cgMLST scheme pairwise allelic mismatch and ML tree branch length

Genetic relatedness among *H. influenzae* isolates was assessed using two core-genome metrics: pairwise allelic mismatches from the cgMLST scheme and branch-length distances from a maximum-likelihood tree. Distance matrices were generated with the PubMLST Genome Comparator plug-in and a custom Python script (https://github.com/artmisk13/cgmlst_hinf), then converted into frequency tables for plotting. Scatter plots were created to visualise the relationship, and where distributions departed from normality, Pearson's correlation test was applied to calculate a P-value. Ordinary least-squares regression was then used to derive a best-fit line and its coefficient of

determination (R^2). All statistical analyses were conducted in R and with Python's statistics libraries; scripts are available at https://github.com/artmisk13/cgmlst_hinf.

2.3. Results

2.3.1. *H. influenzae* genomes from PubMLST database used for cgMLST scheme development and validation.

The reference genome dataset comprised fourteen complete *H. influenzae* genomes, eleven of which were retrieved from the KEGG Organisms: Complete Genomes database (accessed 26 September 2022). Three further clinically important complete genomes were identified through an National Center for Biotechnology Information (NCBI) PubMed search [154, 155]. All associated provenance metadata and genome sequences are accessible via PubMLST (Supplementary Table 2.3).

Of the 986 draft genomes in the development dataset that met the quality-check criteria, 65 were excluded after an initial pangenome analysis with PIRATE, reducing the total to 921. High duplication events were observed in the excluded genomes, which corresponded to outliers in genome length (Appendix 2.1) and were likely to reflect sequencing errors or lower assembly quality. The pangenome analysis was therefore rerun on the remaining 921 genomes, resulting in the identification of 1,063 core genes for the preliminary cgMLST scheme. Next, the 1,411 genomes in the validation dataset were annotated for these core loci, and 35 genomes with fewer than 95% of core genes called were excluded, yielding a final validation set of 1 376 genomes. No significant changes were noted in the distribution of other variables between the initial and final datasets (Supplementary Table 2.4).

Isolates were recovered from across the globe, with the majority originating in North America (356/921 and 548/1376) and Europe (431/921 and 547/1376). Both datasets showed comparable temporal distributions, with most samples collected after 2016. The development dataset contained a higher proportion of clinical invasive isolates (bacteraemia, meningitis and other invasive presentations), reflected by similar counts of non-typeable (542/921) and typeable (379/986) strains, thereby ensuring the scheme's applicability to both NTHi and encapsulated variants (Figure 2.3 (i)–(iv)). Moreover, the validation set introduced 242 (61%) STs and seven (11%) CCs not observed in the development cohort (Supplementary Table 2.4), underscoring its greater genetic diversity. Assembly quality was high in both datasets, with most genomes (893/921 and 1346/1376) featuring fewer than 150 contigs and an N50 above 23,000.

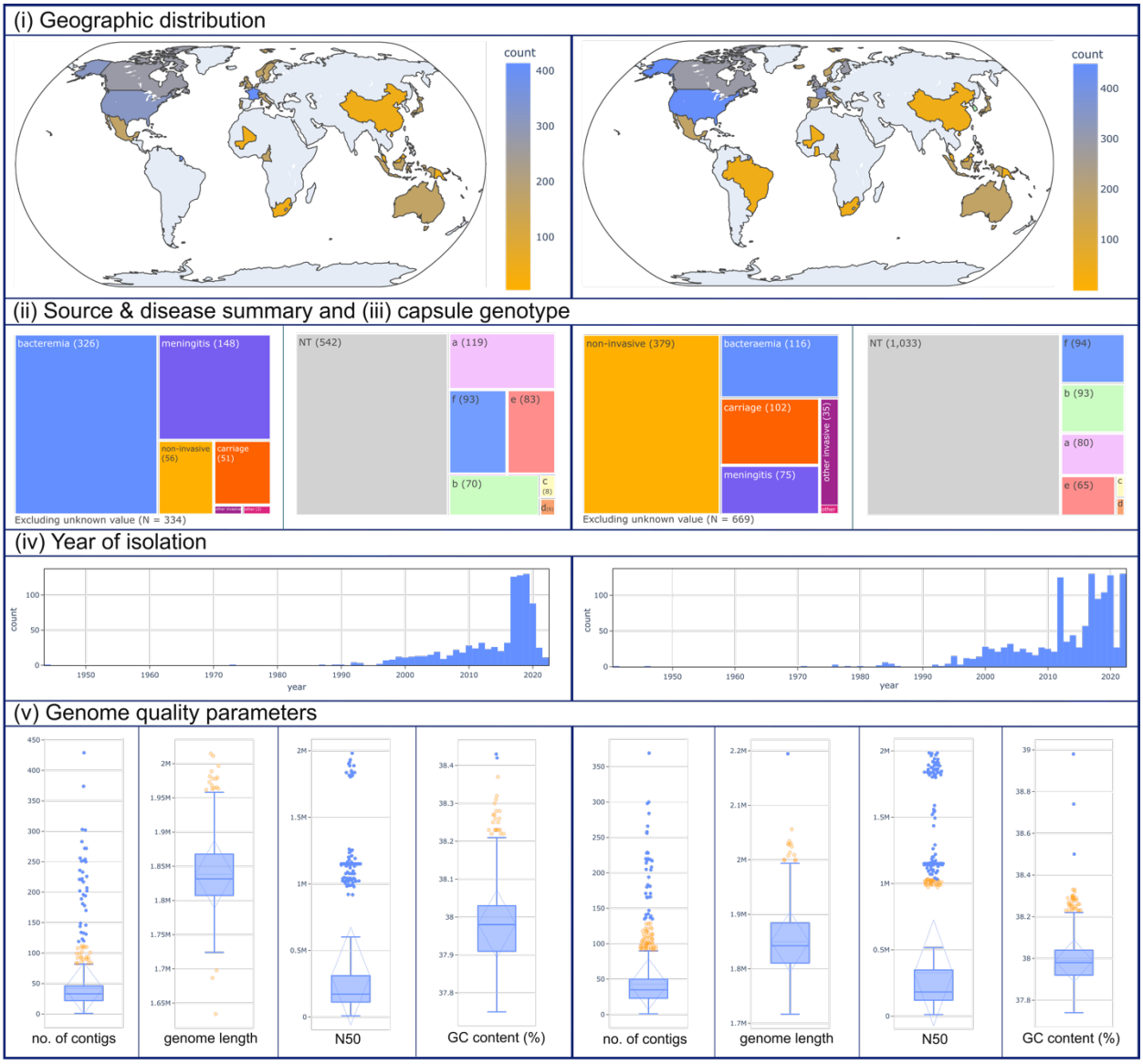


Figure 2.3. Characteristics of the datasets and genomes employed for developing the *Haemophilus influenzae* cgMLST scheme: (a) development (N = 921) and (b) validation (N = 1,376). Reproduced from [156].

2.3.2. One thousand and thirty seven core genes in the validated cgMLST scheme are implicated in important cellular pathways.

The combined pangenome analysis using PIRATE, Panaroo and PEPPAN (Figure 2.2, step a), initially identified 1,392 core genes, of which 144 paralogues were removed. Of the

remaining 1,248 loci, 18 were absent from the PubMLST H. influenzae seq-def database and were subsequently added. After filtering out invalid alleles (Figure 2.2, step b), 185 genes failed to meet the core criterion, leaving 1,063 genes in the draft cgMLST scheme (Supplementary Tables 2.5.1–2.5.5). Validation against the independent dataset (Figure 2.2, step c), further refined the core set to 1,037 genes by excluding 26 loci, 25 of which were present in 90–95% of genomes (Supplementary Tables 2.5.6–2.5.7). All draft genomes were successfully annotated for these core genes using PubMLST's automated annotation process.

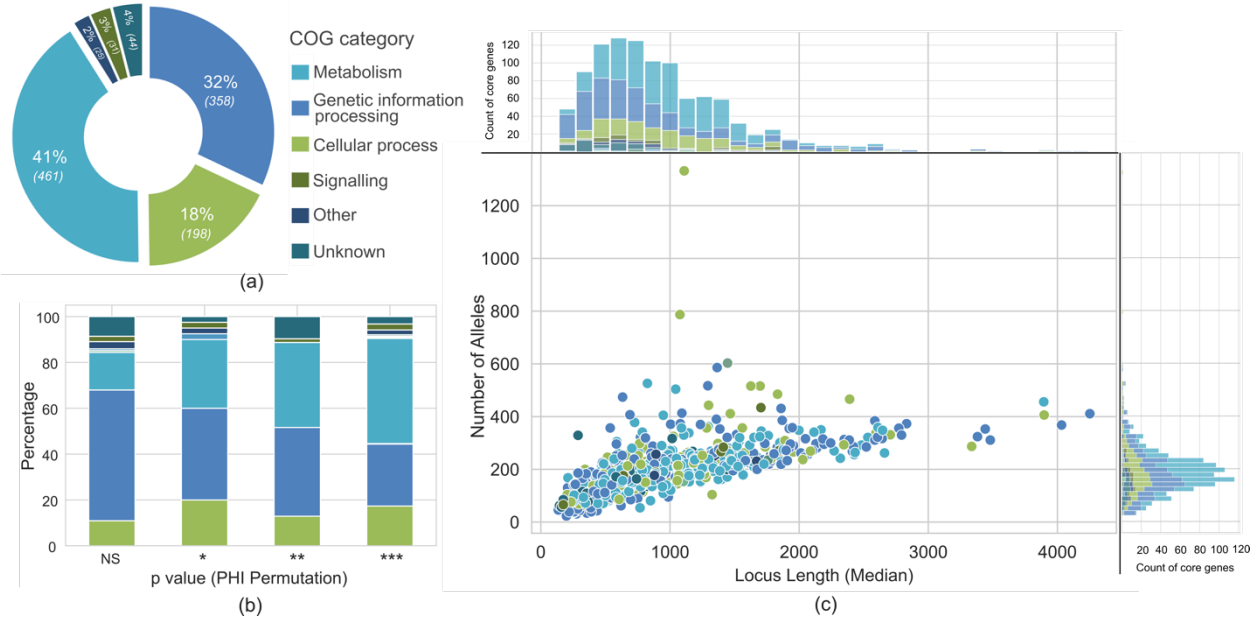


Figure 2.4. Functional classification, recombination, and allele variability analysis of *Haemophilus influenzae* core genes in the cgMLST scheme. (a) Functional classification was achieved with eggnog-mapper, assigning the COG category for each core gene. (b) Intragenic recombination analysis of each core gene based on PHI permutation P-value. Core genes were grouped based on this P-value: NS, non-significant; *0 .01 < p-value < 0.05; ** 0.001 < p-value < 0.01; *** p-value < 0.001. The proportion of each COG category within each p-value group was calculated and shown as the coloured stacks in the bar graph. (c) The number of alleles and their length variation were counted for each core gene. The median locus length and the allele count were plotted and coloured based on the COG category. The upper panel showed the distribution of median locus length and on the right panel, the distribution of allele counts; both were coloured in accordance with the COG category. Reproduced from [156].

In total, 1,024 core genes were mapped to NCBI COG categories, with 37 assigned to more than one category [147]. Of those linked to a single COG, 149 (14%) were initially classified as 'function unknown' (COG S). Upon re-evaluation, 95 of these were reassigned to at least one specific non-S COG based on orthologue hits at higher taxonomic levels. The remaining genes were manually curated using the KEGG orthology hierarchy and/or protein-product searches in the Protein Families Database v95.0 (Supplementary Tables 2.6.1 and 2.6.2). The 13 genes with no initial COG assignment were re-analysed via HMMER without change and were therefore retained in the 'function unknown' (S) category (Figure 2.4a).

Approximately 90% (909/1 037) of loci in the cgMLST scheme exhibited intragenic recombination (PHI p-value < 0.05). When stratified by COG category, most genes in the genetic information processing group (73/128) did not show significant PHI statistics, whereas almost all metabolic genes (405/461) did (Figure 2.4b). Additionally, loci with non-significant PHI statistics were generally shorter and had fewer alleles compared with those showing evidence of recombination (Appendix 2.2).

Allele counts and median locus lengths were positively correlated, though the slope was modest (Figure 2.4). 65% (675/1 037) of core genes had median lengths under 1,000 nucleotides, and 95% (986/1 037) were under 2,000 nucleotides; these sets exhibited median allele counts of 151 and 175, respectively. Notably, some shorter loci displayed exceptional diversity, for instance, HAEM0191 (hypothetical protein) and HAEM1295 (outer membrane protein P5) had 1,332 and 787 alleles, respectively, outstripping other genes of similar length.

2.3.3. Clustering groups of *H. influenzae* genomes based on pairwise allelic mismatches of the core genes reflected their phylogenetic relationship.

Within the validation dataset, 1,320 (95.9%) were allocated cgSTs and subsequently classified into CGCs at the established similarity thresholds (Supplementary Table 2.4.1). The remaining genomes, while lacking cgST assignments, nonetheless carried allelic profiles for 989 to 1,011 out of the 1,037 core loci (95.4–97.5%).

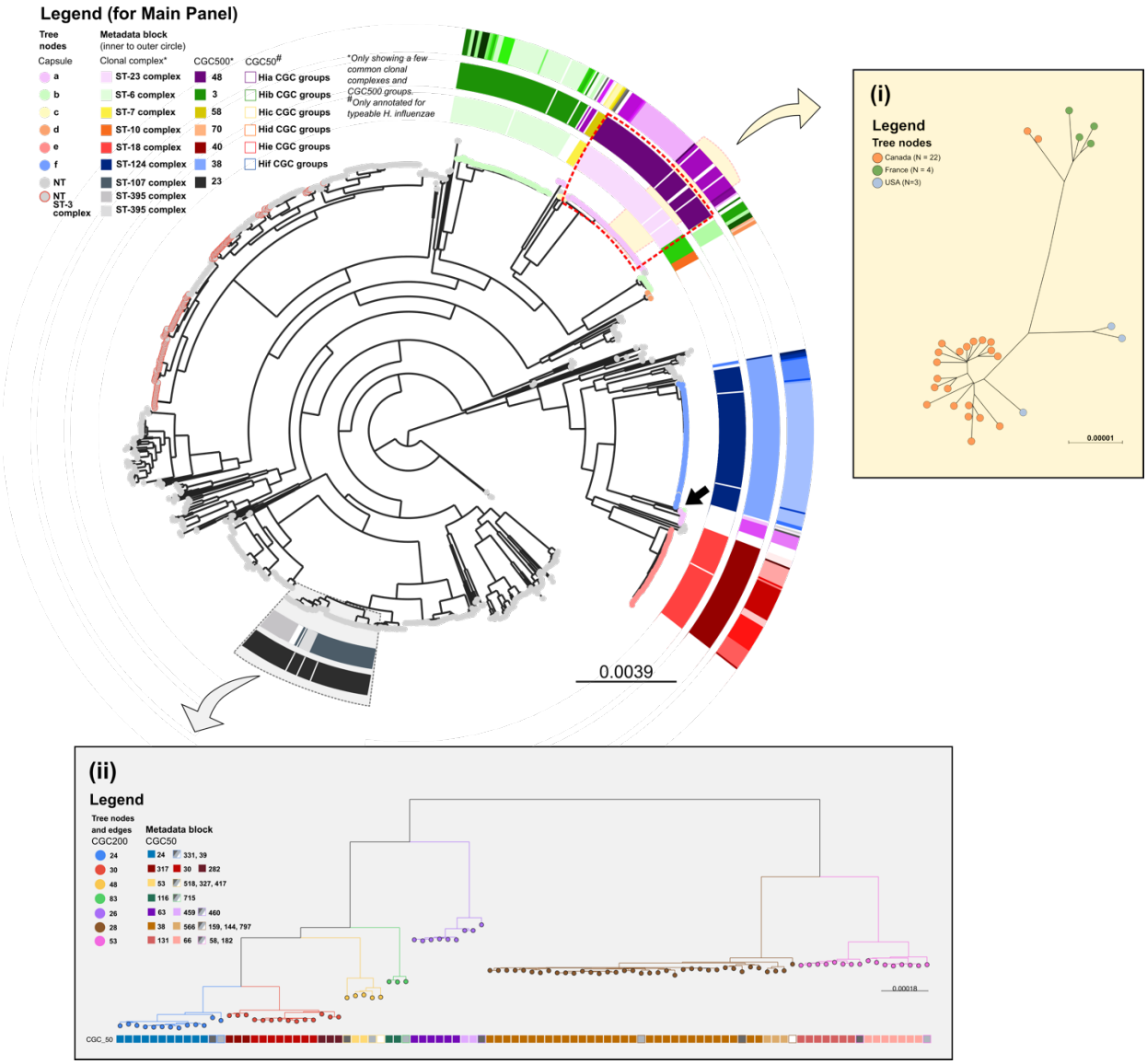


Figure 2.5. Population structure of 1,376 *H. influenzae* genomes from the validation dataset. **Main panel:** a ML phylogenetic tree generated from a concatenated core gene nucleotide sequence alignment. Tree nodes were coloured by capsule type. Each capsule type clustered together, except for *H. influenzae* type a (Hia) and type b (Hib). The innermost metadata block was a CC assignment, which corresponded well with the middle metadata block, representative of the CGC group at the 500 allelic mismatches threshold (CGC500). This correlation was evident for capsulated/typeable *H. influenzae*, but not for NTHi. The outermost metadata block was the CGC group at the 50 allelic mismatches threshold (CGC50), which allowed a more granular distinction of *H. influenzae* clusters. The hierarchal clustering at multiple thresholds was able to reflect the structure of the phylogeny. **Subpanel (i):** A subset of the ML tree in yellow highlight, consisting of 29 Hia isolates CGC50 group 15. Isolates in this group were predominantly from North America, with 20 from Canada clustering closely, a pattern reported previously by Topaz et al [64]. **Subpanel (ii):** The hierarchical structure of the ML tree was also represented using CGC based on multiple thresholds for NTHi. For example, a subset of the ML tree highlighted in grey comprises 97 NTHi within the CGC500 23. Tree edges and nodes were coloured based on the CGC200 groups, which are congruent with the tree topology. The metadata block shows CGC50 groups within each CGC200 group. The CGC50 groups with greyscale colour were singletons within the corresponding CGC200 group. Reproduced from [156].

The population structure of *H. influenzae* can be systematically explored through hierarchical core-genome clustering at multiple thresholds. This classification, derived from core-genome allelic profiles, mirrors the phylogenetic relationships displayed on the ML tree **(Figure 2.5)**. The same clustering pattern is also reproduced on the minimum-spanning tree constructed solely from core-genome allelic differences **(Figure 2.6)**.

Isolates of encapsulated *H. influenzae* sharing identical capsular genotypes were observed to cluster phylogenetically, reflecting their genetic proximity. Core genome clustering groups (CGCs) at multiple allelic-difference thresholds corresponded to individual serotypes (Figure 2.5). For type a (Hia), four principal lineages were resolved, each represented by a distinct CGC500 (numbers 48, 46, 79 and 81), which matched their MLST clonal complex assignments (Figure 2.6 and Supplementary Table 2.7). The most populous Hia lineage, CGC500 48 (ST-23 complex), contained a prominent Canadian subgroup that further coalesced at the 50-allele threshold (Figure 2.5 yellow highlight and subpanel i). A similar structure was evident in serotype b (Hib), where three lineages appeared in both the maximum-likelihood and minimum-spanning trees; two of these aligned with CGC500 clusters and their corresponding CCs (Supplementary Table 2.7). The rarest Hib lineage comprised a single ST-464 isolate (ID 5105) lacking any CC or CGC assignment (Figure 2.5 and **Figure 2.6**, black arrow). As of October 2023, only one additional ST-464 genome (ID 15982) was present in PubMLST; this isolate did not group with other Hib but was more closely related to Hia based on its allelic profile.

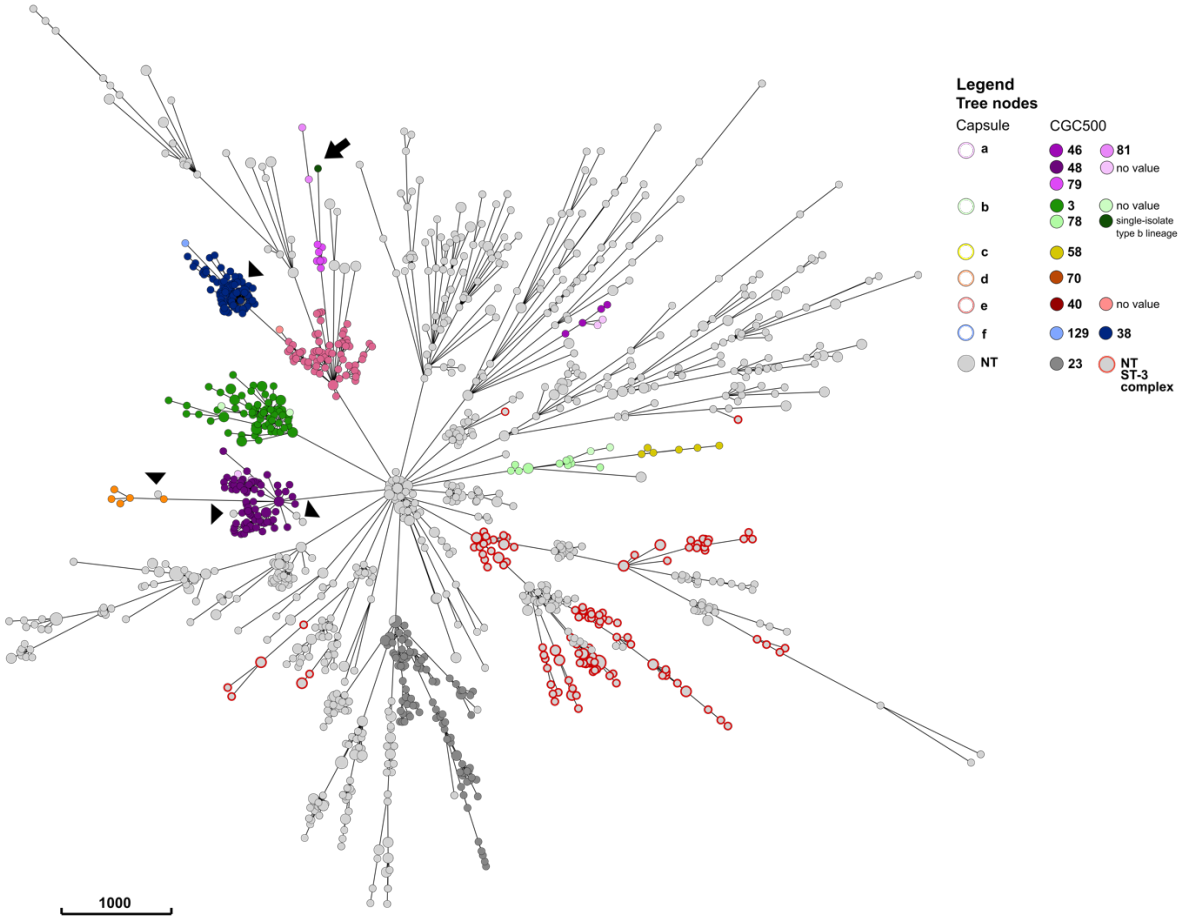


Figure 2.6. MST of the validation dataset (N = 1376 genomes) constructed from the core genome allelic profile. Tree nodes were coloured by capsule type and CGC group at the 500 allelic mismatches threshold (CGC500), using the same colour scheme as **Figure 2.5**. The pattern observed in the phylogeny was replicated in this MST. Isolates belonging to the same capsule type and CGC500 clustered together, e.g. four main lineages for Hia (purple) and three for Hib (green), with one lineage consisting of a single isolate (black arrow). The scattered distribution of NTHi ST-3 complex isolates were also replicated (grey with a red outline). Additionally, 5 NTHi isolates found within the encapsulated clusters were shown in greater detail here (black arrow heads). Reproduced from [156].

NTHi isolates exhibited far greater heterogeneity, forming 110 CGC500 clusters, of which only eight contained at least 30 genomes (Appendix 2.3). Nevertheless, core genome clustering corresponded closely to phylogenetic relationships within NTHi. For example, 97 isolates assigned to CGC500 23 were highlighted on the ML tree (Figure 2.5 grey highlight

and subpanel (ii) and Figure 2.6) , and these partitioned into seven CGC200 subclusters that mirrored the tree topology. In contrast to encapsulated lineages, NTHi clonal complexes did not consistently map onto the phylogeny; the ST-3 complex, for instance, is scattered throughout the tree (Figure 2.5 and Figure 2.6, grey nodes with red outline). This pattern, seen in 19 of 57 CCs in the validation set (Supplementary Table 2.7), demonstrates that CGC groupings provide a more accurate representation of NTHi population structure **(Figure 2.6)**.

Although each serotype typically corresponds to specific MLST CCs, exceptions were noted: the ST-124, ST-210 and ST-422 complexes each contained both NTHi and type f isolates. In contrast, their CGC500 assignments matched their serotype and phylogenetic placement, underscoring cgMLST's advantage over seven-locus MLST. Notably, five NTHi isolates fell within encapsulated CGC500 clusters (Figure 2.5 arrow heads and Supplementary Table 2.7), indicative of rare capsule-loss events in previously encapsulated strains.

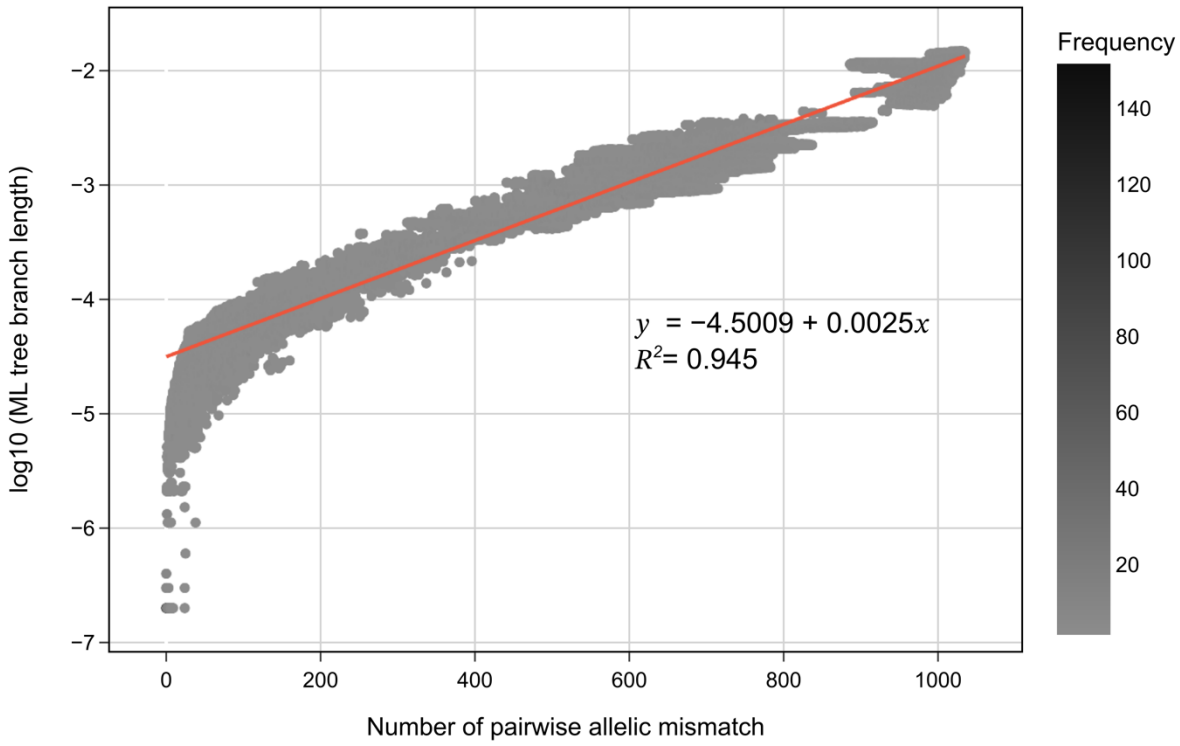


Figure 2.7. Comparison of pairwise allelic mismatch of cgMLST core genes with the log10 branch length values from the ML tree, implemented in the validation dataset (N = 1,376). The more closely related genomes in a pair, the lower the allelic mismatch and log10 branch length value found. The log10 branch length as a function of pairwise allelic mismatch was calculated using the OLS method and the adjusted R^2, a coefficient of determination of the defined function, was also measured. The R^2 value is close to 1, which indicates a strong correlation between allelic mismatches and the branch length. Reproduced from [156].

The concordance between core-genome allelic distances and phylogenetic branch lengths was quantified (**Figure 2.7**). For each genome pair in the validation dataset, the number of allelic mismatches was plotted against the corresponding branch-length value from the recombination-corrected ML tree. An R^2 of 0.945 was obtained, indicating that 94.5% of the variance in branch lengths can be explained by allelic mismatches across the 1,037 core genes. Spearman's rank correlation also yielded a significant p-value (< 0.001). Greater scatter

of points near the origin reveals that very closely related genomes (differing by only a few nucleotides) may not be fully distinguished by allelic mismatch counts.

Finally, two previously employed classification schemes were evaluated for their concordance with both the cgMLST framework and the reconstructed phylogeny. The first, the NTHi-specific pathotype-clade system, was applied to the ML tree of the 1,008 NTHi genomes in the validation dataset (Appendix 2.3). Of these, 160 isolates could not be assigned to any of the six defined clades.

Although the clade classification partially mirrored the phylogenetic topology, intraclade genetic divergence remained substantial. For example, clade VI exhibited a median pairwise allelic mismatch of 991 (range 0–1 028), whereas clade II showed a median of 640 (range 0–766) (Appendix 2.4). We also assessed the biotype system, which is defined by the presence of three enzyme-encoding genes; none of the eight biotypes aligned consistently with CGC groups at any threshold, the phylogenetic structure, or any clinical or demographic variables (Appendix 2.5).

2.4. Discussion

At the time of writing (February 2024), the NCBI Genome Library holds over 560 000 prokaryotic assemblies (https://www.ncbi.nlm.nih.gov/genome/microbes/). The widespread use and decreasing cost of high-throughput sequencing have yielded an unprecedented volume of bacterial genome data [157]. cgMLST employs a gene-by-gene strategy on WGS data to characterise variation within a species or genus. Its advantages include independence from a reference genome, focus on protein-coding loci, and the treatment of each allelic change as a single event, particularly beneficial for highly recombinogenic taxa like *H.*

influenzae. By defining 1,037 core genes, the cgMLST scheme generates a high-resolution population structure that is minimally influenced by recombination.

Based on the published literature, estimates of the H. influenzae core genome varied. Hogg et al. and Eutsey et al. identified 1,450–1,485 'hard core' genes present in 100% of genomes [93, 94]. Pinto et al. reported 1,400 genes shared by at least 95% of over 200 NTHi genomes [54], while Rajendra KC et al. found only 853 core loci among 12,249 pan-genes in 568 NTHi genomes [158]. Gonzalez-Díaz and colleagues observed 1 470–1 627 core genes within each capsular group (10–234 genomes), yet when all 800 capsulated genomes were combined, only 1,037 genes were universally present [65]. Variations in pangenome analysis arise as core size typically declines before plateauing with increasing genome numbers [93], clustering parameters and thresholds influence orthologous group (OG) definitions, sequencing and assembly quality exert further effects [137, 138], and paralogues are not automatically excluded [55-57]. To mitigate these factors, the developed cgMLST scheme in this study incorporated strict genome-quality filters and multiple pangenome tools to enhance paralogue detection, yielding a robust typing scheme that reflects *H. influenzae* phylogeny.

Clustering by cgST aligned closely with the phylogeny inferred via maximum-likelihood analysis of core-genome nucleotide alignments. At 500 allelic mismatches, four H. influenzae type a lineages were recovered, matching those described by Topaz et al. [64], and their predominantly North American sublineage was likewise reproduced at the 50-allele threshold [64]. U.S. surveillance identified two main Hib lineages corresponding to distinct CGC500 groups [3], but a third cluster was indicated by a single ST 464 isolate (PubMLST ID 5105) outside these groups. Two principal Hif lineages also emerged at 500 mismatches, whereas Gonzalez-Díaz et al. reported only one and described three informal clades not seen here [65].

Such differences may arise from variation in population-structure frameworks (species-wide versus Hif-specific core genomes) and phylogenetic methods (core-genome alignment versus reference-based SNPs) [65]. Overall, CGC500 clusters largely corresponded to MLST clonal complexes, though some divergence was anticipated given the approximately 150-fold greater gene count in cgMLST compared to seven-locus MLST. Nevertheless, the observed concordance supports maintaining seven-gene MLST CC thresholds for broader relatedness assessments when WGS is unavailable.

Five NTHi isolates embedded within encapsulated clusters on the ML tree (i.e. sharing the same CGC500 assignments as encapsulated strains) likely represent rare capsule-loss events. The first invasive Hib isolate lacking capsule expression was documented in 2019 [159], and Potts et al. reported similar findings that year in a population-genetic survey of US surveillance isolates [3]. By employing cgMLST and cgST-based clustering, such events can now be detected directly without reconstructing phylogenies.

Lastly, the cgMLST scheme was evaluated against two established classification methods for *H. influenzae*. The first, biotyping, predates molecular techniques and is based on differential production of tryptophanase (indole), urease and ornithine decarboxylase [160]. Slotved et al. demonstrated that enzyme production can be inferred from detection of the corresponding genes, although enzyme profiles did not align with phylogenetic relationships, an observation confirmed here [23]. The second system, specific to NTHi, defines clades using core-genome SNPs, that is, the portions of the reference sequence (isolate 86-026NP, PubMLST ID 5068, NCBI RefSeq GCF_000012185.1) that could be aligned to all other genomes' [61, 158]. Previous studies assigned each NTHi isolate to one clade, with clades I, IV and V forming a monophyletic group [60, 61]; this pattern was not observed in the present work.

These findings indicate that biotyping and NTHi clade typing offer limited discrimination and precision in reflecting H. influenzae phylogeny [41]. In contrast, cgMLST employs a larger gene set in a gene-by-gene approach, yielding a high-resolution population structure free from reference bias.

2.5. Conclusion

In conclusion, the cgMLST scheme introduced here delivers a fine-scale view of H. influenzae population structure and offers a practical tool for microbiology reference laboratories and public health authorities to assess phylogenetic relationships. The assigned cgSTs provide a consistent genome-level nomenclature, representing an advance in typing methods, especially for NTHi, the primary cause of invasive disease. However, clustering genomes solely by single-linkage of core-genome allelic profiles fails to yield stable groupings. Integrating cgMLST with a robust clustering system, such as the Life Identification Number (LIN) code, can support both precise isolate identifiers and stable, reliable multi-order groupings [161]. Nevertheless, the comprehensive characterisation of the bacterial core genome afforded by this scheme serves as a valuable resource for enhancing molecular diagnostics and directing vaccine development.

Chapter 3

Resolving *Haemophilus influenzae* Molecular Epidemiology using Life Identification Number (LIN) Coding

Abstract

Haemophilus influenzae is an occasional human bacterial pathogen for which there is a pressing need for a stable genomic nomenclature for public health and research. We implemented the life identification number (LIN) method based on core genome profiles and implemented this scheme in the PubMLST *H. influenzae* database. Thirteen LIN code thresholds were established: eight for defining clusters from 'megalineage' to 'subclade' levels and five for shallow-level partitioning useful for outbreak investigation. This method assigns LIN codes to each *H. influenzae* genome functioning as a barcode which provides genetic relationship information, and supports public health surveillance, antimicrobial resistance monitoring, and vaccine development.

3.1. Introduction

Haemophilus influenzae is a Gram-negative, pleomorphic bacterium and an accidental human pathogen. Typically, this fastidious organism colonises the human nasopharynx asymptomatically; however, it has the potential to breach the mucosal barrier, resulting in non-invasive diseases such as conjunctivitis, otitis media, or pneumonia, and can invade the bloodstream, causing septicaemia and meningitis. Based on the expression of capsular polysaccharides, *H. influenzae* is categorised into one of seven groups: a, b, c, d, e, f, or

unencapsulated (nontypeable) [95, 162]. While the introduction of *H. influenzae* type b (Hib) conjugate vaccine has significantly reduced the incidence of invasive disease caused by Hib, the overall number of invasive diseases caused by *H. influenzae* has increased, with nontypeable *H. influenzae* (NTHi) now responsible for the majority of cases [20]. Reports from the UK Health Security Agency and the U.S. Centers for Disease Control and Prevention, spanning a 10-year period, indicate that 83% of invasive *H. influenzae cases* in the UK and 72% in the USA were caused by NTHi [21, 163]. Similarly, the Danish national surveillance program reported that between 2014 and 2022, the incidence of invasive NTHi infection was almost seven times higher than that of Hib [67]. Little is known, however, concerning factors determining NTHi invasiveness [164, 165] and vaccine development for NTHi remains limited [109, 111]. In addition, NTHi is increasingly resistant to many antibiotic classes, such as beta-lactams, including third-generation cephalosporins, macrolides, and fluoroquinolones [74, 166]. A total of 68.8% of NTHi invasive isolates collected between 2011-2020 in northern Taiwan were resistant to ampicillin, and 82% of them were resistant to more than three different antibiotic classes- [167]. In addition, 31% of NTHi causing non-invasive infections in paediatric population from South Korea were resistant to ceftriaxone [88]. Therefore, *H. influenzae* is an important pathogen to monitor, with most high-income countries integrating whole-genome sequencing (WGS) as a part of their surveillance programmes [32, 168, 169].

Over the past decade, genomic studies have suggested that *H. influenzae* exhibits some level of population structure [60, 65]. Understanding population structure is crucial for addressing key clinical and research questions, such as determining whether *H. influenzae* cases are part of a potential outbreak, identifying which lineage a given isolate or specimen belongs to, and designing vaccines that work across different capsule types. However, interpreting this population structure is challenging, requiring extensive analyses and

advanced bioinformatic skills. In an era of WGS-based molecular epidemiology, it is essential to have a consistent and high-resolution method for bacterial nomenclature and clustering – one that both reflects the population structure and is easily interpretable. Without such a system, arbitrary naming and grouping [170], such as the terms like 'international clone' [171] that lack genetic context, will persist. This practice complicates the comparison of important circulating isolates, whether for public health needs (e.g., potential outbreaks) or research objectives (e.g., comparative genomic analyses).

Previously, we developed and implemented the core genome multilocus sequence typing (cgMLST) scheme for *H. influenzae* [156]. cgMLST is an expansion of the multilocus sequence typing (MLST) approach, which classifies bacterial genomes into sequence types (STs) based on the allelic profile of a fragment of seven housekeeping genes. These STs can be further organised into clonal complexes (CCs) [47, 48]. The cgMLST scheme for *H. influenzae*, however, assigns a core genome ST (cgST) based on the allelic profiles of 1037 core genes, offering a much higher resolution than MLST [48, 156]. Single-linkage clustering at various arbitrary thresholds based on core genome pairwise allelic mismatches showed a high degree of congruence with the Maximum-Likelihood (ML) phylogeny [156]. Nevertheless, these clusters are unstable due to the potential merging of clusters as genomes with high similarity to different clusters are added [172, 173]. The life identification number (LIN) addresses this issue while integrating the previously developed cgMLST scheme within its framework [174].

The core principle of the LIN method involves assigning a series of numbers (referred to as "LIN code"), to individual genomes based on their similarity to other genomes in a population. This code operates like a barcode, capturing the uniqueness of the genetic makeup

of an individual genome and its relatedness to other genomes in the population at multiple levels [175]. Conceptually, this method can be implemented for any organism, including animals, bacteria [176] and viruses [177]. The original approach for measuring genome similarity in LIN code assignment relied on average nucleotide identity (ANI) [161, 175]. While ANI-based LIN code is effective for distinguishing bacterial isolates at the species level, it is less suited for strain-level differentiation. For example, in *Bacillus anthracis*, ANI-based LIN code can separate significant isolates at similarity thresholds ranging from 99.99% to 99.9999% ANI, but ANI values at these levels are difficult to interpret meaningfully [175]. Furthermore, in highly-recombining species such as *H. influenzae*, the number of core genes (i.e. genes shared by most or all isolates) is much smaller compared to the pan-genome size [54, 65], making whole-genome ANI less reliable, as it may overstate dissimilarities among isolates. Integrating the core genome concept through cgMLST scheme with the LIN method overcomes these limitations. cgMLST-based LINcodes are defined using the core genome pairwise allelic mismatches at various thresholds, which enables bacterial isolates to cluster at different levels of resolution [174]. This approach has been successfully applied for the *Klebsiella pneumoniae* species complex (Kpsc) [174] and *Streptococcus pneumoniae* [178].

This study (1) established LIN code thresholds for *H. influenzae* based on the previously developed cgMLST scheme, (2) integrated the LIN code system into the public PubMLST database (https://pubmlst.org/organisms/haemophilus-influenzae), and (3) compared LIN code groupings with existing clustering methods, creating a user-friendly nickname nomenclature for congruent clusters to translate the otherwise machine-readable codes into easily interpreted labels [179].

3.2. Method

3.2.1. Choosing datasets

Two independent datasets were used to define core genome LIN code partition levels. Dataset 1 consisted of high-quality, publicly available draft genomes (N = 2,297) from the PubMLST database as previously described [156] and was utilised to evaluate the genetic discontinuity present and identify putative partition levels with phylogenetic reconstruction used to validate each threshold. Dataset 2 consisted of a collection of 5569 core genome allelic profiles and cgSTs, collectively termed as "core genome profile", obtained from the cgMLST scheme of *H. influenzae* [156]. Profiles were defined automatically for every genome deposited in the PubMLST database (accessed 15th February 2024). These were used to validate statistical methods.

3.2.2. Identification of LIN code partition levels

There are two main groups of LIN code thresholds. The first group defines broader population structures, such as lineages and clonal groups. These thresholds were systematically identified and validated,. The second group consists of thresholds used for strain-level identification and is particularly useful for detecting outbreak clusters. These thresholds were based on previous studies of *H. influenzae* outbreaks [180, 181], allowing flexibility for adaptable cluster definitions [174]. Here, LIN code thresholds are denoted as "t" followed by the number of allelic mismatches, for example, t110 represents a threshold at 110 allelic mismatches, facilitating consistent genomic comparisons across different levels of genetic relatedness.

Distribution of pairwise allelic mismatches for genetic discontinuity observation

Initially, putative LIN code partition levels were identified based on the genetic discontinuity observed from the distribution of core genome pairwise allelic mismatches (PAM) using the draft genome dataset. PAM was generated as a distance matrix using the Genome Comparator plugin on PubMLST. In the pairwise comparison, missing allele designations were treated as allelic differences. As PAM reflects genetic dissimilarities, it can be translated to similarity level using the following formula:

$$Similarity\ level = \frac{N - PAM}{N} \times\ 100\%$$

PAM : core genome pairwise allelic mismatches

N : total number of core genes in a cgMLST scheme

A frequency table was subsequently generated using the PAM distance matrix by stacking it into a dataframe consisting of three columns (isolate 1, isolate 2, PAM) using pandas Python package. To determine quantitatively the partition level, k-means clustering using sklearn Python package was applied to the frequency table (https://github.com/artmisk13/DPhil_thesis_Hinf). A histogram was constructed from the frequency table and annotated by the k-means clustering result for visual evaluation. This process was repeated to several subsets of the draft genomes in the dataset, which included NTHi-specific CCs (ST-3 complex; ST-11 complex; ST-139 complex; ST-107 complex; and, ST-57 complex), two Hib-specific clonal complexes (ST-6 complex and ST-222 complex), one *H. influenzae* type a (Hia)-specific clonal complex (ST-23 complex), and one *H. influenzae* type f (Hif)-specific clonal complex (ST-124 complex). Within each capsule type, the CCs chosen were those contained the largest number of isolates.

3.2.3. Threshold validation using statistical testing

The *H. influenzae* core genome profile was clustered using a single-linkage hierarchal classification method across various thresholds as defined in the previous section. Subsequently, clustering results were assessed using several statistical metrics, including Silhouette score, adjusted Wallace (aW) coefficient, and normalised mutual information (NMI). Thresholds with low Silhouette score, indicating poor clustering specificity, or highly variable aW coefficients, suggesting sensitivity to data variation, were excluded. Furthermore, if two or more thresholds produced highly similar clustering results, as indicated by high NMI score, only one threshold was retained to eliminate redundancy.

1. **Silhouette score:** this measures how similar a member of a cluster is to other members of the same cluster compared to members of different clusters (i.e. cohesiveness). The silhouette score ranges from -1 to 1; a value close to 1 indicating a perfect fit, while 0 means presence of overlapping clusters and a negative value signifies a wrong assignment of data point(s), as a different cluster is more similar. An arbitrary threshold commonly set for silhouette score is 0.5, with a score > 0.5 suggesting a high-quality clustering [182, 183]. Silhouette scores were generated using MSTclust program (https://gitlab.pasteur.fr/GIPhy/MSTclust) [184].
2. **Adjusted Wallace coefficient (aW):** is a measure of agreement between clustering produced by two different methods [185]. In this study, aW was employed to indicate the stability of clustering following data perturbation. Therefore, for a given *H. influenzae* core genome allelic profile dataset, $aW_{a \to b}$ was the probability that two cgSTs clustered together after data perturbation (b)

if they have been classified together (a). The aW score was computed using MSTclust program (https://gitlab.pasteur.fr/GIPhy/MSTclust) [184]. Two values were given following the computation: aW1 and aW2, which take into account false positive and false negative pairwise agreements after data perturbation, respectively.

3. **Normalised mutual information (NMI):** is a measure of similarity between two different clusters by combining homogeneity and completeness aspects. The NMI statistic was used to compare clustering results based on two different thresholds, with a value close to one indicating a higher similarity of the clustering results [186, 187]. For this purpose, we included all possible thresholds, ranging from 0 to 1030 allelic mismatches. This step addresses the redundancy of using LIN code thresholds that are close to each other, which could result in the same cluster. The NMI is measured using the phierCC program (https://github.com/zheminzhou/pHierCC) [173]. (Appendix 3.4).

These statistical approaches were employed to determine LIN code thresholds; however, thresholds were not defined solely based on this. The quality of clustering is highly dependent on the characteristics of the dataset and the specific objectives of the clustering. Therefore, results of the statistical tests were plotted alongside the PAM distribution to further refine the selection of previously identified thresholds.

3.2.4. LIN code implementation on PubMLST *H. influenzae* database and the LIN code prefix interpretation

Finally, LIN code was implemented in the PubMLST *H. influenzae* database, which involves the following steps: 1) extraction of core genome profile data from the typing database and isolates; 2) construction of core genome minimum-spanning trees (MSTs); 3) determining the optimal order for adding "new" genomes (i.e. genomes without LIN codes); 4) LIN code assignment [174, 188]. Since Dataset 1 in this study was deposited in PubMLST database, the final LIN codes were also assigned to these genomes.

Each genome was assigned a LIN code, which consists of a series of numbers (Figure 3.9). The total count of digits in a LIN code corresponds to the number of thresholds used. Each number represents a grouping or "bin", providing information on core genome similarity, which increases progressively from the left-most to the right-most bin. A subset of the LIN code can be generated by including the left-most number and extending up to a specified endpoint, referred to as the "LIN code prefix" in this study [188]. Apart from the left-most number, each subsequent number in the LIN code prefix can be interpreted as a subgroup of the preceding bin.

Although the LIN code provides a rapid phylogenetic overview, it can be difficult to interpret. To improve usability, we introduced a set of nicknames for its hierarchical prefixes, namely megalineage, superlineage, lineage, and clonal group (CG). In Chapter 2, single-linkage core genome clustering (CGC) at multiple thresholds yielded two key insights: typeable *H. influenzae* isolates fall into distinct CGCs, and CGC500 (clusters defined by ≤ 500 allelic mismatches) generally correspond to MLST CCs. Accordingly, megalineage, superlineage, and lineage-level prefixes were nicknamed by serotype (e.g., Megalineage a,

Megalineage b), with numeric suffixes added when a serotype encompasses multiple megalineages (e.g., Megalineage a1 Megalineage ML a2). Additionally, prefixes congruent with CCs were labelled as CG followed by the corresponding CC number.

To identify key *H. influenzae* clusters at each LIN code partition, phylogenetic analyses were performed. A maximum-likelihood (ML) phylogeny of *H. influenzae* draft genomes was inferred from their core-genome alignment using RAxML, with recombination subsequently accounted for via ClonalFrameML. [100, 151]. The tree was visualised using iTOL [189] and annotated with both the LIN code prefix and the corresponding LIN code nickname. A comprehensive literature review was also conducted to identify *H. influenzae* clusters of clinical and public-health importance. Isolates from these clusters were then assigned cgSTs and LIN barcodes to illustrate the scheme's utility in addressing relevant clinical and public-health questions.

Lastly, the concordance between pre-existing, widely used typing methods, such as MLST, and the proposed LIN code-based clustering was evaluated across each partition threshold. The pre-existing methods included capsule genotype [152], MLST (ST and CC) [48], ribosomal MLST (rMLST) clustering at 5, 10, and 20 allelic mismatches [140, 190], PopPunk cluster [191], and NTHi pathotype [60], with the latter applied only to the NTHi subpopulation. The adjusted Rand index statistic was used to assess the level of agreement between the LIN code clusters and the pre-existing methods, with scores approaching 1 indicating high concordance (Appendix 3.2).

3.2.5. Average nucleotide identity

Because the LIN code concept was initially developed using ANI percentage as threshold level, we evaluated the range of pairwise ANI value for genomes in the same cluster at multiple partitioning thresholds. Pairwise comparison of the core genome average nucleotide identity (ANI) of *H. influenzae* draft genomes in the dataset was done utilising fastANI [192]. We summarised the central tendency (median) and dispersion (minimum and maximum) of the ANI percentage for every genome pair and evaluated the differences in these measurements for genome pairs within the same cluster at each LIN code threshold (Appendix 3.3).

3.3. Results

3.3.1. Thirteen LIN code partitioning levels were identified to define clusters of *H. influenzae* genomes with 12.25% to 99.9% similarity level

H. influenzae draft genomes in this study were internationally dispersed, with the majority (81%) originating from Europe or North America. While the disease or health status origin was unknown for 57% of the isolates, 55% of those with complete provenance data were from invasive disease cases (e.g., meningitis and bacteraemia). There were 545 (19%) unique MLST STs out of 2,830 STs identified in the PubMLST database at the time of analysis (23 February 2024), with 2,128 (75%) STs lacking whole-genome sequence data or having inadequate sequencing quality (Supplementary Table 3.1). The core genome profile dataset contained 5,569 unique cgSTs, 2,200 of which were represented in the draft genome dataset (Supplementary Table 3.2).

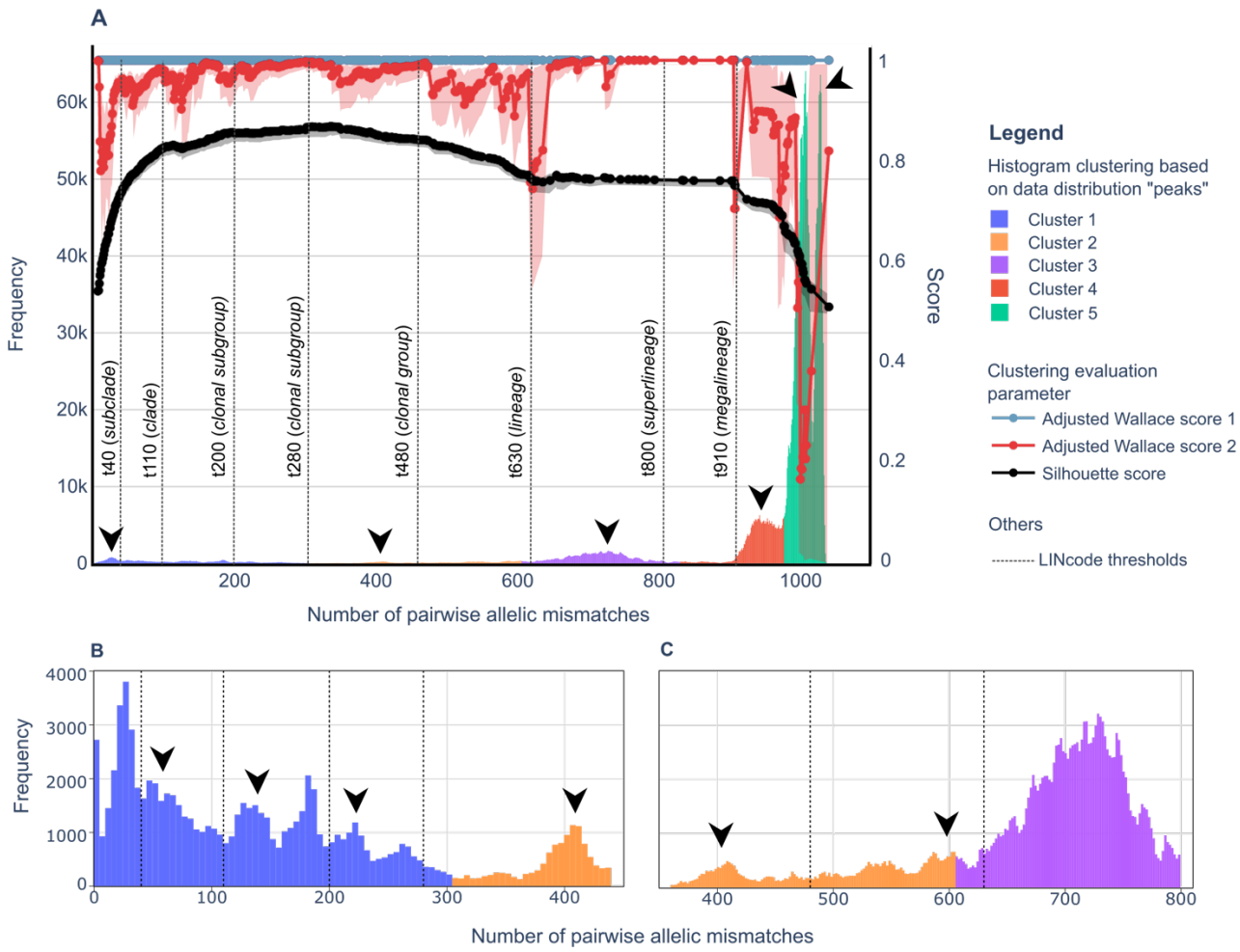


Figure 3.8. Distribution of pairwise allelic mismatches (PAM) for 2,297 high-quality H. influenzae draft genomes and statistical evaluation for clustering at multiple allelic mismatch thresholds. (A) Multimodal PAM distribution curve with six apparent peaks (black arrowheads). The curve was coloured based on the data modes identified using k-means clustering method and annotated by three statistical parameters: silhouette score (solid black line), adjusted Wallace score 1 (dark blue), and adjusted Wallace score 2 (red). Eight cgMLST-based thresholds were concluded: t910 (12.25% similarity), t800 (22.85%), t630 (39.25%), t480 (53.71%), t280 (73%), t200 (80.71%), t100 (89.39%), and t40 (96.14%). (B) and (C) showed the zoomed-in distribution graph from (A) in 0-450 range and 350-800 allelic mismatches. Subset analyses of the draft genomes were conducted to generate similar curves, helping to identify additional partitioning thresholds that were not clearly visible on the primary curve (Appendix 3.1).

The distribution curve of PAMs of 1,037 core genes in the *H. influenzae* cgMLST scheme depicts genetic discontinuity within the bacterial population structure (Figure 3.8). The curve was multimodal, with six prominent peaks observed at the 1020, 1000, 940, 700, 400, and 30

mismatche thresholds, each corresponding to 1.6%, 3.6%, 9.4%, 32.5%, 61.4%, 97.1% similarity, respectively (Figure 3.8, black arrowheads). From PAM distribution curves of subsets of *H. influenzae* genomes (Appendix 3.1), four additional modes occurred at 500, 250, 160, and 50 allelic mismatches, reflecting 51.8%, 75.9%, 84.6%, and 95.2% similarity. The specific allelic mismatch levels surrounding these modes were provisionally defined as LIN code thresholds and subsequently underwent statistical validation (Supplementary Table 3.3). Subsequently, eight thresholds were included — t910 (12.25% similarity), t800 (22.85%), t630 (39.25%), t480 (53.71%), t280 (73%), t200 (80.71%), t100 (89.39%), and t40 (96.14%) — for defining LIN code partition levels and prefixes (Figure 3.8 and Appendix 3.1 dashed lines). These thresholds equal to 12.25%, 22.85%.

The finalised LIN code implemented on PubMLST *H. influenzae* database consisted of eight partition levels and five additional 'shallow' partition levels. This shallow-level classification will be especially beneficial to distinguish outbreak clusters. Five shallow partition levels were chosen: 10, 5, 2, 1, and 0 allelic mismatches or 99, 99.5, 99.8, 99.9, and 100% core genome similarity [180, 181]. Therefore, there were a total of 13 LIN code thresholds implemented on PubMLST. The LIN code was automatically assigned to all *H. influenzae* genomes uploaded in the database with a defined cgST. Final LIN code as shown in PubMLST for the draft genomes in this study was documented in the Supplementary Table 3.1.

3.3.2. Decoding LIN code prefixes: A guide to understanding and converting machine-readable barcode into user-friendly nomenclature

Any *H. influenzae* genome uploaded to PubMLST is assigned a cgST and a LIN code if it has fewer than 25 core genes missing allele designations [156]. A LIN code consists of 13

numbers, each indicating a bin for a specific core genome similarity value. Except for the first number, every number in the bin represents a subgroup of the bin on its left and therefore, can only be interpreted in the context of the previous number. For example, the number in the superlineage column (Figure 3.9) is interpreted as follows:

- 0: This number indicates a subgroup of megalineage "0." Therefore, the isolate on the purple branch belongs to megalineage 0 and superlineage with the LIN code prefix "0_0." It shares at least 12.25% and 22.85% core genome similarity with other genomes in the same megalineage and superlineage, respectively.
- 5 and 6: These represent two subgroups of megalineage "1." Isolates on the red, blue, and yellow branches belong to superlineage "1_6," while the isolate on the green branch belongs to superlineage "1_5." All of these isolates share at least 12.25% similarity since they are part of megalineage 1 but have less than 12.25% similarity with the isolate on the purple branch.

This rule can be implemented for each LIN code bin. Consequently, the higher the similarity level of a bin, the longer the LIN code prefix assigned to it. While this system contains the genomic similarity information for every genome, it is not immediately understandable. To overcome this issue, a "nickname" can be attached to a LIN code prefix. For instance, within superlineage 1_6 in Figure 3.9, three distinct lineages were identified and assigned the nicknames "Lineage 1," "Lineage 2," and "Lineage 3". All isolates on the pink branch belonged to the Lineage 1 and this nickname carried the LIN code prefix information of 1_6_2. Similarly, clonal groups within the Lineage 1 were nicknamed "Clonal Group 1" and "Clonal Group 2". For demonstrative purposes, the numbers in the nicknames were assigned in sequential order.

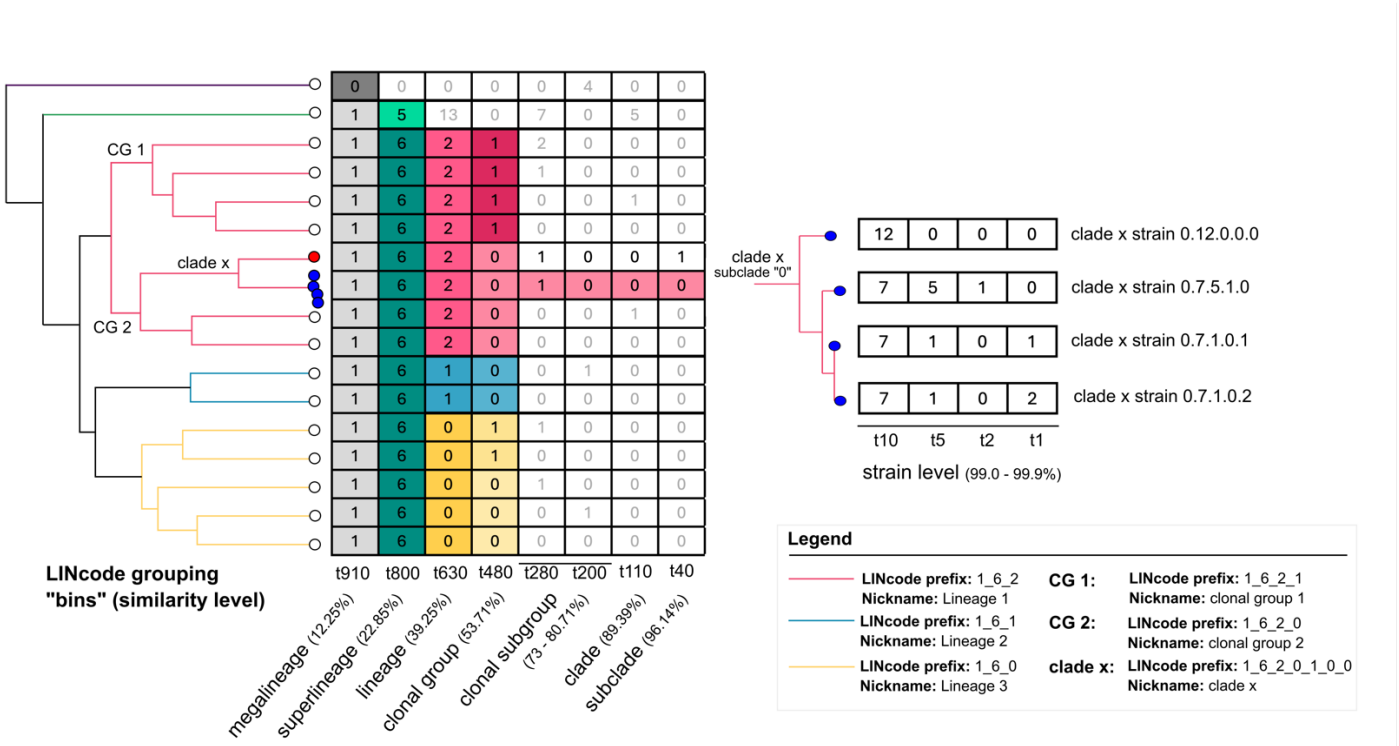


Figure 3.9. Illustration of LIN code grouping and nickname nomenclature implementation using partitioning thresholds defined for *H. influenzae*. The left side of the figure shows a dendrogram representing a phylogenetic relationship among 20 hypothetical genomes. The table in the middle contains series of numbers from LIN code assigned to each genome, from megalineage to subclade "bin" with increasing genetic similarity. Except for the first number, every number in the bin represents a subgroup of the bin on its left. Because of this, LIN code prefixes always included all the numbers to the left, up to a specified level. "Nicknames" such as "Lineage 1" (pink branches), "Lineage 2" (blue branches), and "Lineage 3" (yellow branches) can also be attached to LIN code prefixes, simplifying the nomenclature. The dendogram on the right is a magnification of the clade x subclade "0". For each isolate in this group, the strain level LIN code numbers were shown.

Alternatively, when a specific cluster of isolates had been reported to be clinically or publicly significant, the name previously used for this cluster can also be adapted to the nickname attached to a LIN code prefix. This exemplified by the node denoted as "clade x" in the dendrogram (Figure 3.9). The five isolates belonging to this node had the same LIN code prefix up to the clade grouping bin, which was 1_6_2_0_1_0_0. In this hypothetical scenario,

a previous report had identified this cluster and named them “clade x” and the nickname is therefore attached to the LIN code prefix.

The last LIN code grouping bin is for strain differentiation, which consists of five additional LIN code thresholds (i.e. shallow-level partitioning). As an example, a cluster of four isolates (Figure 3.9 blue nodes) within the clade x also belonged to the same subclade. Strain level notation starts with the subclade number, followed by the rest of the LIN code. Using LIN code bins at strain level, it can be concluded that:

- One isolate was different from the three only by 10 core genes (different number at “t10” level);
- One of those three isolates had 5 allelic mismatches (different number at “t5” level) compared to the rest; and finally,
- The two most similar isolates had only 1 allelic mismatch (different number at “t1” level).

This level of detailed differentiation is especially valuable for outbreak detection and epidemiological investigations, where identifying subtle genetic variations can provide crucial insights.

3.3.3. Ten major lineages of global collection of *H. influenzae* genomes were identified

Partitioning based on the LIN code barcoding system shows high congruence with the core genome ML tree topology, indicating that this clustering system reflects the *H. influenzae* population structure (Figure 3.10, Supplementary Table 3.1). This consistency was observed at all partition levels, both across the entire *H. influenzae* genome dataset and within smaller

subsets of the dataset. To illustrate this, the ML tree (Figure 3.10) was annotated by LIN code prefixes at lineage (Figure 3.10A) and clonal group level (Figure 3.10B).

Ten major lineages, each containing more than 50 genomes, were identified. Four of these lineages were associated with specific *H. influenzae* capsule types, while the remaining lineages were comprised solely of NTHi genomes. Although a small number of NTHi genomes (n = 1–15) were present within capsule-associated lineages, the NTHi-associated lineages were composed exclusively of NTHi genomes. Nicknames were assigned to lineages associated with typeable *H. influenzae*. With the exception of Hia, clustering at megalineage, superlineage and lineage level was identical for each capsulated *H. influenzae* and therefore, in this case, their nicknames for these three partition levels were identical (Supplementary Table 3.4).

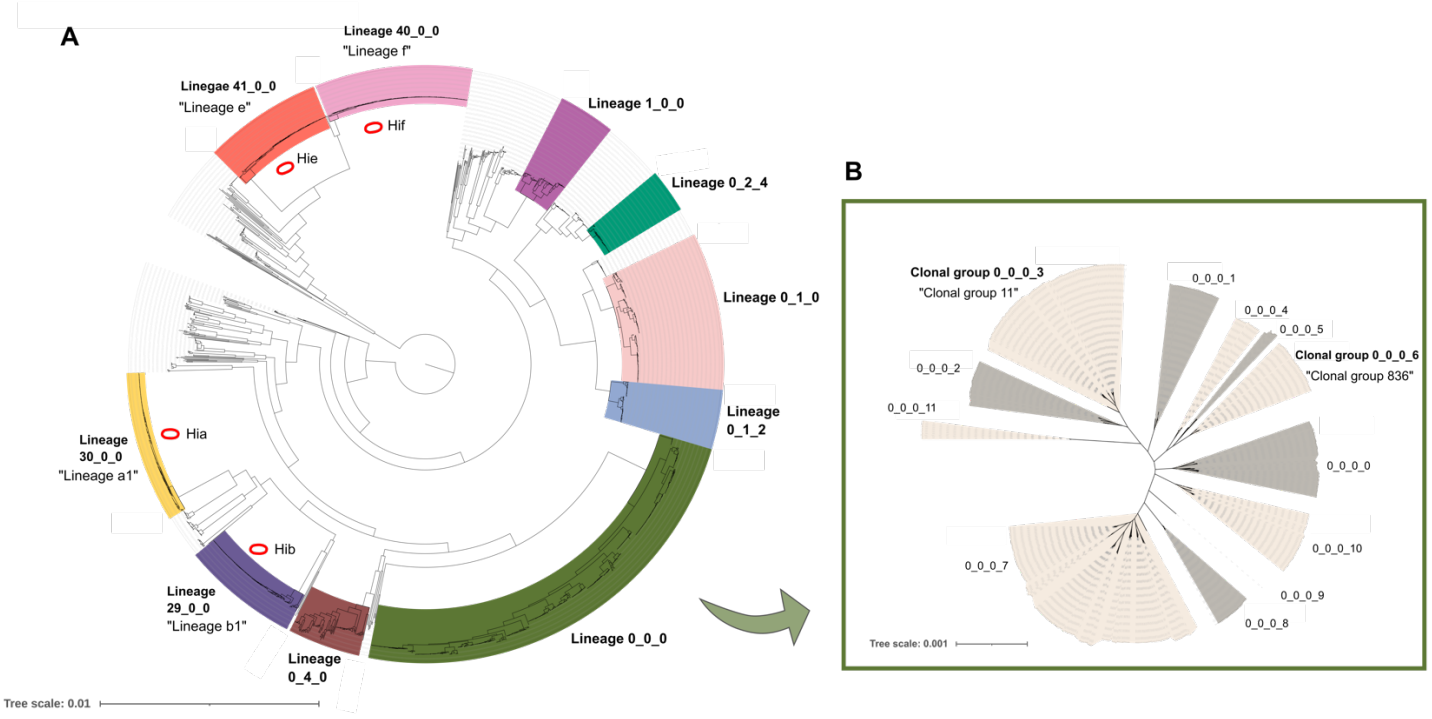


Figure 3.10. A Maximum-Likelihood (ML) phylogeny of 2,297 *H. influenzae* genomes annotated by LIN code clustering at lineage and clonal group level. **(A)** The main ML tree with the 10 major lineages coloured. Genomes belonging to the same lineage had at least 39.25% core genome similarity. Lineages

consisted mainly of capsulated isolates were indicated by a red circles followed by the capsule type: Hia, Hib, Hie, and Hif for *H. influenzae* type a, b, e and f. **(B)** A subset of the tree in the main panel, consisting of genomes within Lineage 5 (n = 554). This tree subset was annotated by LIN code group at clonal group level, with each alternating colours marking a different clonal group. Two clonal groups were nicknamed as Clonal Group 11 and Clonal Group 836 due to its exclusive association with the ST-11 and ST-836 complex, respectively.

Because previous report on *H. influenzae* core genome clustering at 500 allelic mismatches showed a high congruence with MLST CC [156], LIN code-based clustering at each threshold was compared to previously defined clustering methods (Appendix 3.2). The comparison resulted in an adjusted Rand index score of 0.84 between LIN code clonal group and MLST CC. Therefore, when a clonal group consisted entirely of genomes from a specific CC, the CC name was incorporated into the nickname [178, 179]. As an example, Lineage 5 in the ML tree (Figure 3.10) was magnified to depict the partitions at clonal group level. Within this lineage, 12 clonal groups were identified, two of which were exclusively associated with a CC. The first group, "Clonal Group 11," was associated with the ST-11 complex and linked to the LIN code prefix 0_0_0_3, while “Clonal Group 836” was associated with the ST-836 complex and linked to the LIN code prefix 0_0_0_6. Data for other clonal groups associated with specific CCs and nicknamed accordingly is provided in Supplementary Table 3.4.

3.3.4. LIN code clustering bins at clade, subclade, and strain level are useful for epidemiological tracking and investigation.

Higher-level thresholds are useful for providing an overview of *H. influenzae* population structure. Here, implementation of LIN code prefixes and the nickname nomenclature was showcased using real-life data of *H. influenzae* type a (Hia) isolated in Canada from 2001 to 2018 (N = 121) [181]. All genomes in this Canadian dataset were also part of the draft genome

dataset used in this study. An ML phylogeny was reconstructed for the Canadian genomes, using an additional reference genome (GCF_020736045.1) to root the tree (Figure 3.11). The majority of genomes (n = 103) belonged to Clonal Group 23, characterised by short branches that were not readily distinguishable from the overall tree topology. LIN code grouping at clade level revealed seven distinct clades, one of which had been previously identified in the literature as Clade 1.1 [64].

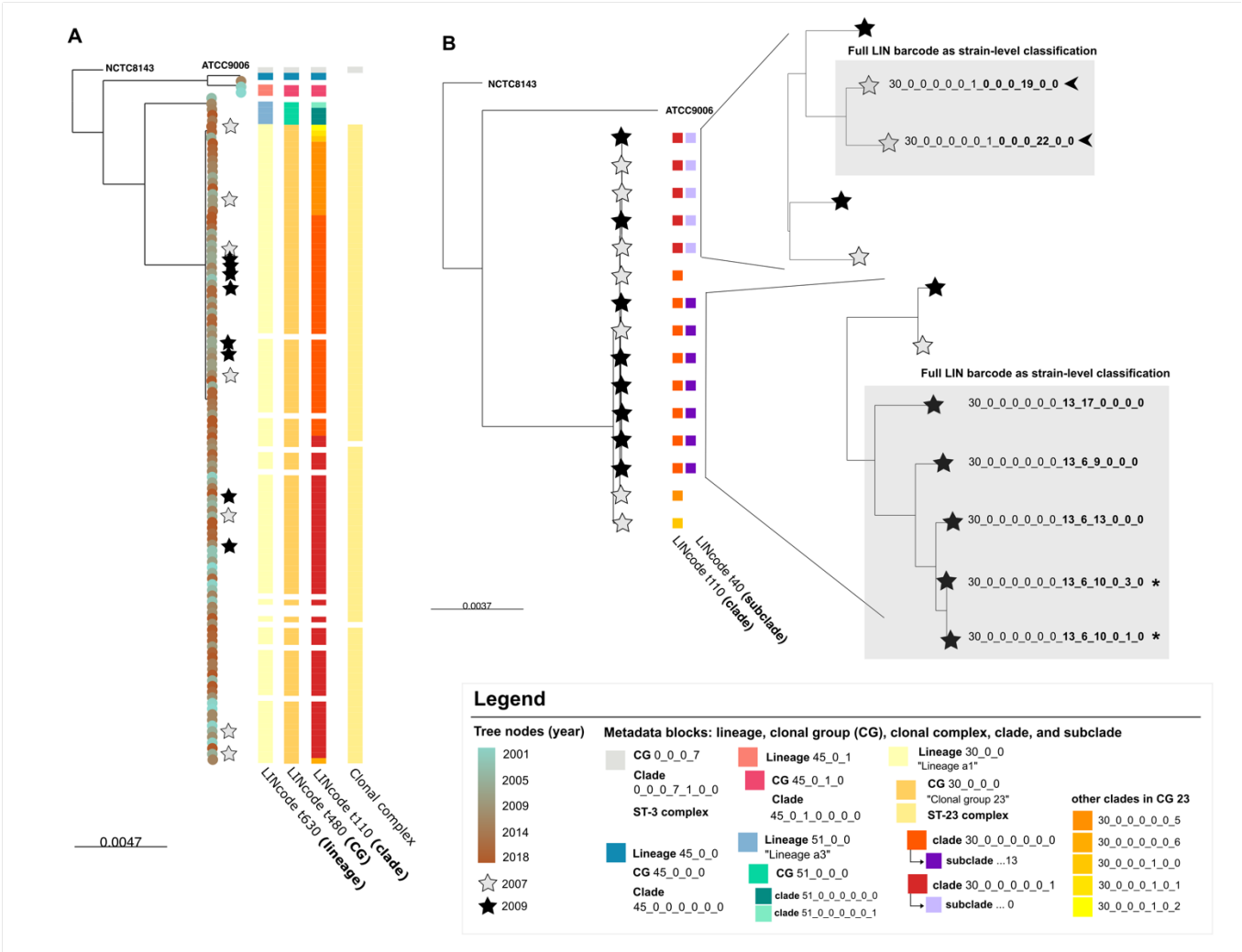


Figure 3.11. Implementation of LIN code clustering at clade, subclade, and strain level to the ML phylogeny of 123 *H. influenzae* genomes as described by Tsang et al [181]. **(A)** The main ML tree with nodes coloured by year of isolation and four metadata blocks (left to right): LIN code lineage, clonal group, clade and MLST CC. Excluding the reference genomes (NCTC8143 and ATCC9006), there were three distinct lineages, the lineage 30_0_0 ("Lineage a1"), 51_0_0 ("Lineage a3"), and Lineage 45_0_1

(nickname not given, but part of the "Superlineage a2"). Lineage a1 was the largest and contained genomes from ST-23 complex. All ST-23 complex isolates belonged to the CG 30_0_0_0, therefore given the nickname CG 23. Isolates from year 2007 (grey star) and 2009 (black star) were subset. **(B)** A subset of the main ML tree, comprising genomes isolated in 2007 and 2009. Two subclades (no nickname was given) were identified. Within each subclade, pairs of nearly identical genomes were observed (grey highlight, asterisk and arrowheads), suggesting a potential epidemiological link.

To demonstrate the use of the LIN code system for strain-level identification, a tree subset was created using genomes isolated in 2007 and 2009 (Figure 3.11A star symbols and Figure 3.11B). We assessed clusters at the subclade bin (t40, 96.14% similarity), and genomes within the same subclade were further analysed for strain-level differentiation. Two dominant subclades emerged, with most Hia isolated in 2009 (6/8) belonged in one subclade (Figure 3.11B dark purple blocks). These isolates were different with each other by a maximum of 10 core gene alleles and two of them (PubMLST ID 16677 and 16679) had identical core genome alleles except for one gene (Figure 3.11B asterisk). Similarly, two isolates (PubMLST ID 16668 and 16744) from the other clade, both isolated in 2007, differed by only two core genes (Figure 3.11B arrowheads).

Similarly, the LIN code method was implemented for NTHi genomes causing invasive infection among adults with HIV in Atlanta, U.S.A between 2008-2018 [70]. The phylogeny was reconstructed using a neighbour-joining approach and annotated with LIN code prefixes (Appendix 3.5). The LIN code method reproduced the original results while offering additional insights. As previously noted, in this dataset two closely related clusters were identified at LIN code lineage level, the Lineage 26_0_2 and Lineage 26_0_1, each displaying a clear temporal association.

Lineage 26_0_1 comprised primarily 2017–2018 isolates of ST 164. However, a PubMLST query (September 2024) showed that not all genomes in this lineage belonged to ST 164: of

the four non-ST 164 isolates, two shared five or more MLST alleles with ST 164, while the other two did not. This discordance highlights the limited resolution of seven-locus MLST in a highly recombinogenic species, where horizontal gene transfer can alter housekeeping loci. Within this lineage, 19 of 21 isolates formed a homogeneous subclade (LIN code prefix 26_0_1_0_1_0_0_1), which we have designated "Subclade Atlanta 1" because, as of 15 April 2025, no other PubMLST entries belonged to it. These isolates differed by only one allele among 1,037 core genes (99.9 % similarity), strongly suggesting a recent transmission event, a finding corroborated by epidemiological data. [70].

Conversely, Lineage 26_0_2 encompassed isolates from the ST-519 clonal complex. Within this lineage, a single clonal group (26_0_2_0) was identified and accordingly nicknamed "Clonal Group 519." A pronounced temporal split at the subclade level separated pre-2013 and post-2013 isolates: those collected before 2013 formed "Subclade Atlanta 3" (LIN code prefix 26_0_2_0_0_0_0_0), consisting exclusively of Atlanta, USA isolates and suggesting localised transmission. As of September 2024, only five additional isolates in PubMLST were assigned to Lineage 26_0_1 and twelve to Lineage 26_0_2, all from Europe. Both lineages merit close monitoring given their association with invasive disease in immunocompromised hosts and their potential to spark outbreaks.

3.4. Discussion

H. influenzae remains a significant pathogen, particularly because current vaccines only protect against Hib, leaving other capsule types and NTHi—linked to severe diseases like septicemia in the elderly—unaddressed [20, 21]. With an aging global population, particularly in developed nations, this poses a growing public health concern. Advances in sequencing

technology now offer the tools to explore *H. influenzae* population genomics more effectively, crucial for understanding its invasiveness and guiding vaccine development. The cgMLST-based LIN code system enhances this effort by providing a unique genomic "barcode" for high-resolution isolate identification and a stable clustering framework [174, 179]. This method can be integrated with pre-existing classification systems, preserving links to historical groupings through its nickname nomenclature while providing high-resolution genomic insights.

The LIN code was initially developed based on whole-genome similarity using average nucleotide identity (ANI), with partition thresholds reflecting biologically relevant relationships [175]. The cgMLST-based LIN code follows the same principle but uses core genome allelic mismatch to define thresholds. This method has been successfully applied to *K. pneumoniae* and *S. pneumoniae*. For *K. pneumoniae*, four key thresholds (t610, t585, t190, t43) distinguish isolates at species, subspecies, sublineage, and clonal group levels, plus five additional thresholds (t10, t7, t4, t2, t1, t0) for epidemiological investigation [174]. Similarly, *S. pneumoniae* LIN code, based on 1222 core genes, includes five higher-level thresholds (t1180, t750, t540, t160, t25) and six shallow ones (t15, t8, t4, t2, t1, t0), covering 3% to 100% similarity. [178].

The LIN code thresholds for *H. influenzae* in this study span similarity levels from 12% (t910) to 100% (t0). Thresholds below 10% were excluded, as they are typically used for distinguishing species or subspecies [174, 178]. Additionally, each microorganism group has a unique evolutionary history that influences the appropriate range of genome similarity for clustering purpose. Moreover, the pairwise ANI of *H. influenzae* genomes within clusters

revealed that higher allelic mismatch thresholds corresponded to lower ANI, consistent with the original implementation of LIN code using ANI [175].

Population genetics studies on *H. influenzae* have aimed to identify significant lineages or clones linked to outbreaks or invasive disease. For instance, Musser et al. identified two "clonal divisions" of capsulated *H. influenzae* before the WGS era, grouping most Hia, Hib, and all Hic and Hid in division I, and the minority of Hia, Hib, and all Hif in division II, with Hie as intermediate [193, 194]. With the cgMLST-based LIN code proposed in this study each capsule type is distinguishable even at the megalineage level. For Hic, Hid, Hie, and Hif, there is only one megalineage per capsule type, indicating less than 12% similarity between genomes of different capsules. Hia and Hib have three urlineages each, reflecting previous findings on major and minor Hia [64] and Hib “clades” [3]. Additionally, there were no subdivisions at the superlineage or lineage levels within all capsule group-associated urlineages, except for Hia and Hif. This indicated a limitation of sampling of available *H. influenzae* genomes; however, could also suggest a high degree of genetic similarity among most *H. influenzae* genomes within the same capsule group [65]. Four out of the ten major lineages with assigned nicknames in this study were associated with capsulated *H. influenzae*.

Previous studies identified four "clones" of invasive Hia, each dominated by a specific MLST-defined ST: ST 4, 23, 62, and 372 [170]. This finding has been confirmed by subsequent population genomics evaluations [64, 181]. The MLST scheme remains valuable when WGS is unavailable and provides a stable naming (ST) and clustering (clonal complex) at a broader level. Accordingly, in developing the LIN code system, its relationship with MLST was carefully considered. In our phylogeny reconstruction (Figure 3.11), the “clones” correspond to three LIN code CGs (CG 4, 23, and 62) with nicknames based on their dominant STs and

one without a nickname (LIN code prefix 45_0_0_0). While ST-62 and ST-372 both belong to the same megalineage, ST-4 and ST-23 are not as closely related, challenging earlier assumptions that grouped ST-4 and ST-23 together in "clonal division I" and ST-62 and ST-372 in "clonal division II" [170].

While mapping established lineages or clades to the cgMLST-based LIN code is straightforward for encapsulated *H. influenzae*, this is not the case for NTHi due to its vast diversity. For instance, Fong *et al.* identified 32 NTHi isolates from invasive cases in New South Wales (2017-2022), linked to 22 different STs [195]. Similarly, Slotved et al reported 581 NTHi isolates through the National Danish Surveillance Programme exhibiting 148 distinct STs [23, 67]. Although these STs can be translated into LIN code clonal subgroups or clades, they belong to various lineages, superlineages, and urlineages. Most invasive NTHi STs fall within Lineages 5 and 6 of the largest megalineage 0, which includes isolates from invasive and non-invasive sources, thus limiting associations with invasiveness. Nevertheless, the cgMLST-based LIN code offers a reliable framework for assessing genetic relatedness and consistent cluster reporting across multiple thresholds, while avoiding the use of ambiguous term for to refer specific cluster in a literature [171]. This system can also support future comparative genomics in identifying variants linked to invasiveness or antibiotic resistance.

The cgMLST-based LIN code has proven to be both accurate and efficient for distinguishing between outbreak and non-outbreak cases in public health investigations. To illustrate this, we demonstrated the application of LINcodes in three previous studies, including one suspected invasive Hia outbreak. Terrat et al. reported two meningitis cases at the same hospital in California within a week, raising concerns of an outbreak [196]. Analysis of the two genomes (PubMLST IDs 16540 and 16541) revealed that they belonged to distinct

urlineages (45 and 30), confirming no outbreak link. This leads to the key question: which specific LIN code threshold is optimal for reliably distinguishing true outbreak cases from unrelated infections?

The second report [181] originated from invasive *H. influenzae* surveillance programme in Canada, which has described multiple invasive Hia outbreaks [197, 198]. While epidemiological tracking was not performed, the LIN codes suggested a possible transmission chain. Hia isolates from the same year tended to cluster within the same **subclade bin** (t40, 96.14% similarity), with several genomes showing near-identical alleles across the core genome. A similar pattern was observed in a third report from Atlanta, USA, involving invasive NTHi strains [70]. Geospatial analysis revealed two clusters that aligned with LIN code clustering at the lineage threshold (t630, 39.25%), with a clear temporal pattern emerging at the **subclade** level. These observations suggested that the subclade bin could be used to indicate potential outbreaks, although more data are needed to establish a definitive threshold. A recent CDC report described an invasive NTHi outbreak in which four isolates showed zero SNP differences after excluding highly recombinant regions, though exclusion criteria were unclear [31]. The LIN code method can streamline outbreak analysis by providing a consistent framework, avoiding methodological variations. While 100% similarity (0 SNPs or allelic differences) confirms an outbreak, minor allelic differences—up to 10 mismatches (99% similarity)—should also be considered, especially in the highly recombinogenic *H. influenzae* population [199].

With the LIN coding method, each *H. influenzae* isolate will be assigned a barcode, providing genealogy information that is readily interpretable. Depending on their needs, users can 1) determine whether their isolates belong to specific groups of interest by examining the

barcode prefixes on the left side, or 2) investigate whether their isolates are closely related to each other by inspecting the right end of the barcode. There are some advantages and limitations of the cgMLST-based LIN coding. For genomes within a bacterial species that continuously underwent homologous recombination, the gene-by-gene approach using comparison of allelic profile is preferable for estimating evolutionary relationship [103, 117]. Additionally, when the cgMLST scheme is integrated to and maintained in a publicly available database such as PubMLST, it is highly accessible to public health authorities and researchers worldwide and can be implemented to other platforms. Nevertheless, the main limitation of this method is the prerequisite of well-developed cgMLST scheme and this scheme is applicable only for a single species or genus. Therefore, cgMLST-based LIN code is interpreted within a context of a specific species or genus, rather than as a universal nomenclature system applicable across all organisms [175]. Lastly, the use of multiple thresholds in defining partitions for the LIN code is highly dependent on the population structure of a species. *H. influenzae* shows some evidence of genetic discontinuity pattern which guided the decision on the thresholds. This pattern was clearer in some other bacteria, such as those belonging to Kpsc [174], but can be hazy for another, such as *S. pneumoniae* [178].

3.5. Conclusion

The cgMLST-based LIN code method offers a robust and stable framework for the genomic classification of *H. influenzae*. By integrating high-resolution cgMLST scheme with a multilevel clustering approach, this, not only reflects the population genetic structure of the species, but also provides a standardised means of comparison across diverse datasets. We demonstrate the congruence between LIN codes and the established genealogy, confirming its

value in identifying and differentiating lineages and sublineages within *H. influenzae*. The ability of LIN code to generate unique, barcode-like identifiers for individual isolates, combined with its compatibility with pre-existing classification systems, ensures that it can be integrated into ongoing public health surveillance and research efforts. Finally, the LIN code system improves our understanding of *H. influenzae* population dynamics, bolstering efforts in antimicrobial resistance monitoring and guiding future vaccine developments.

Chapter 4

Variants in surface-exposed proteins are associated with invasive non-typeable Haemophilus influenzae infection: A genome-wide association study

Abstract

Non-typeable *Haemophilus influenzae* (NTHi) has emerged as a primary cause of invasive *H. influenzae* infections in the post-Hib vaccine era, causing significant morbidity and mortality globally. This study utilised publicly available genomes divided into two independent datasets for initial (n = 1975 genomes) and validation GWAS (n = 1165 genomes). A reference-free k-mer-based approach was applied, suitable for detecting variants in both core and accessory genes, including highly variable regions. Initial GWAS identified 622 significant k-mers, predominantly in non-coding regions or accessory genes, associated with the invasive phenotype. Of these, 256 k-mers were confirmed in the validation dataset, with 146 showing significant associations in further statistical analyses. The logistic regression model, built using 19 independent k-mers, achieved an accuracy of 66.1% and an AUC of 74.54%, highlighting the predictive value of these variants. Notably, genetic variants were mapped to genes encoding porin and TonB-dependent receptors (TBDRs), as well as tRNA and rRNA genes, suggesting their potential roles in immune evasion, iron acquisition, and protein synthesis under host-imposed stress. Recombination analyses revealed two hotspots coinciding with these genes, indicating that horizontal gene transfer might contribute to the dissemination of invasive traits. These findings enhance the understanding of NTHi pathogenicity and can inform vaccine development strategies targeting these genetic determinants.

4.1. Introduction

Haemophilus influenzae is a fastidious, pleomorphic Gram-negative coccobacillus that exclusively inhabits the human host, primarily residing in the upper respiratory tract as a commensal member of the microbiota, yet it can also act as an accidental pathogen. It is classified based on the expression of capsular polysaccharides into six serotypes (a to f) with an additional unencapsulated, or nontypeable (NTHi), group [200]. Clinically, *H. influenzae* is associated with a broad spectrum of presentations ranging from non-invasive infections, such as otitis media and sinusitis [1, 2], to invasive diseases like meningitis, epiglottitis, orbital cellulitis, and septicaemia. This spectrum of manifestations highlights its significance in clinical settings, with bacterial biological factors playing a crucial role in determining the outcome of infection [201].

Following the worldwide implementation of the *H. influenzae* type b (Hib) polysaccharide-conjugate vaccine, Hib was rapidly controlled as the predominant cause of invasive *H. influenzae* infections, notably causing a sharp decline in Hib meningitis in children [202, 203]. In the postvaccination era, however, NTHi has emerged in carriage anda leading cause of invasive disease [204, 205]. Epidemiological data consistently demonstrate that the majority of invasive cases are now attributable to NTHi, which was responsible for 64% and 78% of all invasive *H. influenzae* disease in Europe [206, 207] and South Africa [208], respectively, with global case fatality rates for invasive NTHi infection exceeding those of Hib [20].

Unlike the encapsulated *H. influenzae* from which almost all circulating strains are recovered from invasive disease, NTHi can be associated with carriage or recovered from

respiratory tract infection or invasive disease. In the past 2 decades, studies of NTHi have focused on specific high-risk populations such as those with chronic obstructive pulmonary disease (COPD) [158, 209-211] and cystic fibrosis (CF) [212, 213]. These studies are typically interested in investigating virulence factors (VFs) associated with acute respiratory tract symptoms such as pneumoniae and bronchitis, or persistent airway infections [158, 209, 210, 212, 213]. VFs were defined as "gene products that enable a microorganism to: 1) colonise a host niche, 2) proliferate, and 3) cause tissue damage or systemic inflammation" [11, 12]. The latter includes factors that facilitate invasion through epithelial and endothelial barriers, such as the LnlA surface protein in *Listeria monocytogenes,* which interacts with human intercellular E-cadherin [214]. Previously defined VFs (and their encoding genes) which were identified in COPD-associated NTHi strains include glycosyl transferase (*lic2B*) [213], hemoglobin-haptoglobin binding-like protein (*hgps*) [213], high-molecular-weight proteins (*hmw*) [210, 213, 215], and IgA protease (*igB*) [209, 213]. Nevertheless, no study has investigated, on a genomewide- scale, the factors encoded in the NTHi genome that are associated with invasive disease. With the increasing accessibility of whole-genome sequencing (WGS) technology and its integration into public health surveillance programs [112], a growing number of *H. influenzae* genome sequences are now available in public databases. This can be leveraged to conduct large-scale comparative genomic analyses using available NTHi genomes, both from invasive and non-invasive cases to reveal genetic determinants differentiating the two groups.

H. influenzae is naturally competent for DNA uptake from the environment and integrate it *into its* genome, through horizontal gene transfer (HGT) [216]. HGT accounts for much of the genetic diversity observed in *H. influenzae* [54, 95], especially NTHi. Up to 40% of the Hib genome has been found to exhibit significant sequence variation [95]. The first

pan-genome analysis of the species revealed that NTHi genetic diversity of was at least ten times greater than that of Hib [54]. As *H. influenzae* inhabits the human nasopharynx—an environment populated by many other bacterial species [217] and protected by multiple layers of immune defences [218]—its natural competence enables it to acquire beneficial genetic material, enhancing its fitness and competitive advantage. The continuous exchange of genetic material, coupled with intrinsic mutations driven by positive selection, may contribute to the emergence of invasive NTHi strains capable of breaching the mucosal barrier and evading immune responses. However, the specific genetic variants associated with invasive NTHi infection remain incompletely understood.

Identifying variants associated with NTHi invasive disease would, not only enhance our understanding of the biology of NTHi, but also provide valuable insights for vaccine development. Proteins encoded by variable genes that promote bloodstream invasion are typically immunologically accessible, making them promising vaccine candidates [219, 220]. If conserved regions within these accessible proteins can be identified—or if a mosaic design combining both variable and conserved regions is employed [219, 221]—it may be possible to develop a vaccine effective against all NTHi strains, and potentially all *H. influenzae*. The most recent attempt to develop a recombinant-protein NTHi vaccine was discontinued due to limited efficacy in reducing the frequency of acute exacerbations in COPD patients [222]. This vaccine formulation included three surface-exposed proteins: Protein D, Protein E, and Pilin A [223]. While this presents a significant challenge, it also highlights an opportunity to explore alternative strategies, including novel vaccine components that target biological mechanisms underlying NTHi invasive disease.

This study aimed to identify the genetic determinants underlying the phenotypic outcome of invasive disease in the NTHi population through genome-wide association study (GWAS) experiments. Originally developed for human genome research, GWAS has since been adapted for bacterial populations [224]. In our experiments, we employed a k-mer-based approach, which is particularly well-suited for highly diverse bacterial species like *H. influenzae [225].* This method is reference-free and capable of detecting allele variants in both core and accessory genes, including those that are highly variable, such as phase-variable genes [215].

4.2. Methods

4.2.1. Phenotype definition and dataset curation of publicly available genomes

The phenotypes of interest were the phenotypic outcome of NTHi invasive disease. NTHi isolates were considered to have this phenotype if they were clinical isolates from normally sterile body sites, which encompass blood, cerebrospinal fluid (CSF), pleural cavity, joint, peritoneal cavity, and internal organ, such as brain and lymph nodes [226]. NTHi from all other isolation sources including nasopharyngeal swab in asymptomatic subjects or eye and middle ear exudate, sputum, and bronchoalveolar lavage (BAL) [227] in symptomatic patients were grouped as "non-invasive" phenotype.

This study utilised publicly available *H. influenzae* genomes divided into two different datasets. No genome was shared between the two datasets. The first dataset, referred to as "dataset 1", was utilised for the initial GWAS (see Part 4.2.4. Bacterial GWAS pipeline) and curated from PubMLST *H. influenzae* isolate database (accessed June 2024) and the NCBI Sequence Read Archive (SRA) (accessed July 2024). The NCBI SRA search was designed to

balance the number of isolates between phenotype groups. When more sequence reads were available than needed for a given phenotype, a pseudo-random number generator was used to select the required subset (https://docs.python.org/3/library/random.html) (https://github.com/artmisk13/DPhil_thesis_Hinf). Sequence reads from NCBI were screened for species definition and phenotype data availability. Only reads with conclusive species assignment as *H. influenzae* and professed phenotype data were assembled using SPAdes v4.0.0 (https://github.com/ablab/spades) [228]. The resulting genome assemblies went through initial quality control step utilising Quast v5.2.0 (https://github.com/ablab/quast) [229] and subsequently uploaded to PubMLST for MLST, ribosomal MLST (rMLST), and core genome MLST (cgMLST) loci detection and allele designation. rMLST was utilised for species identification and potential contamination detection [140]. The cgMLST profile was used to define the population structure by providing the core genome alignment for phylogeny reconstruction as well as defining lineage through life identification number (LIN) coding approach, as described in Chapter 3 [174].

The pool of all public genomes was quality-controlled using the following criteria [156]:

1. Genome length 1.4 Mb – 2.45 Mb.
2. rMLST support score of ≥ 85%.
3. Complete MLST and rMLST designation.
4. 'Good' annotation status of the cgMLST scheme, indicating that at least 1012/1037 core genes had an allele designation. This was the minimum number of core genes that had to be successfully designated an allele number for a genome to get a cgST assignment [119]. The threshold was established to accommodate incomplete assemblies inherent to short-read sequencing, while remaining stringent enough to

exclude poorly sequenced or assembled genomes. As a result, only genomes with an assigned cgST were eligible to receive a LINcode for lineage definition.

The second dataset ("dataset 2") was employed for the validation GWAS (see Part 4.2.4. Bacterial GWAS pipeline) and curated through NCBI SRA search (accessed October 2024). The same principal and workflow in dataset 1 curation was also implemented, with two additional steps: First, only sequence reads deposited with an associated published research article were included; and, second, after quality control processes as described above, all assembled genomes were subjected to a deduplication step. Isolate or strain name and sequence read accession number, combined with location and isolation date data, were used to identify potential duplicates.

H. influenzae genomes in all datasets which fulfilled the criteria described above were processed for *in silico* capsule type genotyping and capsulated genomes were excluded from downstream analyses.

4.2.2. Capsule type genotyping, genome annotation, and lineage definition.

Capsule type was defined by a *H. influenzae* capsule prediction tool, hicap [152], and NTHi genomes were defined as genomes without any evidence of genes in and insertion sequences associated with the capsule operon. All NTHi draft genome assemblies were annotated using Prokka v1.14.6 [145] with default settings while reference genomes were annotated by the NCBI Prokaryotic Genome Annotation Pipeline, as downloaded from https://www.ncbi.nlm.nih.gov/datasets/genome/ (accessed October 2024).

H. influenzae lineages were defined using core genome LIN codes (Chapter 3). The GWAS pipeline recommends using pyseer (section 4.2.4), and the PopPunk cluster system.

However, as shown in section 3.3.3 and Appendix 3.2, PopPunk [191] for *H. influenzae* was consistent with cgLIN clonal subgroups, which would result in a **greater number** of clusters, each with a **smaller sample size.** Accordingly, NTHi lineages were defined by clustering at the cgLIN superlineage level, where genomes within a cluster differed by no more than 800 core-gene allelic mismatches, in order to provide a broad overview of the population structure. To avoid confusion, cgLIN superlineage clusters will be referred to as "lineages" throughout this chapter.

4.2.3. Phylogenetic, recombination, and pangenome analyses

Maximum-Likelihood phylogeny of NTHi genomes was reconstructed from the alignment of core genes included in the *H. influenzae* cgMLST scheme [156], with GTRGAMMA model using RaXML v8.2.12 [151]. To account for recombination effects and produce recombination analyses the resulting phylogenetic tree was processed through ClonalFrameML v1.12 [100]. Two recombination statistics were produced, the R/theta, a ratio of rates of recombination and mutation, and r/m, a ratio of effects of recombination and mutation. These statistics were then compared between the two different phenotypic groups. Concurrently, reference-based alignment (reference genome: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000931575.1/) was constructed in order to visualise the recombination blocks. The alignment tool used was SKA aligner v2 [230] embedded in Gubbins recombination tool [99], followed by phylogeny reconstruction using RaXML GTRCAT model and correction for recombination.

To assess the content and annotation of NTHi pangenome, PIRATE v1.0.5 a pangenome reconstruction tool was used [132]. Default settings were chosen for percentage identity

thresholds for pangenome reconstruction. PIRATE clustered coding sequences (CDSs) across NTHi genome using this set of thresholds into a unit known as "gene families". A gene family refers to a group of homologous coding sequences that share a specific function, as determined by the functional annotation tool eggnog-mapper [146]. In this study, for the sake of clarity, the term "gene" is used in place of "gene family" throughout the chapter. The pangenome size between different phenotype groups were also assessed.

4.2.4. Bacterial GWAS pipeline

GWAS experiments were conducted using pyseer v1.3.11 [225] with a k-mer-based approach and implementing random effects models. Firstly, unitigs were generated and counted using unitig-counter [231]. Subsequently, a "capsule region masking" step was implemented before running the GWAS experiments (https://github.com/artmisk13/DPhil_thesis_Hinf). This step involved excluding the k-mers associated with the capsule biosynthesis region from the GWAS analysis to prevent them from disproportionately driving associations due to their strong link with invasiveness [232]. By masking this region and the flanking sequence that: 1) might contain the insertion element IS1016 which has been described to be associated with invasive phenotype even in the absence of capsule genes [233, 234], or 2) in linkage disequilibrium [235], we aimed to identify other genetic factors that contribute to invasiveness independently of the capsule operon, thereby providing a clearer picture of additional pathogenic mechanisms.

The step-by-step workflow is depicted in Figure 4.12 There are two major GWAS experiments: initial and validation GWAS. In the initial GWAS, there were 200 parallel GWAS runs and within each run there were 100 *H. influenzae* genomes for each phenotype category

(Figure 4.12 step B.1 and B.2). These genomes belonged to dataset 1 and were chosen and assigned to each GWAS run randomly using Numpy Python package (https://numpy.org/) (https://github.com/artmisk13/DPhil_thesis_Hinf). This step was essential to prevent p-value inflation of rare genetic variants that might otherwise appear significant due to their presence in only a few isolates with the phenotype of interest [224]. To correct for the population structure, a kinship matrix was derived from the core genome phylogeny and included in each experiment. Lineage effects were calculated with Q-Q plots visualised (Appendix 4.1), and the lineage was defined using the superlineage partition level of the *H. influenzae* cgLIN. For each run, a significance threshold was measured using a Bonferroni correction and only k-mers below this threshold were annotated back to *H. influenzae* genomes. NTHi strain 477 from NCBI Reference Sequence database (NCBI Assembly accession GCF_000931575.1) was utilised as the reference for annotating k-mers; however, all draft genomes in dataset 1 were also included to account for genetic variants which were not present in the reference. Identified k-mers in this step are summarised as a box plot. A Manhattan plot was constructed on the reference genome to visualise their location in the genome [225].

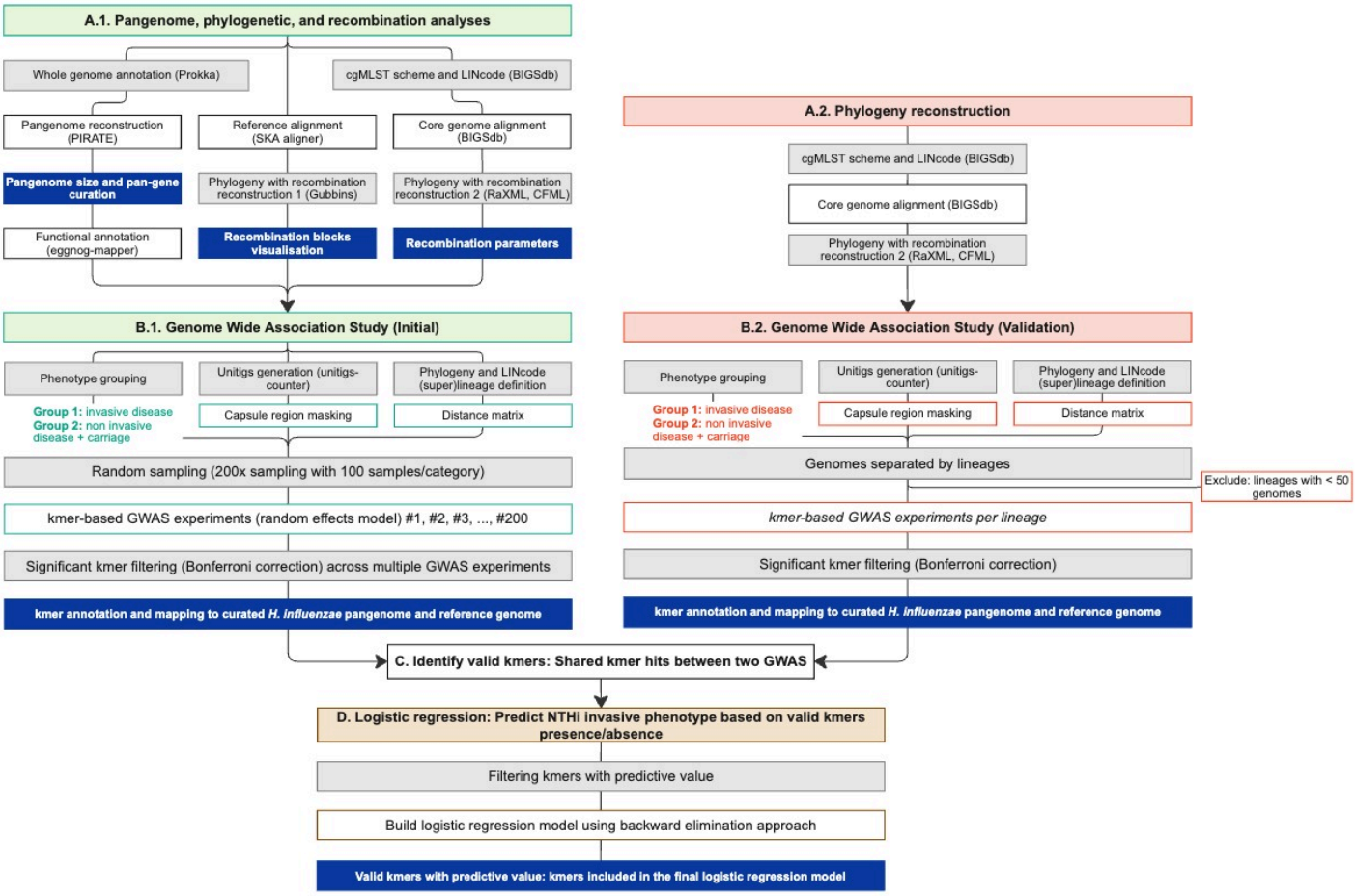


Figure 4.12 The bacterial GWAS pipeline for identifying genetic variants associated with phenotypic outcome of invasive disease in the non-typeable *H. influenzae* population. **(A)** Preparation of input file before GWAS: **A.1.** Pre-processing stage before the initial GWAS: Results of pangenome and phylogenetic with recombination reconstruction for analyses were fed as input files for the GWAS step for kmer hits annotation and population structure correction, respectively. **A.2.** Pre-processing stage before the validation GWAS: Because the validation GWAS used an independent dataset, dataset 2, phylogeny reconstruction was conducted for this dataset. **(B)** The step-by-step kmer-based GWAS approach. Both initial (**B.1**) and validation (**B.2**) GWAS involved multiple experiments; however, in the initial GWAS these were done for different, randomly assigned subset of the dataset while in the validation GWAS experiments were done for each specific lineage. **(C)** Valid kmer hits were defined by comparing the two GWAS results. **(D)** Developing a logistic regression model to predict invasive outcome of NTHi infection based on presence/absence of valid kmers.

Although the main steps in the validation GWAS analysis were similar to the initial GWAS, there was one significant modification. To confirm that the significant k-mers identified in the initial GWAS were specifically associated with a phenotype rather than

lineage, the GWAS experiment was conducted within each *H. influenzae* lineage [236]. Consequently, an additional filtering step was introduced. *H. influenzae* draft genomes in dataset 2 were clustered according to their lineage groups. Only genomes belonging to **large lineages** (N ≥ 50) were included in the GWAS experiment. An independent GWAS analysis was conducted for each of these lineages.

Valid k-mers (Figure 4.12 step C) were defined as those that: 1) were identified in both initial GWAS and validation GWAS experiments; and, 2) showed a positive association with invasive disease. The presence and absence of all valid kmers were determined for all genomes in dataset 1 and 2 using the Genome Comparator plug in on PubMLST [49], with 100% identity and 100% length criteria. The resulting presence/absence matrix was processed for further statistical testing and modelling.

4.2.5. Statistical testing and phenotype prediction using logistic regression

Here, a prediction model based on the presence/absence of k-mers associated with NTHi invasive disease was constructed. The statistical significance of all valid k-mers in both datasets 1 and 2 was retested with chi-square test using Scipy statistics package in Python (https://scipy.org/). K-mers with p value < 0.1 passed the threshold to be included as features in the logistic regression model. To improve model accuracy, there were two additional selection steps: 1) correlation matrix to check for multicollinearity; and, 2) feature importance calculation using random forest method [237].

The correlation matrix was constructed using the corr function from Pandas Python package, implementing the Pearson test to compute correlation coefficient (https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html). The result was

grouped based in the gene/region the kmers annotated to and transformed into distance matrix. Average linkage hierarchical clustering was conducted to identify clusters of highly correlated kmers with distance threshold of 0.3 (equals to correlation coefficient threshold of 0.7). The first kmer in a cluster will be kept as 'representative' and the rest were excluded from downstream process.

The second features selection step was done using permutated feature importance metrics from Random Forest method. We utilised a RandomForestClassifier package from scikit-learn and rfpimp package [238]. This process was run twice, with and without an additional column of random numbers, to introduce additional variability into the dataset and therefore, indicates the stability/consistency of the features. Only features with positive importance score in both iterations were included in the logistic regression model (Equation 4.1).

$$\text{logit}(P) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n$$

Equation 4.1. Logistic regression equation. Logit(P) is the log odds ratio of probability of an outcome (invasive phenotype) compared to probability of non-outcome (non-invasive phenotype). β_0 is the coefficient on the constant term, β_{1-n} are the coefficients on the independent variables (i.e. each valid kmer), and x_{1-n} are the independent variables, 1 if present, 0 if absent.

The logistic regression model was built using the "statsmodels" package (https://www.statsmodels.org/stable/index.html). The dataset (combination of dataset 1 and 2) was split into training (80%) and testing (20%) dataset, to fit and test the model. Metrics to evaluate the model included McFadden's Pseudo R^2, Akaike information criterion (AIC), area under the curve (AUC) score, and p-values of each feature in the model. After the initial model was generated, features with p-value of < 0.1 were excluded and backward elimination approach was implemented to find the best combination of features. This approach involved iterative removal of one feature at a time until the AIC score or accuracy of the model

increased or decreased, respectively [239]. In this process, a significance threshold of 0.05 was chosen as a basis of choosing the feature to remove. A confusion matrix and the receiver operating curve (ROC) were generated for the final logistic regression model (https://github.com/artmisk13/DPhil_thesis_Hinf).

4.2.6. Mapping kmers with predictive value

All k-mers included in the logistic regression model were considered predictive of the invasive disease phenotype. These kmers were mapped to the full-length gene from a NTHi reference genome using Geneious Prime software. In addition, corresponding protein regions that might be affected by nucleotide variations carried by the kmers were also visualised using AlphaFold 3-dimensional protein structure prediction [240]. Biological function of these regions was evaluated based on the information from InterPro database. Finally, all k-mers in non-CDSs and intergenic regions with predictive value were mapped to one of the NTHi reference genome used in this study.

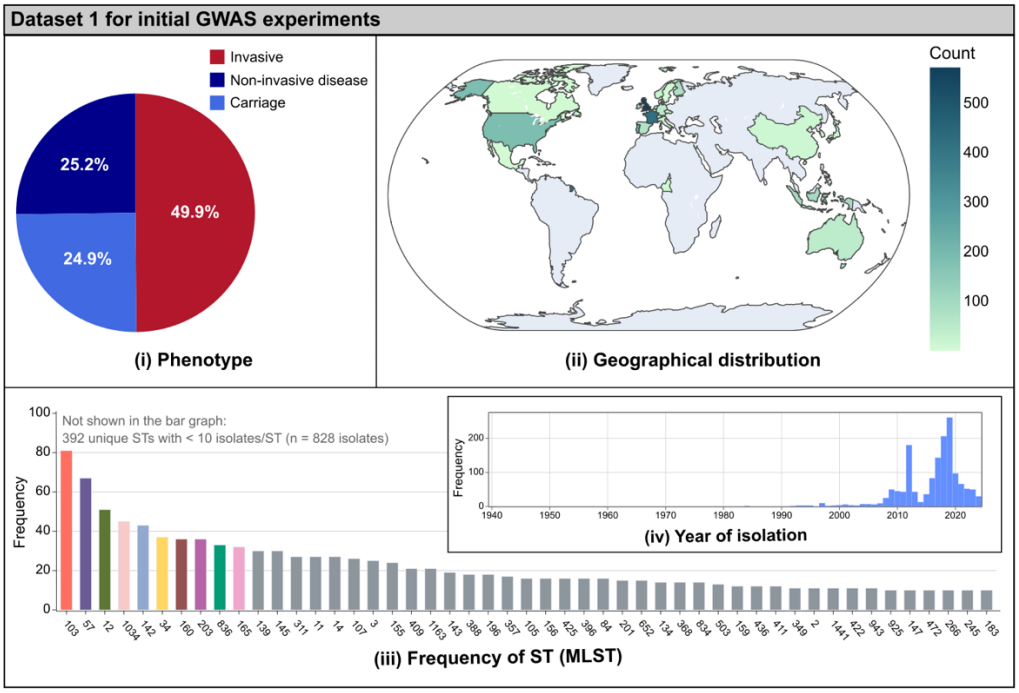
4.3. Results

4.3.1. Each GWAS dataset consisted of more than 1000 high quality, globally distributed non-typeable *H. influenzae* draft genomes

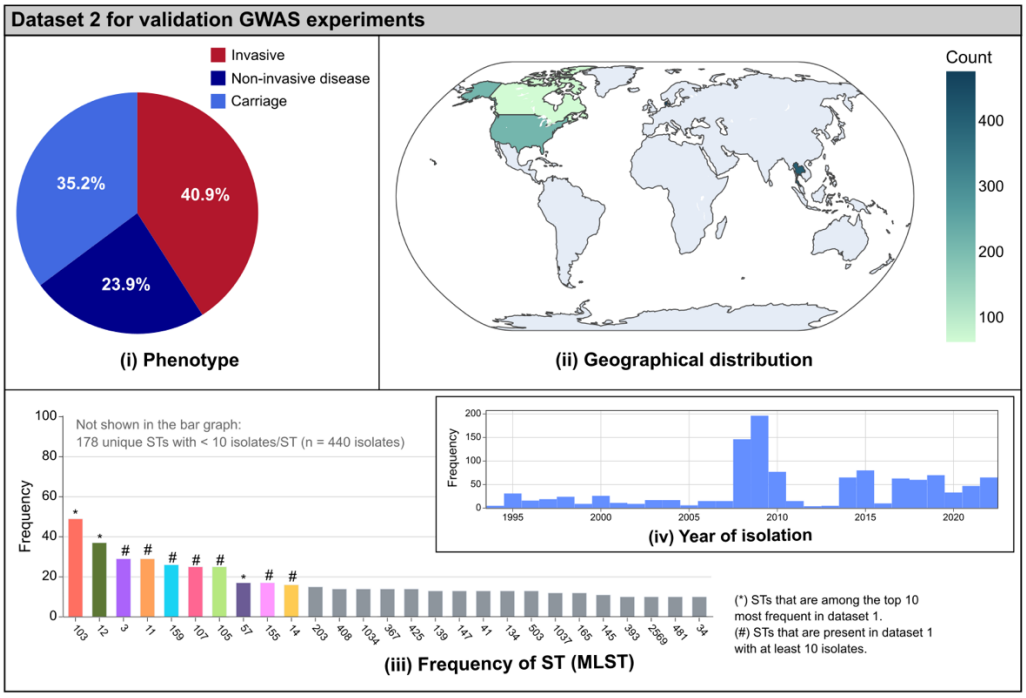
Dataset 1 consisted of 1975 NTHi genomes and originated from five different continents and 52 different countries (**Figure 4.13**a-ii, Supplementary Table 4.1); however, most of NTHi in dataset 1 were isolated in Europe (n = 1606, 81%) and only 5 isolates were from Africa. An equal percentage of genomes belonged to either the invasive (49.9%) or non-invasive (50.1%)

phenotype group. Within the non-invasive group, there was a similar number of NTHi isolated from disease and carriage (**Figure 4.13**a-i). According to 7-locus MLST, there were 441 unique STs in dataset 1, with 239 of them assigned to a single NTHi genome (**Figure 4.13**a-iii, Supplementary Table 4.1). This reflected the known diversity among NTHi population captured by the MLST scheme.

1165 NTHi genomes were curated in dataset 2. These genomes originated from four studies on *H. influenzae* population genomics in carriage and disease [67, 142, 241]. Each of these studies took place in four different countries, Denmark (n=477), Thailand (n=410), USA (n=214), and Canada (n=64) (**Figure 4.13**b-ii). The distribution of different phenotype group was similar in dataset 2; however, there were more genomes within the non-invasive (59.1%) phenotype group (**Figure 4.13**b-i). As many as 205 unique STs were assigned to NTHi genomes in dataset 2. The three largest STs in dataset 1 were in the top ten largest STs in dataset 2 (**Figure 4.13**b-iii, Supplementary Table 4.2).



(A)



(B)

Figure 4.13. Frequency of phenotype and STs, geographical distribution, and year of isolation of two NTHi datasets for GWAS experiment and model evaluation. (A) Dataset 1 (B) Dataset 2.

Several parameters were used to evaluate the genome assembly quality, including the number of contigs, genome length, N50, L50, and the percentage of GC content. NTHi genomes from the three different datasets were of high quality, with similar range and median of all the quality check parameters (**Table 4.6**).

Table 4.6. Summary of genome assembly quality parameters of the two NTHi datasets.

Quality parameters	**Dataset 1** Median (min-max)	**Dataset 2** Median (min-max)
Number of contigs	44 (1-1176)	111 (1-473)
Genome length	1,849,251 (1,634,229-2,433,349)	1,863,391 (1,758,333-2,168,449)
N50	134,476 (11,265-2,085,715)	117,352 (15,620-1,901,558)
L50	5 (1-49)	5 (1-40)
% GC	38 (37.68-42.19)	38.03 (37.73-38.59)

4.3.2. *H. influenzae* genomes from invasive and non-invasive group had a similar pangenome size.

The pangenome reconstruction tool identified a total of 7032 genes in the pangenome of NTHi in dataset 1. There was a small difference in the size of pangenome between NTHi genomes isolated from invasive, compared to non-invasive cases (**Figure 4.14**a). Core genes were defined as those present in more than 95% of the isolates, while cloud genes were identified as those present in less than 15%. Shell genes were classified as those present in 15% to 95% of the isolates [132]. While the difference in the number of core and shell genes was fewer than 25, NTHi from non-invasive group had 480 more genes in the cloud genome. Nevertheless, when the non-invasive group was divided further into “carriage” and “non-invasive disease” isolates, the total size of the pangenome was reduced, indicating that isolates

within the same phenotype group were more similar in the composition of their pangenome (**Error! Reference source not found.**b).

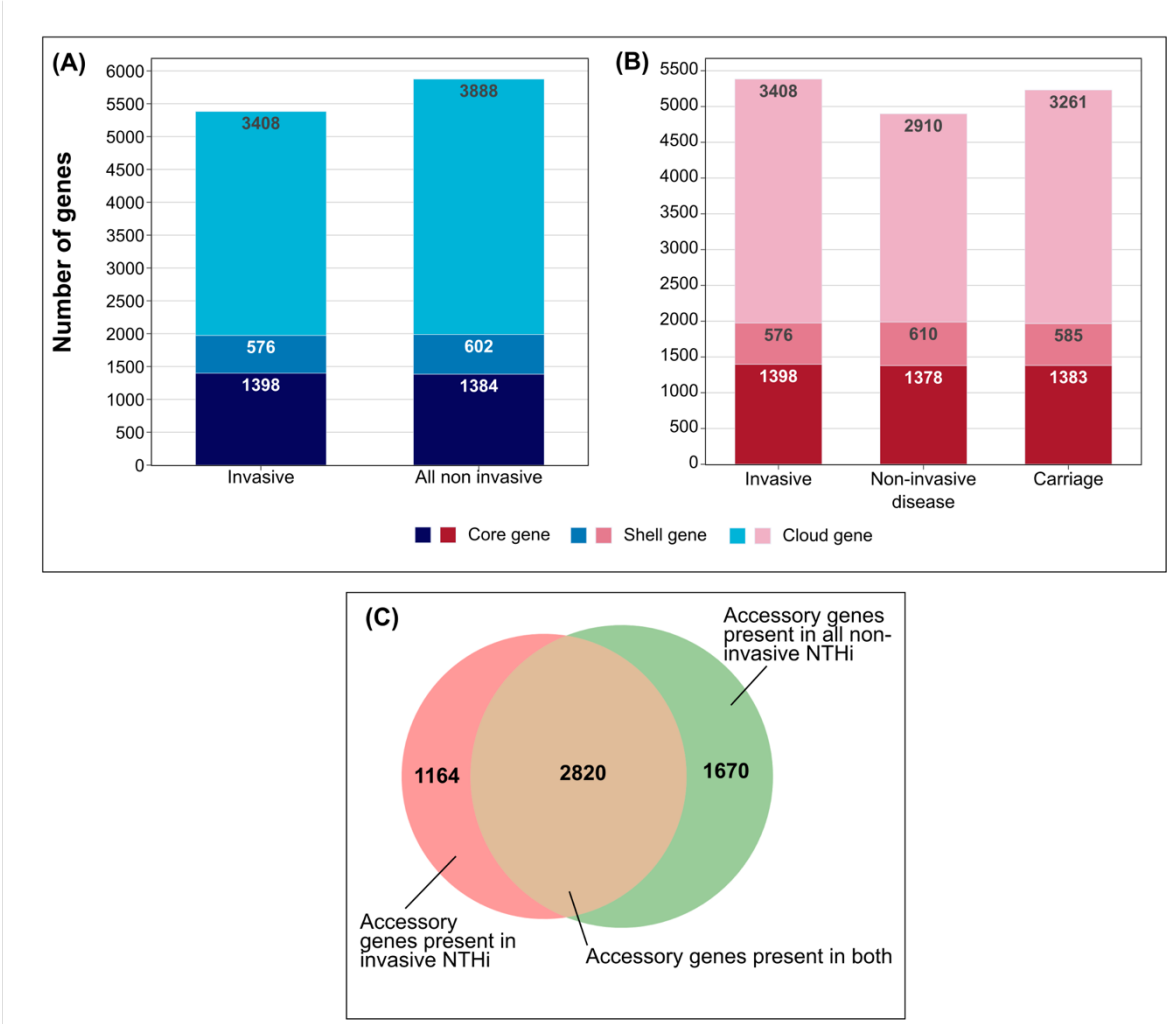


Figure 4.14. The pangenome size of NTHi in dataset 1 based on their phenotypic group. (A) Non-invasive disease and carriage phenotypes were combined into a single group ("all non-invasive"). (B) All non-invasive group was separated into non-invasive disease and carriage group. (C) Venn diagram of shared and non-shared accessory genes between invasive and "all non-invasive" NTHi. Core gene: genes present in > 95% of the NTHi isolates in dataset 1, shell gene: genes present in > 15% but < 95% of the NTHi isolates in dataset 1, cloud gene: genes present in < 15% of the NTHi isolates in dataset 1.

A total of 1378/1398 core genes (98.6%) in NTHi from invasive group were also detected as core in the non-invasive group (Supplementary Table 4.3). The remaining 20 genes were

still present in the non-invasive group, but as part of the accessory genome. Similarly, six genes that were core in the non-invasive group were classified as accessory genes in the invasive group. These findings highlight the known stability of *H. influenzae* core genes in terms of presence and absence, but do not rule out the possibility that different alleles of core genes may be associated with invasive potential. Among the 26 differentially present core genes between the two phenotype groups, more than half (n = 15) were genes involved in various bacterial metabolic pathways, including inorganic ion, carbohydrate, and coenzyme metabolism, as well as energy production.

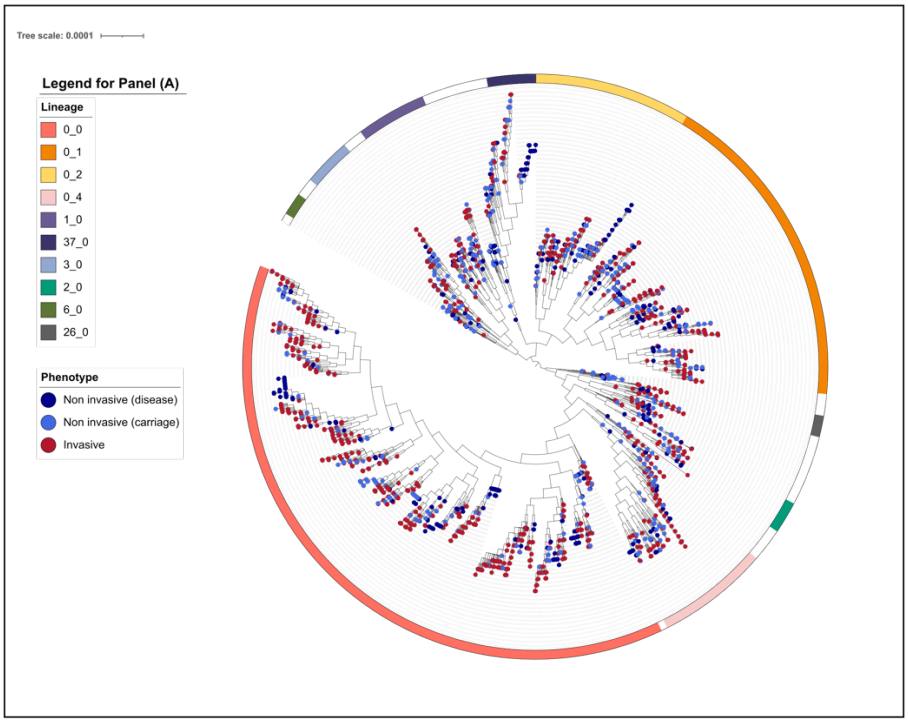
Collectively, 3984 genes were in the accessory genome of the invasive NTHi isolates and, unlike the core genes, only 2820 of them (70.8%) were also accessory in the non-invasive isolates (**Figure 4.14**c). As many as 1158 (29.1%) of invasive-associated accessory genes were not present, while 6 were present as core genes, in the non-invasive NTHi. Likewise, out of 4490 accessory genes among the non-invasive NTHi isolates, 1650 (36.8%) were not present and 20 were present as core genes in the invasive isolates. As many as 1511 (37.9%) and 1402 (31.2%) genes in the accessory genome of invasive and non-invasive isolates, respectively, did not have any functional annotation.

4.3.3. Recombination events were detected in most parts of NTHi genomes across different lineages, with comparable effects of recombination between two phenotypic groups.

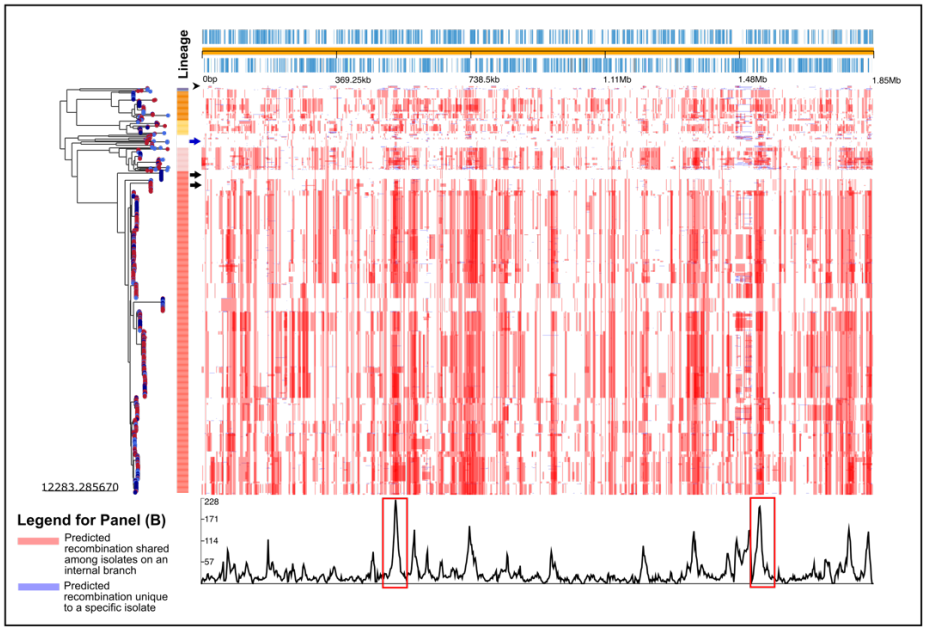
Using *H. influenzae* cgLIN superlineage partition level, there were 80 lineages among 1975 NTHi in dataset 1. Genomes belonging to the same lineage had at least 22.85% similarity in their core genome. 1702 (86.2%) genomes belonged to one of the ten largest lineages, as

annotated in the ML phylogeny (Figure 4.15a). However, as many as 62 lineages had less than 10 genomes as their member.

Isolates from both invasive and non-invasive phenotype groups are distributed throughout the phylogeny without clear clustering in specific lineages (Figure 4.15a). In the five largest lineages (each with over 50 genomes), invasive isolates comprise 40–56%, indicating both phenotypes are similarly represented. Some smaller lineages consist exclusively of either invasive or non-invasive isolates; however, the limited number of isolates in these lineages ($n < 5$) prevents any definitive conclusions about their association with invasiveness.



(A)



(B)

Figure 4.15. (A) The Maximum-Likelihood phylogeny of NTHi in dataset 1 using the core genome alignment. Tree nodes were coloured based on phenotype group and the circular color strip indicates different lineages as defined by cgLIN nomenclature. The lineage legend was sorted based on the number of genomes in descending order. **(B)** Recombination blocks in the NTHi. The reconstructed phylogeny was based on the whole-genome alignment to the NTHi reference genome. Predicted recombination events shared among isolates on an internal branch were coloured as red blocks and

those who are unique to specific isolate were colored as blue blocks. **Arrowhead:** distinct recombination pattern in lineage 1_0, **Arrows:** fewer recombination events in some subsets of the NTHi isolates, **Red squares:** Location in the NTHi genome with the most frequent recombination events.

Figure 4.15b illustrates the recombination blocks in NTHi genomes from dataset 1, based on whole-genome alignment to the NTHi reference genome. Recombination events are widespread across NTHi genomes; however, two distinct "hotspots" showed notably higher frequencies of recombination (Figure 4.15b, red squares). These hotspots were located within genes encoding a TonB-dependent receptor (TBDR) and a porin, with the surrounding regions also affected by recombination.

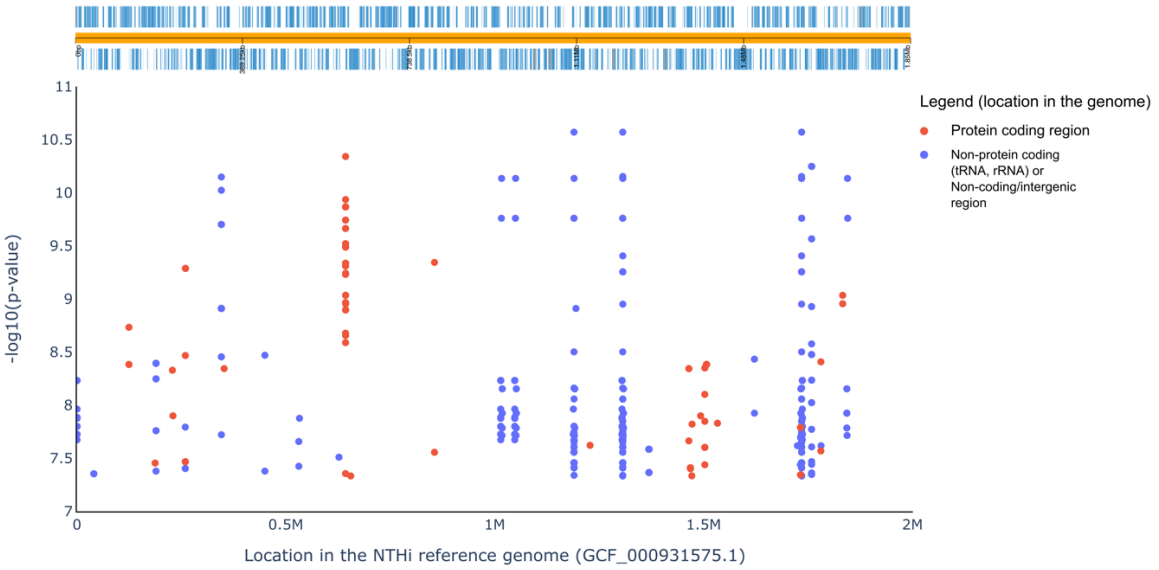
Overall, recombination in NTHi occurred without a consistent pattern linked to specific lineages. However, there were some exceptions. For instance, isolates within lineage 1_0 exhibited a distinct recombination pattern compared to the rest of the dataset (Figure 4.15b, arrowhead). Additionally, subsets of isolates in lineage 0_0 displayed fewer recombination events (Figure 4.15b, black arrows). A similar trend was observed for isolates belonging to smaller lineages (Figure 4.15b, blue arrow). In contrast, lineages other than 0_0 showed more frequent recombination in the terminal branches. These observations confirm that NTHi undergoes extensive recombination, which likely facilitates the horizontal transfer of genetic variants associated with traits such as invasiveness.

Two parameters of recombination, R/theta and r/m, were available for 1494 (75.6%) isolates. No single substitution was found for the rest of isolates. The two recombination parameters were compared between invasive and non invasive group. Due to non-normal distribution of the data (Appendix 4.2) with Kolmogrov-Smirnov pvalue < 0.001, the comparison was done using Mann Whitney U test. There was only a small difference (p value

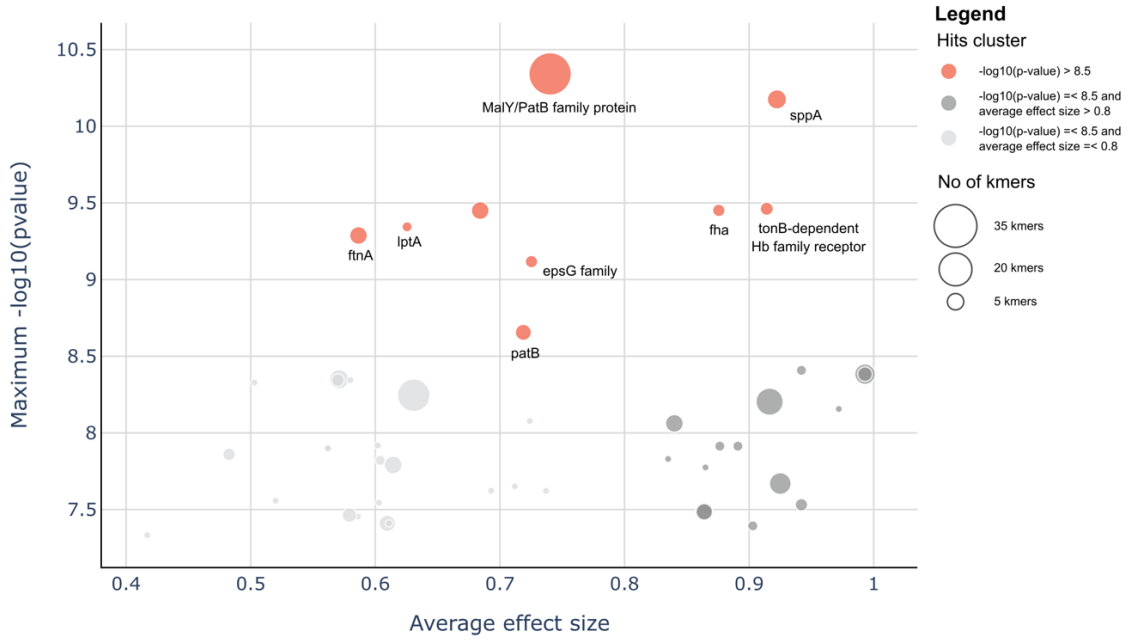
< 0.001) in R/theta between the two phenotype groups, with the non-invasive slightly higher than the invasive group with the median (minimum – maximum) value of 0.468 (0.26-5.67) and 0.428 (0.18-2.19), respectively. However, there was a more pronounced difference (p value < 0.001) in r/m between those groups, 5.64 (1.74-55.28) for non-invasive and 4.56 (1.75-57.2) for invasive group. This finding indicated that even though the rate of recombination/mutation in both groups were similar, recombination had relatively higher effect compared to mutation in non-invasive group.

4.3.4. Genetic variants associated with invasive phenotype group in the initial GWAS were mostly detected in the accessory genome of NTHi

All kmers identified for each of the 200 parallel GWAS runs were filtered and only those with adjusted p-value of less than the Bonferroni corrected threshold for that run were processed further. Collectively, there was a total of 622 unique significant kmers and all of them were annotated to either the reference genome or draft genomes in the dataset 1. 191 kmers (30.7%) were annotated to 47 protein-coding sequence (CDS) and 33 of these CDS were in the accessory genome. The rest of significant kmers (n = 431, 69.3%) were in non-protein coding, such as tRNA or rRNA genes, or intergenic regions (**Figure 4.16**a).



(A)



(B)

Figure 4.16. Annotation of significant kmers associated with NTHi invasiveness to a reference and high-quality draft genomes. **(A)** Manhattan plot of kmers annotated to the NTHi reference genome (NCBI Assembly: GCF_000931575.1). **(B)** Protein-coding sequence (CDS) in NTHi genomes with detected genetic variants associated with invasiveness.

Kmers annotated to CDS are summarised in **Figure 4.16**b. Each circle represents one CDS and its size is proportional to the number of kmer hits found in that CDS. There were 9 CDS with the most significant kmers (p-value < $3.17e^{-9}$) and the gene names, their products and functions were detailed in **Table 4.7**. The *ftnA* gene exhibited the lowest average effect size of 0.59, indicating that the presence of k-mer variants in this gene increased the probability of an NTHi genome being invasive by 59%. The complete list of regions in the NTHi genomes to which significant kmers annotated to are available from the Supplementary Table 4.4.

Table 4.7 Protein-coding genes with the most significant kmers associated with NTHi invasive phenotype.

Gene name / Preferred name	Protein accession / Gene family ID[a]	Product	Function	COG[c]
***malY/patB* family**	WP_011272389.1	MalY/PatB family protein	catalyze the conversion of cystathionine to homocysteine	E
sppA	*sppA*	signal peptide peptidase	to cleave the remnant signal peptides left behind after protein secretion and cleavage by signal peptidases	OU
ftnA	*ftnA*	non-heme ferritin	iron-storage protein	P
lptA	WP_042594392.1	phosphoethanolamine transferase	modification of lipid A with phosphoethanolamine cellulose	S
-	2463_01794	-	a family of proteins that are present in crAss phage[b]	-
***epsG* family**	2463_00147	EpsG family protein	involved in biofilm formation	S
patB	*patB*	Aminotransferase class I and II	catalyze the transfer of amino groups between amino acids and keto acids	E
fha	3292_01713	protein with haemagglutination activity domain	a key component for bacterial adhesion to host cells	UW
Ton-B dependent family receptor	4400_01531	ton-B dependent family receptor	capture and transport iron bound to iron-carrying proteins (hemoglobin, transferrin, and lactoferrin) across the bacterial cell wall	P

a) Gene ID: PIRATE pangenome tool assigns gene ID for each genome.

[b] Functional annotation using eggnog-mapper did not yield any positive result. Function was determined through sequence search in Pfam database.

[c] COG: cluster of orthologous group. E: Amino acid transport and metabolis; O: Posttranslational modification, protein turnover, chaperones; U: Intracellular trafficking, secretion, and vesicular transport; P: Inorganic ion transport and metabolism; S: Unknown function; W: Extracellular structure.

4.3.5 Validation GWAS confirmed 16 genetic regions associated with invasive NTHi.

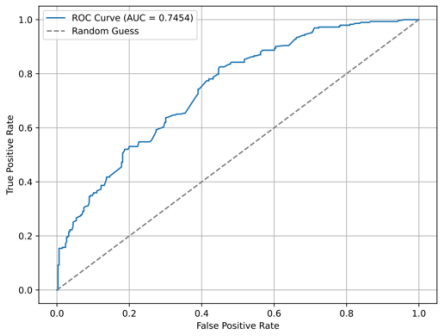
Through the validation GWAS, collectively there were 4294 kmers from 6 GWAS runs in 6 lineages. The NTHi lineage with most kmers (n = 3477) detected was lineage 0_4, while within lineage 3_0 there was 0 identified kmer associated with invasive NTHi. These kmers were annotated to the same set of reference and high-quality draft genomes in initial GWAS to allow for compatible comparison between the two GWAS stages. A total of 5 protein-coding and 11 non-CDS regions in NTHi genomes were consistently associated with the invasive phenotype in both the initial and validation GWAS. Across these regions, 256 unique k-mers were identified, representing genetic variants linked to NTHi invasive disease. To determine valid k-mers, the presence of these kmers was systematically evaluated in NTHi genomes in both dataset 1 and 2, resulting in 159 kmers which had positive association with invasive disease. Eventually, 146 of them had statistically significant p-values in the chi-square test and were advanced for further evaluation in logistic regression analysis.

Best Model Summary:
Generalized Linear Model Regression Results

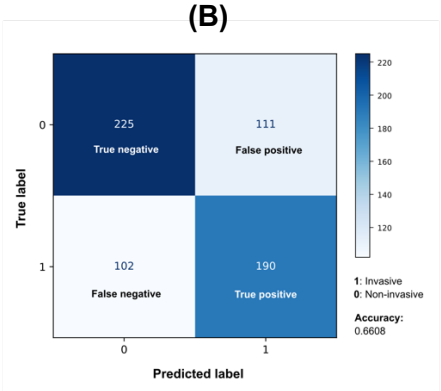
Dep. Variable:	y	No. Observations:	2511
Model:	GLM	Df Residuals:	2491
Model Family:	Binomial	Df Model:	19
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1471.8
Date:	Tue, 04 Feb 2025	Deviance:	2943.6
Time:	23:53:26	Pearson chi2:	4.42e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.1889
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.1688	0.213	-14.885	0.000	-3.586	-2.752
cds-WP_044331820.1_1;init	1.7188	0.189	9.086	0.000	1.348	2.090
cds-WP_005656589.1_1;val;0_1	0.5306	0.101	5.271	0.000	0.333	0.728
sixA_19;val;0_1	-0.2330	0.100	-2.332	0.020	-0.429	-0.037
lldD_148;init	0.3036	0.105	2.894	0.004	0.098	0.509
sixA_6;init	2.7060	0.452	5.990	0.000	1.821	3.592
g00001_1;3289_01621_7;init	0.3521	0.119	2.956	0.003	0.119	0.586
ftnA_20;val;0_1	0.4298	0.100	4.296	0.000	0.234	0.626
lldD_197;init	0.6947	0.151	4.603	0.000	0.399	0.990
lldD_338;init	0.3113	0.116	2.675	0.007	0.083	0.539
rimP_314;init	1.6334	0.449	3.638	0.000	0.753	2.513
cds-WP_044332714.1_3;init	0.5739	0.178	3.232	0.001	0.226	0.922
cds-WP_044332714.1_43;val;0_1	0.3725	0.103	3.609	0.000	0.170	0.575
cds-WP_044332714.1_26;val;0_1	0.5381	0.141	3.803	0.000	0.261	0.815
rimP_370;init	1.1353	0.494	2.298	0.022	0.167	2.104
g00001_1;hgpA_2_2;val;0_4	0.2979	0.175	1.705	0.088	-0.044	0.640
tdeA_372;init	1.6745	0.516	3.244	0.001	0.663	2.686
cds-WP_044331820.1_260;init	-0.3400	0.127	-2.667	0.008	-0.590	-0.090
g00001_1;4400_01531_8;init	0.6356	0.304	2.094	0.036	0.041	1.231
cds-WP_044333040.1_2;init	0.7737	0.351	2.205	0.027	0.086	1.462

(A)



(B)



(C)

Figure 4.17. (A) Summary of the best logistic regression model using backward elimination. **(B)** Receiver operating characteristic (ROC) curve of the best logistic regression model across multiple thresholds. **(C)** Confusion matrix of phenotype prediction made by the model in the test dataset, using probability threshold of $\geq$ 0.5.

Kmers were grouped based on the genetic regions they were annotated to and checked for multicollinearity by constructing a correlation matrix for each group. After clustering based on the correlation coefficient, only 29 kmers were retained (Supplementary Table 4.5). Subsequently, according to feature importance score measures using permutation approach (Supplementary Table 4.5), 25 were included in the construction of the logistic regression model. The final model was constructed using an iterative backward elimination approach to identify the optimal model with the lowest AIC or highest accuracy. The best model, with an AIC of 2983.5, predicted whether genomes in the test dataset originated from invasive

infection cases with 66.1% accuracy (Figure 4.17c). The test dataset comprised a randomly selected 20% of the combined datasets 1 and 2. The prediction results were mapped onto the phylogenetic trees of datasets 1 and 2, revealing both correct and incorrect predictions throughout the phylogeny (Appendix 4.4). he final logistic regression model included a total of 19 predictive k-mers, which were annotated to 3 CDSs, 2 tRNA genes, 2 rRNA genes, and 6 intergenic regions (

Table 4.8).

Table 4.8 Valid kmers in the final logistic regression model and their location in the NTHi genome.

Gene/Preferred name	Protein accession/Gene ID[a]	kmer names[b]	Product	COG[c]
Protein-coding genes				
porin	WP_044332714.1	cds-WP_044332714.1_26;val;0_1 cds-WP_044332714.1_3;init cds-WP_044332714.1_43;val;0_1	porin	I
hgpA	3289_01621 4400_01531	g00001_1;3289_01621_7;init g00001_1;4400_01531_8;init g00001_1;hgpA_2_2;val;0_4	• hemoglobin (Hb) or Hb/haptoglobin receptor	P
ton-B dependent family receptor	WP_044333040.1	cds-WP_044333040.1_2;init	• ton-B dependent family receptor	
sixA	WP_005652786.1	sixA_6;init	phosphohistidine phosphatase	T
Non-protein-coding genes				
tRNA-Met	NTHI477_RS01680	rimP_314;init rimP_370;init	tRNA-Methionine	-
tRNA-Glu	NTHI477_RS05660 NTHI477_RS06230 NTHI477_RS08480	cds-WP_044331820.1_1;init	tRNA-Glutamate	-
23s rRNA	NTHI477_RS06225 NTHI477_RS00005 NTHI477_RS04860 NTHI477_RS05025 NTHI477_RS08485	lldD_148;init lldD_197;init	23s ribosomal RNA	-
16s rRNA	NTHI477_RS08475	lldD_338;init	16s ribosomal RNA	-
Intergenic region				
Between WP_005656589.1 and WP_224056840.1		cds-WP_005656589.1_1;val;0_1 tdeA_372;init	-	-
Between NTHI477_RS08480 and NTHI477_RS08485		cds-WP_044331820.1_260;init	-	-
Between WP_005650594.1 and WP_074031221.1		ftnA_20;val;0_1	-	-
Between WP_044331820.1 and WP_044331826.1		lldD_338;init	-	-
Between NTHI477_RS06235 and WP_042593485.1		lldD_338;init	-	-
Between WP_005652786.1 and WP_038441256.1		sixA_19;val;0_1	-	-

a) Gene ID: PIRATE pangenome tool assigns gene ID for each genome.
b) Kmer names were arbitrary, as a reference to compare with the logistic regression model summary from Figure 4.17. Some kmers were annotated in multiple non-protein-coding genes and/or intergenic region.
c) COG: cluster of orthologous group. I: Lipid transport and metabolism; P: Inorganic ion transport and metabolism; T: Signal transduction mechanisms.

The 19 kmers included in the final logistic regression model were independent of each other and demonstrated significant associations (p-value < 0.1) with invasive NTHi phenotype. The McFadden pseudo-R^2 was 0.19, close to 0.2, a threshold which represented an excellent model fit while a value close to 0 indicated that the model did not have predictive value [242]. Eight kmers annotated to a protein coding region was mapped to the complete CDS and visualised in the 3D protein structure using AlphaFold (**Figure 4.18**). On the other hand, the 11 kmers in non-protein coding sequences or intergenic regions were aligned to the complete *H. influenzae* reference genome and visualised as the circular genome (Appendix 4.3).

Two out of three products of the CDSs were located in the *H. influenzae* cell outer membrane. One belonged to a porin superfamily which functions as channels that share a beta-barrel structure to outer membrane protein commonly found in Gram-negative bacteria but different in strand and shear number [243]. The three kmers associated with NTHi invasive disease were located in the beta-barrel chain. The other protein was a receptor for hemoglobin (Hb) and Hb-haptoglobin complex from the human host, which is crucial for *H. influenzae* for heme uptake across the bacteria cell wall. This protein belongs to a large and highly diverse family of TonB-dependent receptors (TBDRs) characterized by two distinctive motifs—the TonB box and the TonB C-terminal box—as well as two structural domains: the TBDR plug and the beta-barrel [243]. As depicted in **Figure 4.18**c, the variable region of this receptor is primarily located within the TBDR domain, where two k-mers associated with NTHi invasive disease were identified. Nevertheless, one kmer was mapped to the TBDR plug and the other in the signal peptide, which tags the mature protein to the correct cellular location [243].

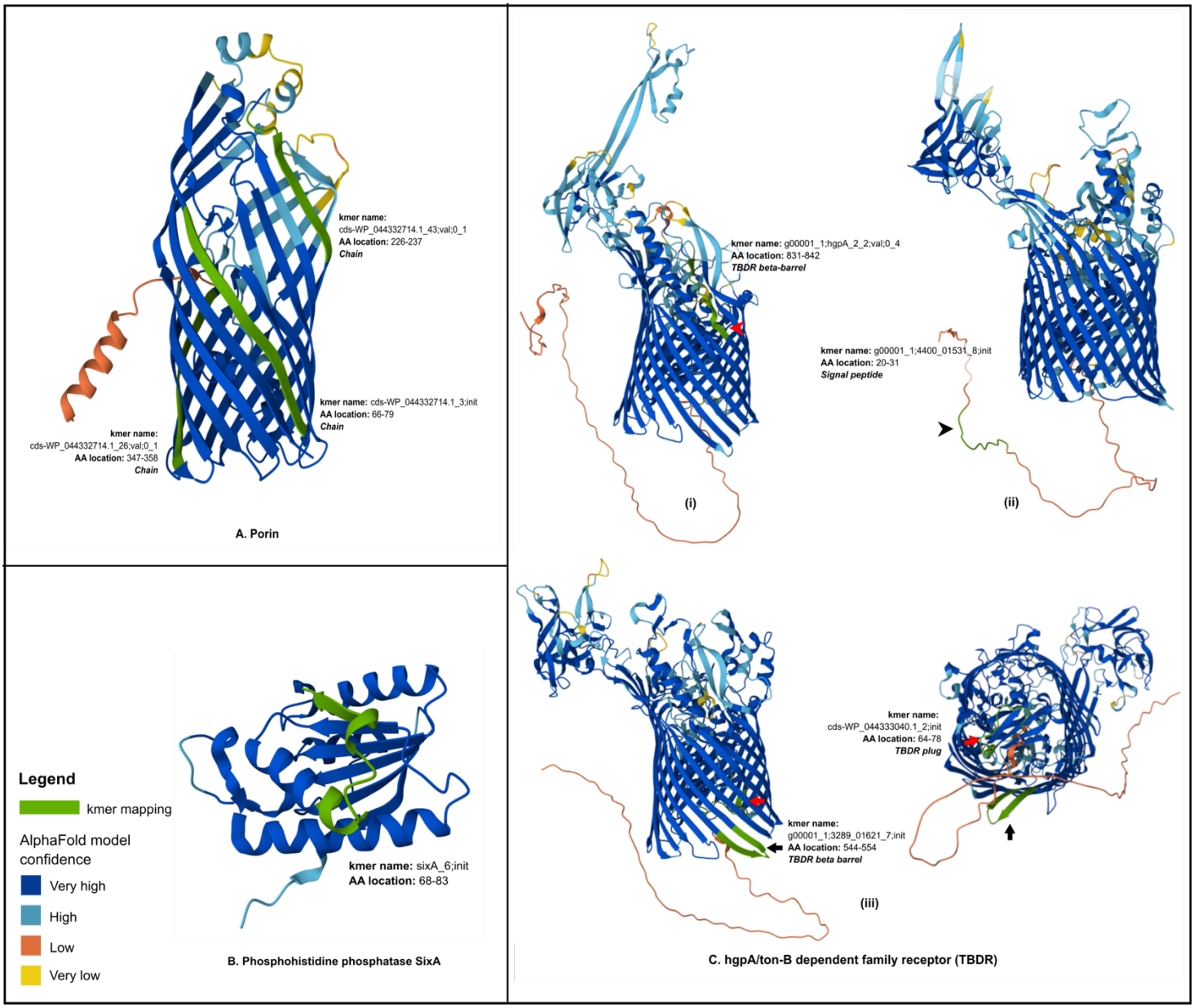


Figure 4.18. Regions in the coded proteins affected by nucleotide sequence variants associated with NTHi invasive phenotype identified through the GWAS pipeline. All 3-Dimensional protein model was predicted by AlphaFold. (A) Gram-negative porin superfamily (Uniprot accession: A0A2S9S1T1). (B) Phosphohistidine phosphatase (Uniprot accession: P44164). (C) Ton-B dependent receptor, with different allelic variants and protein conformation, each with nucleotide similarity percentage of 45-49% to one another: (i) Uniprot accession: Q9ZA21, (ii) Uniprot accession: A0A7H5FVA, (iii) Uniprot accession: Q9KIV1.

4.4. Discussion

The rising number of invasive disease cases due to NTHi necessitates a deeper understanding of the factors that enable NTHi to invade the bloodstream. A key step toward this goal is identifying genetic variants associated with the invasive phenotype. In our study, we addressed this question by performing comprehensive comparative genomic analyses using a k-mer-based GWAS, supported by detailed population structure characterisation and enriched with recombination analyses.

Proteins facilitating immune evasion and mucosal penetration are key to invasive disease. In this study, accessory genome of invasive NTHi was only 3.5-13.2% larger than those isolated from non-invasive disease or carriage. However, previous studies in other bacterial species have not conclusively established a link between pathogenicity and genome size. This suggests that evaluating the specific content and function of accessory genomes may be more informative than merely assessing their size [244, 245].

Our findings suggest that the larger accessory genome size in invasive NTHi may be due to the presence of additional virulence-related genes. However, the identification of accessory genes unique to each group implies that these genes have specific functions that facilitate bacterial adaptation to distinct niches. The current GWAS was conducted using a reference-free, k-mer-based approach, which allows for the identification of diverse genetic determinants associated with specific phenotypes. This includes the presence or absence of genes, allelic variations caused by single nucleotide polymorphisms (SNPs), and insertion-deletion events [224]. Furthermore, this method can also detect sequence variations in non-coding regions [225].

Previous studies have suggested that *H. influenzae* has an open pangenome [54, 61, 93], primarily due to its natural competence for taking up naked DNA from the environment, as well as other forms of horizontal gene transfer [96, 244]. This capability significantly contributes to the high genetic diversity observed within the *H. influenzae* population, particularly among NTHi strains, as demonstrated by the recombination analyses in this study. Recombination events were detected across nearly all regions of the NTHi genome, with two distinct recombination hotspots identified. These findings will be explored in greater detail alongside the results of the GWAS experiments.

Smaller groups within an NTHi lineage in our study exhibited distinct recombination patterns. Recombination dynamics often vary among lineages of the same bacterial species [246], as observed in other human pathogens such as *Listeria monocytogenes* [247] and *Neisseria meningitidis* [248]. This variability can be partly explained by sampling biases in the natural population, as well as the degree of ecological specialisation of specific lineages. Adaptation to new ecological niches can either increase or decrease the rate of recombination [249]. For instance, recombination is virtually absent in the highly specialised human pathogen *Mycobacterium tuberculosis*, whereas pathogenic lineages of *Escherichia coli* with enhanced virulence exhibit higher recombination rates [249]. Our findings suggest a slightly lower rate and impact of recombination in invasive NTHi. It is possible that recombination enables NTHi to exchange genetic variants that enhance invasiveness; however, once strains adapt to an invasive lifestyle, the opportunity for recombination may diminish, leading to a reduced recombination rate due to limited interactions with other lineages.

Just over 600 unique k-mers in NTHi were significantly associated with the invasive phenotype in the initial GWAS, each showing an average effect size greater than 0.35. This

total is considerably lower than the typical number of significant k-mers detected in GWAS studies of other human pathogens [250, 251], with the most significant hits and largest effect sizes mapping to genes encoding proteins that are plausibly linked to invasiveness. One key factor contributing to this outcome is the accurate lineage definition provided by cgLIN [179] which effectively corrected for population structure—a major challenge in microbial GWAS, where population structure can lead to false-positive associations between genetic variants and phenotypes [224, 225]. Another reason for the observed results is the use of a sub-sampling approach in the initial GWAS, where a randomly selected subset of 200 genomes was analysed. This strategy enhances the reliability of the findings by focusing on associations that are consistently detected across multiple subsamples, making them more likely to be true positives [252, 253]. Additionally, a report by Coll *et al.* on the highly recombinogenic human pathogen *Enterococcus faecium* suggested that a sample size of 200 is sufficient to detect variants with moderate effect sizes (OR = 0.5) that are present in at least 10% of the population [254]. NTHi exhibits a recombination level so high that it reduces overall population diversity [255] and increases the likelihood that alleles carrying specific traits will be commonly present.

Another common cause of false-positive associations in microbial GWAS is the issue of multiple testing. To mitigate this, the significance threshold for associations was determined using Bonferroni correction, and validation GWAS was performed in an independent cohort [224, 236]. All unique k-mers that were positively associated with invasive disease in both the initial and validation GWAS were assessed for their predictive value in a logistic regression model aimed at predicting NTHi invasiveness. The model achieved an AUC of 74.54% and an accuracy of 66%, based on the presence or absence of 19 independent k-mers, excluding those with evidence of co-correlation. Notably, all excluded highly correlating k-mers were

overlapping within the same intra- or intergenic regions, indicating an absence of linkage disequilibrium suggestive of epistatic interactions [256]. The false-positive and false-negative predictions made by the model can be partly attributed to the inherent limitations of the cross-sectional design of this GWAS study, where most NTHi genomes were sequenced only once per subject, providing a snapshot in time. An NTHi strain with the potential for invasiveness might have been sequenced at a point when it was causing only a non-invasive infection. Additionally, host susceptibility plays a crucial role: an NTHi strain with invasive potential may fail to invade the bloodstream if the host's immune system is particularly robust, resulting in false-positive predictions Conversely, a strain considered "benign" might still invade the bloodstream if a predisposing condition, such as COPD, compromises the respiratory tract mucosal barrier [257], leading to a false-negative prediction. Nevertheless, previous bacterial GWAS studies on human pathogens that constructed prediction models for complex phenotypes based on GWAS results reported AUCs ranging from 58% to 84%, demonstrating that the AUC achieved in this study is within the expected range for predictions based solely on bacterial factors [250, 258, 259].

The 19 k-mers with predictive value in the logistic regression model for predicting invasive phenotype were mapped to 3 CDSs, 2 tRNA genes, 2 rRNA genes, and 6 intergenic regions. All 3 CDSs were part of the core genome of NTHi. Three of these k-mers were annotated to a gene encoding a porin superfamily protein, and when translated into the corresponding amino acid sequences, they were non-overlapping and located in three distinct regions of the beta-barrel structure. This protein functions as an aqueous channel that facilitates the diffusion of small hydrophilic molecules across the outer membrane of Gram-negative bacteria. Loop 3 of the beta-barrel forms a constriction point, allowing the passage of molecules with a molecular mass typically below 600 Da, including sugars, amino acids,

and certain antibiotics such as beta-lactams, tetracyclines, chloramphenicol, and fluoroquinolones [243, 260]. A specific type of porin previously identified in NTHi, OmpP1, has been predicted to bind to arachidonic acid (AA), an inflammatory modulator with bactericidal effects. Loss-of-function mutations in the AA docking site of OmpP1 have been associated with resistance to this effect [261]. Although none of the three variants detected in our GWAS were located directly within the AA docking sites, one was in close proximity, while the other two were situated in the beta-barrel segment, potentially facilitating hydrophobic interactions essential for trimerisation [243, 262]. The trimeric assembly has been shown to be crucial for function in the well-studied porin OmpF from *E. coli* [263]. Therefore, it is plausible that allelic variations in the gene encoding the porin superfamily, which are associated with NTHi invasiveness, may lead to loss of function or reduced stability of the protein, thereby enhancing resistance to immune responses or even antibiotic action.

Four genetic variants associated with NTHi invasive phenotype were mapped to three genes encoding proteins with the same function, a Hb and Hb-haptoglobin TBDR. Proteins with this function are involved in diverse heme acquisition systems and do not necessarily share high sequence identity. *H. influenzae* possesses three major iron/heme acquisition systems: the Hb-haptoglobin binding proteins (*hgp*) [221], the heme-haemopexin (*hxu*) system [264], and the transferrin-bound iron protein (*tbp*) system [265]. All three were part of the core iron/heme modulon for the species, which was preferentially expressed when iron/heme source is scarce [265]. Consequently, it is common for *H. influenzae* isolates to carry multiple genes encoding TBDRs. For instance, the complete NTHi genome used as the reference in this study (NCBI accession GCF_000931575.1) contains seven genes encoding TBDRs, with amino acid sequences ranging from 714 to 998 residues. The three genes associated with the kmer variants were, however, all encoded TBDR within the *hgp* acquisition system. A previous

report characterised the diversity of *hgp* gene family, dividing it into seven groups with approximately 50% pairwise sequence identity [221]. Consistently, the three genes shared 45-49% amino acid sequence identity among them.

Two *hgp* genes, *hgpB* and *hgpC*, are known to be phase-variable, as indicated by the presence of a simple DNA sequence repeat tract, $CCA_{(n)}$, in the 5' region [221]. This mechanism allows *H. influenzae* to regulate the expression of these two TBDRs in a biphasic manner, enabling adaptive responses to environmental changes [221, 266]. Previous fragment length analyses showed that nearly 60% of invasive NTHi isolates had either *hgpB* or *hgpA* in a phase-varied ON state [221]. This DNA repeat translates into an XPTX tetrapeptide repeat of unknown functional significance, although structurally, it is located in the signal peptide region of the TBDR [243]. Among the three *hgp* genes identified in our GWAS experiments, two contained the in-frame repeat tract, albeit of different lengths (**Figure 4.18**c(ii)-(iii)). Notably, one k-mer variant was located within the $CCA_{(n)}$ tract (**Figure 4.18**c(ii)), reinforcing its potential role in the pathogenesis of invasive NTHi. The other two variants were mapped to repeat-containing *hgp* genes but in different regions: one in the TBDR beta-barrel within the periplasmic space and another in the TBDR plug domain. In its resting state, the plug domain blocks the barrel and specifically binds iron chelates, such as siderophores, on the extracellular side. Upon binding, the plug undergoes a conformational change that permits the passage of the chelate through the barrel [267]. The last variant was mapped to a TBDR lacking the tetrapeptide repeat, located in the extracellular-facing region of the beta-barrel. The amino acid sequences of surface-exposed regions of TBDRs encoded by *hgp* genes displayed high variability, which likely results from accumulated point mutations driven by strong selective pressure [221].

Seven out of eight intragenic variants in associated with invasive NTHi infection were located within two genes encoding highly immunogenic surface-exposed proteins (i.e. porin and TBDR). The accessibility of these proteins to the immune system imposes constant selective pressure on NTHi, leading to highly variable regions within these genes. Mutations that enable bacteria to evade host immune responses are positively selected, increasing the likelihood of survival and contributing to the invasive NTHi phenotype. Moreover, mutation is not the sole driver of genetic variation in these genes. Recombination analyses of the NTHi population revealed two recombination hotspots, both encompassing the porin and TBDR genes. The exchange of genetic material through recombination may allow variants associated with invasiveness to spread to other strains, which helps explain the lack of a clear association between the invasive phenotype and the bacterial phylogeny. Despite their variable expression and high diversity in certain regions, both porin and *hgp* are part of the core genome and encode highly immunogenic proteins, making them strong vaccine candidates if carefully designed. Targeting immune-accessible yet conserved regions of these proteins could enhance vaccine efficacy [221, 268]. As an example, the *N. meningitidis* serogroup B vaccine includes the phase-variable protein NadA and highly diverse porins as components of its outer membrane vesicle formulation [269]

The last kmer variants associated with invasive NTHi infection was mapped to the *sixA* gene which encodes for a phosphohistidine phosphatase. This enzyme is part of the nitrogen phosphotransferase system (PTS) which transfers a phosphate moiety from phosphoenolpyruvate to a histidine protein homologue, NPr. This system works in parallel with the sugar PTS counterpart, but exerts downstream regulatory functions through phosphate transfer [270, 271]. At the time of the writing, cellular process which is under the control of nitrogen PTS in *H. influenzae* has not been reported; however, in *E. coli*, it regulates

potassium transport [270] and the absence of *sixA* gene resulted in growth defect [271]. As this gene is part of the core genome of NTHi, the regulatory functions of this phosphatase are likely crucial for the bacterium's survival, with certain variants potentially providing a fitness advantage, such as enhancing survival in hostile environments.

Five k-mers identified in this study were mapped to two tRNA genes and two rRNA genes. tRNA genes encode tRNA molecules that carry specific amino acids and possess anticodons complementary to corresponding codons in mRNA. In rapidly dividing bacterial species, multiple copies of tRNA genes are often present to accelerate translation and, consequently, cell division [272]. The two NTHi tRNA genes associated with invasive phenotypes, tRNA-Met and tRNA-Glu, are each present in triplicate in the *H. influenzae* reference genomes. However, our GWAS findings revealed a difference between them. Two k-mer variants were mapped exclusively to one of the tRNA-Met genes (GenBank ID: NTHI477_RS01680 or LK401_RS08200), with one variant at the 5' end and the other at the 3' end of the sequence, the latter being the amino acid attachment site. The k-mers did not map to the other two tRNA-Met genes because of their high sequence variability relative to each other, except for the anticodon sequence, which remains conserved since methionine is encoded by only one codon. A previous study on *Pseudomonas fluorescens* demonstrated that in hostile environments, compensatory duplication of a tRNA gene can occur without point mutations in any of the copies [273, 274]. One possible explanation for the observed patterns in the NTHi population is that the additional tRNA-Met copies arose through horizontal gene transfer (HGT) rather than duplication, with one copy providing a translational efficiency advantage that facilitates bacterial survival and invasiveness in the human bloodstream. Alternatively, the duplication of the tRNA-Met gene followed by mutations could explain

these findings, as has been observed in the human pathogen *Burkholderia thailandensis* as a mechanism for infection persistence [275].

In contrast, one k-mer variant was mapped to all three tRNA-Glu genes, which are identical to each other in the NTHi reference genome (Appendix 4.5). This finding raises the question of whether the absence of the k-mer indicates the actual absence of known tRNA-Glu genes. To investigate this, a BLAST search of the complete tRNA-Glu gene sequence was performed against all datasets in this study, revealing that isolates lacking the k-mer variant did not have any detectable tRNA-Glu genes. This could be due to the high stringency of tRNA gene detection methods, which may fail to identify unrecovered tRNA-Glu genes that exist in nature. However, if we exclude this possibility, the absence of tRNA-Glu in some NTHi isolates might be explained by functional replacement through alternative mechanisms, such as modified wobble bases, which enable other non-glutamate tRNAs to recognize multiple codons [276].

16s and 23s rRNA were the two rRNA genes with identified variants associated with invasive NTHi and they are an essential part of protein synthesis ribosomal machinery [277]. No previous report has evaluated variants in these rRNA genes in association with the invasive potential of a bacterial species. Nevertheless, mutations in the 16s and 23s rRNA had been known to be the underlying mechanism of tetracycline and macrolides resistance, respectively, in many bacterial pathogens such as *E. coli* [277, 278], *N. gonorrhoeae* [279], *Mycoplasma genitalium* [280], and *M. bovis* [281]. While higher risk of invasive infection has been associated with antimicrobial resistance, in general, these two traits evolved independently [282].

4.5. Conclusion and future directions

In conclusion, our kmer-based, reference-free GWAS approach has identified a range of genetic determinants associated with invasive disease among NTHi isolates. Our findings suggest that invasive NTHi strains tend to possess a slightly larger accessory genome, potentially harbouring extra virulence-related genes that facilitate bloodstream invasion and immune evasion. Notably, our analyses identified significant associations between unique k-mers and genes encoding porin superfamily proteins, TBDRs (particularly within the *hgp* system), as well as variants in tRNA and rRNA genes. These genetic elements may contribute to the invasive phenotype by facilitating immune evasion, enhancing iron scavenging systems, and supporting efficient protein synthesis under host-imposed stresses. Additionally, our recombination analyses revealed two recombination hotspots which coincided with genes associated with the invasive phenotype, potentially explaining the spread of invasive traits across different lineages. However, our study is limited by its cross-sectional design, which provides only a snapshot of the bacterial population at a single time point, and by potential host-related factors that were not captured in our genomic analyses. Future longitudinal and multi-faceted investigations are warranted to fully elucidate the dynamic interplay between genetic determinants, recombination processes, and host susceptibility in the pathogenesis of invasive NTHi infections.

Chapter 5

The population genetics of *Haemophilus influenzae* isolates from Indonesia in carriage and disease

Abstract

Haemophilus influenzae remains an important cause of respiratory and invasive disease, with the rise of non-typeable *H. influenzae* (NTHi) and increasing antimicrobial resistance posing significant challenges to public health. This study investigated the population genetics of *H. influenzae* isolates from Indonesia, from both carriage and invasive disease, using a high-resolution core genome multilocus sequence typing (cgMLST) scheme and life identification number (LIN) code barcoding system. A total of 113 isolates, confirmed through qPCR and whole-genome sequencing, were analysed for serotype, antimicrobial susceptibility, and virulence factor (VF) profiles. Genomic data were integrated with phenotypic antibiotic resistance testing and compared against a curated global dataset (N = 2,869). The majority of the Indonesian isolates (94.7%) were non-typeable, and cgLIN analysis revealed that Indonesian *H. influenzae* were highly diverse, distributed across 21 megalineages. Invasive isolates did not form a single clade, highlighting the polyclonal nature of disease-causing NTHi strains. A total of 24/52 of the isolates tested showed resistance to at least one antibiotic, particularly ampicillin, mediated either by the acquisition of beta-lactamase gene or mutations in the *ftsI* gene. Resistance to chloramphenicol and fluoroquinolones was also detected and attributed to *catII* and mutations in *gyrA*, *parC*, and *parE*, respectively. The study underscores the utility of cgMLST and cgLIN for high-resolution surveillance and the importance of including carriage isolates to capture the full spectrum of strain diversity. These findings

establish a genomic framework for future public health surveillance and antimicrobial stewardship efforts in Indonesia and other LMICs.

5.1. Introduction

Haemophilus influenzae is a Gram-negative coccobacillus that colonises the upper respiratory tract (URT) of humans, which are its only known natural host. While colonisation is generally asymptomatic, *H. influenzae* is also an opportunistic pathogen capable of causing a wide range of diseases, from localised infections such as otitis media, sinusitis, and conjunctivitis, to life-threatening invasive diseases such as bacteraemia, meningitis, and pneumonia. Based on the presence or absence of a polysaccharide capsule, *H. influenzae* can be classified into encapsulated strains (serotypes a to f) and unencapsulated or non-typeable *H. influenzae* (NTHi) [2, 283]. Among encapsulated types, serotype b (Hib) historically accounted for the majority of severe invasive disease in children, especially meningitis [202, 203].

The global implementation of Hib conjugate vaccines has drastically reduced the incidence of Hib disease in many countries [203], including Indonesia, where the vaccine has been part of the national immunisation program since 2013 [284, 285]. However, the decline of Hib has coincided with a relative increase in disease caused by other serotypes (particularly *H. influenzae* type a [Hia] and type f [Hif]) and more notably by NTHi. NTHi has now emerged as the most frequent cause of invasive *H. influenzae* disease worldwide [205]. This pattern is observed across different regions worldwide. Reports from the European Centre for Disease Prevention and Control (ECDC) showed that most *H. influenzae* invasive disease cases were attributed to NTHi (78%) or non-Hib encapsulated isolates (13%) [206, 207]. According to a

surveillance program in South Africa in 2018, 64% of invasive *H. influenzae infections* were due to NTHi [208].

The current challenges posed by *H. influenzae* represent a double burden: an increasing number of invasive infections are now caused by non-vaccine-preventable strains, while antimicrobial resistance (AMR) among circulating strains continues to rise. *The bacterium* is listed as one of twelve bacterial pathogens monitored by the World Health Organisation (WHO) Global Antimicrobial Resistance Surveillance System (GLASS) [71]. In particular, strains resistant to ampicillin (AMP) are designated as a "medium-priority" threat in the WHO Bacterial Priority Pathogen List (BPPL) [72]. This problem is particularly imminent among NTHi. Multiple reports previously found that resistance to AMP was highest in NTHi [22, 25] and collectively NTHi was resistant to a wider range of antibiotics classes [22]. Based on WHO "AWARE" Classification tools, AMP is still classified as "Access" antibiotics, which means they have lower resistance potential, along with other antibiotics such as amoxicillin-clavulanic acid (AMC) and chloramphenicol (CHL) [286]. Meanwhile, fluoroquinolones and third-generation cephalosporins belongs to "Watch" antibiotic group because these antibiotics have higher resistance potential and hence, should be used with caution [286]. In *H. influenzae*, t high-level levofloxacin and moxifloxacin (MOX) resistance has been found[287], as well as cefotaxime (CTA) [59] and ceftriaxone (CTR) [88].

The dual threat of increasingly invasive and resistant NTHi complicates both prevention and treatment strategies, especially in settings with limited surveillance infrastructure. In many low- and middle-income countries (LMICs), including Indonesia, the growing burden of NTHi remains largely unmonitored and uncharacterised. Routine diagnostic infrastructure is often inadequate to culture and characterise this fastidious bacterium, and molecular

diagnostics and serotyping using PCR are not widely implemented. Consequently, comprehensive data on the burden, genetic diversity, and AMR profiles of circulating *H. influenzae* strains are lacking. Previous studies in Indonesia have mainly focused on carriage rate in specific settings, without molecular characterisation [39, 288, 289]. There are currently no genomic surveillance studies that describe the broader spectrum of *H. influenzae*, especially those involving both invasive and carriage isolates.

Studying the population genomics of bacterial pathogens in both carriage and disease is essential for understanding the full spectrum of strain diversity and transmission dynamics. Carriage isolates serve as reservoirs for virulent and antimicrobial-resistant clones, often preceding clinical disease and outbreaks. Including carriage data also informs vaccine design and surveillance strategies by identifying lineages with epidemic potential before they emerge in clinical settings [61, 290]. In Chapter 2 and 3, a high-resolution core genome multilocus sequence typing (cgMLST) scheme for *H. influenzae* followed by integration of the life identification number (LIN) code hierarchal clustering and nomenclature system was established, allowing for accurate delineation of the bacterium population structure despite the source of isolates.

The aim of this chapter was to investigate the population genetics of circulating *H. influenzae* in Indonesia from carriage and disease using the cgMLST scheme and cgLIN code. Additionally, provenance, clinical data, and antimicrobial susceptibility testing (AST) results are analysed in relation to the underlying bacterial population structure.

5.2. Method

5.2.1. Design, setting, and primary dataset

This was a descriptive analytic study that utilised archived *H. influenzae* isolates and associated data obtained from previous projects. Each of these projects involved surveillance conducted in different populations across various geographic regions in Indonesia (**Table 5.9**).

Table 5.9. *H. influenzae* surveillance projects from which isolates in this study originated.

No	Project primary aim	Study population characteristics	Number of archived isolates	Reference
1	Investigation of the NP carriage rates, serotype distribution, and antimicrobial susceptibility of *Streptococcus pneumoniae* in children under 5 years of age in Kotabaru, South Kalimantan, Indonesia, in 2019	Healthy young children aged 0–59 months during monthly visit to a community health centre.	42	[291]
2	Assessing *S. pneumoniae* carriage rate and serotype distribution before and after PCV13 vaccination (2+1 schedule) in Lombok Island March 2018 – June 2019.	2-month-old healthy infants during routine visit to a primary care facility for scheduled vaccination. Subjects were followed-up prospectively	43	[292, 293]
3	Evaluation of changes in *S. pneumoniae* carriage rate among children in a hospital outpatient clinic, before and after PCV13 vaccination.	Asymptomatic children > 6 and < 18 years-old during routine visit at a hospital outpatient clinic who were offered PCV13 based on underlying medical conditions [294, 295].	22	[296]
4	Building a capacity for *S. pneumoniae* and *H. influenzae* identification and isolation at a tertiary hospital in Surabaya, Indonesia through passive surveillance.	Hospitalised patients of any age with suspected invasive infection, indicated for bacterial culture.	10	[297]
	Total		117	

NP: nasopharyngeal, PCV13: 13-valent pneumococcal conjugate vaccine

Some steps (Figure 5.19), including sample size determination, sample collection, selection of nasopharyngeal (NP) swabs for conventional *H. influenzae* identification, and isolate storage, were completed prior to the conception of the current study. As this study relied solely on archived isolates, the original NP swab samples were not accessed, and all associated metadata were fully de-identified.

5.2.2. *H. influenzae* isolation, identification, and serotyping

H. influenzae isolates used in this study were isolated from different specimen types, including NP swab, blood, and pleural fluid. Below are the details of isolation procedure from each specimen:

1. NP swab
    - NP swab specimens were stored in skim milk-tryptone-glucose-glycerol (STGG) media, allowed to thaw at room temperature (25°C), then vortexed for 10-20 seconds.
    - 100 µl of the sample was transferred to supplemented chocolate agar plate (sCAP) media and air-dried for several minutes. The sample was spread onto the media using the 4-quadrant streak method and incubated for 18-24 hours at 37°C with 5% CO_2.
2. Blood
    - Blood specimens were inoculated into a blood culture bottle and incubated according to the manufacturer's protocol (BD BACTEC FX).
    - Blood culture bottles positive for growth were examined using Gram staining. Preliminary identification of *H. influenzae* was made when pleomorphic, Gram-negative cells were observed under the microscope. Pleomorphism refers to the presence of bacilli, coccobacilli, and filamentous forms within the same Gram-stained specimen, observed during a single microscopic examination.

- Sample from positive blood culture bottles were cultured on sCAP media and incubated for 18-24 hours at 37°C with 5% CO_2.

3. Pleural fluid
    - Pleural fluid specimen was processed immediately upon arrival. 1.5 ml of pleural fluid was centrifuged at 1000 g for 15 minutes. The supernatant was discarded until 100 µl remained.
    - A total of 20 µl of concentrated pleural fluid was used for Gram staining and 20 µl was inoculated on sCAP media then incubated at 35-37°C with 5% CO_2 for 18-22 hours.

Subsequently, colonies growing on the sCAP media were observed and tests further to conclude microbiology identification of *H. influenzae*:

1. Colony morphology observation: *H. influenzae* colonies will appear grayish, mucoid, without haemolytic activities, and have a distinctive indole smell.
2. Gram staining observation: *H. influenzae* appears as Gram negative pleomorphic cells.
3. Colonies suspected to be *H. influenzae* were tested for oxidase activities and XV factor dependence:
    3.1. Oxidase test: A pure colony was picked using a 1 µl inoculating loop and streaked onto a BBL DrySlide Oxidase (±3-4 mm). Positive oxidase activity was indicated by the appearance of a purple colour on the BBL DrySlide Oxidase within 20 seconds.
    3.2. XV factor dependence test: A bacterial suspension in 2 mL NaCl 0.85% with the turbidity equal to McFarland standard 1.0 was prepared and ensured its homogeneity. The suspension was lawn on trypticase soy agar (TSA) media

using cotton swab. XV, V, and X discs were placed onto the media with 2 cm distance between each disc and from the edge of the petri dish. The media was incubated in a 5% CO_2 incubator at 37°C for 18-20 hours. *H. influenzae* identification was verified if there was growth only around XV disc.

4. The suspected colonies were also subcultured on sCAP media and incubated at 35-37°C with 5% CO2 for 18-22 hours. Subsequently, these isolates were also stored as stock at -80°C in the STGG medium.

All *H. influenzae* isolates initially identified through microbiological testing were further tested for definitive identification using a molecular technique targeting the *hpd* gene detection via quantitative real-time polymerase chain reaction (qRT-PCR). Only isolates that tested positive for *hpd* were considered confirmed *H. influenzae* [298] and subsequently subjected to serotyping by detecting capsule region genes, also using qRT-PCR [10, 299]. The primers and qRT-PCR conditions used for *hpd* confirmatory testing and serotyping were provided in Appendix 5.1.

5.2.3. Antibiotic susceptibility testing (AST)

The disk diffusion method was performed to evaluate the susceptibility of *H. influenzae* isolates for antibiotics often prescribed in the outpatient and inpatient setting in Indonesia, including AMP, amoxicillin-clavulanic acid (AMC), chloramphenicol (CHL), fluoroquinolones (ciprofloxacin [CIP] and moxifloxacin [MOX]), and third-generation cephalosporin (cefotaxime [CTA] and ceftriaxone [CTR]) [300]. Experiments were conducted following the Clinical and Laboratory Standards Institute (CLSI) protocol 2022 [301]; however, due to the unavailability of the *Haemophilus* Test Medium (HTM) in the country, sCAP medium was used as an alternative. Previous studies have reported the use of sCAP for disk diffusion

testing of *H. influenzae* and found no significant impact on results for the antibiotics tested in this study, although the medium's dark colouration can hinder accurate measurement of inhibition zones [302-304]. To address this limitation, all zone measurements were independently verified by two observers (see list of contributors). Below were the details of the testing conditions:

1. Colony suspension equals to 0.5 McFarland standard prepared using colonies cultured overnight (20-24 hour).
2. A lawn culture was prepared by spreading the bacterial suspension evenly across the surface of a 100-mm agar plate using a sterile swab.
3. Four antibiotics disks were placed on the plate, ensuring adequate spacing from the plate edge and between disks to avoid overlapping zones of inhibition.
4. Incubation at $35^oC \pm 2^oC$ with 5% CO_2 for 16-18 hours.

Complete inhibition zone measurement was done by visual inspection. The zone margin was the edge without no obvious, visible growth. Phenotypic beta-lactamase test was not possible to the unavailability of reagents in the country.

5.2.4. Whole-genome sequencing and post-assembly analyses

Bacterial DNA was extracted using the DNeasy Blood and Tissue® Kit according to the manufacturer's protocol (Qiagen, Carlsbad, CA, USA). The Nextera XT DNA Library Prep® kit was used for library preparation before paired-end WGS (2x250) on a MiSeq (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. FastQ reads were evaluated for quality and trimmed using FastQC v.0.11.9 and Trimmomatic v.0.32 pipeline, respectively [305, 306]. Draft genomes were assembled utilizing the SPAdes v.3.15.5 toolkit with the setting optimized for 2x250 bp reads [228]. Quality assurance and quality control (QA/QC) for draft

genome assemblies were verified using QUAST. Statistics for assembly quality (number of contigs, genome length, GC content, N50, and L50) are reported in Supplementary Table 5.1.

Genome assemblies and relevant, de-identified isolate characterization data were uploaded to the PubMLST website (http://pubmlst.org) [49]. All genomes were evaluated for definitive species identification and potential contamination by characterizing the ribosomal multilocus sequence typing (rMLST) profile. rMLST is a typing method that unifies bacterial genealogy and typing by cataloging variations in 53 bacterial ribosome protein subunit *(rps*) genes. This approach is capable of categorising bacterial sequences into taxonomic groups at all hierarchical levels [140]. The seven-locus MLST profile and clonal complex (CC) were automatically assigned on the PubMLST website. The website also integrated *hicap,* a *H. influenzae* serotype prediction tool based on the evaluation of capsule operon in the genome (i.e. capsule genotype), and concordance between qRT-PCR serotype and capsule genotype was assessed [152].

Additionally, the *H. influenzae* cgMLST scheme and the associated LINcode which were developed as a part of the thesis project (Chapter 2 and Chapter 3) was already implemented in the PubMLST database at the time of the analysis for this chapter. Therefore, all *H. influenzae* isolates in this study were annotated for 1,037 core genes in the cgMLST scheme. If 1,012 of these genes were present in a genome (i.e. maximum number of missing core gene of 25), it was assigned a core genome ST (cgST) and subsequently the LIN code barcode. Two thresholds, the megalineage and superlineage, were chosen to describe the population structure of the *H. influenzae* isolates collection from Indonesia.

H. influenzae whole-genome annotation of VF-related genes was performed using the curated Virulence Factor Database (VFDB) across different bacterial species, which consisted

of a total of 32,809 VF-related genes. In the VFDB, VF-related genes for *Haemophilus* sp. were curated from several reference genomes: *H. influenzae* 86-028NP (NC_007146), PittEE (NC_009566), PittGG (NC_009567), Rd KW20 (NC_000907), *H. ducreyi* 35000HP (NC_002940), *H. somnus* 129T (NC_008309) and 2336 (NC_010519). The first three *H. influenzae* reference genomes were NTHi [11, 12]. A BLAST search with 80% minimum alignment and 90% minimum nucleotide identity threshold was conducted and genes found in at least 1 isolate was collectively included in a "VF scheme" [307]. This scheme was processed through chewBBACA allele calling algorithm to conclude the presence/absence of the VF-related genes [55].

Antimicrobial resistance (AMR) determinants in *H. influenzae* isolates were identified using the Resistance Gene Identifier (RGI) tool from the Comprehensive Antibiotic Resistance Database (CARD) [308]. This tool detects both acquired resistance genes, such as beta-lactamases, and chromosomal mutations associated with intrinsic resistance, provided these variants have been curated in the database. To supplement the CARD-based analysis, known resistance-associated genes and mutations for beta-lactams, phenicols, and fluoroquinolones previously reported in *H. influenzae* were also evaluated (Chapter 1, Table 1.4). This includes a search for point mutations in the *ftsI* gene that encodes for the penicillin-binding protein 3 (PBP3) in the case of beta-lactam resistance and in the *gyrA, parC,* and *parE* gene which confers resistance to fluoroquinolone antibiotics. All four genes were defined and curated in the *H. influenzae* PubMLST database as *ftsI*, HAEM1394, HAEM1650, and HAEM1649. *ftsI* alleles of beta-lactamase negative beta-lactam resistance isolates were evaluated through pairwise alignment with wild-type allele from two *H. influenzae* reference genomes, strain FDAARGOS_1560 (NCBI accession GCF_020736045.1) and Rd KW20 (GCA_000027305.1). Similarly, *gyrA, parC,* and *parE* alleles were assessed for fluoroquinolone-resistant isolates.

Collection of *H. influenzae* plasmid sequences was obtained from the NCBI Nucleotide database through searching using the following keywords: "(Haemophilus influenzae) AND "Haemophilus influenzae"[porgn:_txid727]" with the filter "Genetic compartments: Plasmid" chosen. It resulted in 47 plasmid sequences, including one integrative conjugative element (ICE). Through this experiment, I intended to conclude the presence/absence of full-length plasmid; therefore, percentage nucleotide similarity threshold of 95% and minimum length alignment of 80% to the full-length plasmid were implemented. Additionally, as the *H. influenzae* plasmid sequences did not include all transposons, BLAST search of transposons sequences from NCBI were done. A total of 7 transposons sequence were retrieved (NCBI accessions: NC_011409, MN106410, MN106411, MN106412, OW736085, OW736566, OW737369) [309, 310]. The result of plasmid detection was compared to AMR genes detection for congruence.

Core genome alignment, based on concatenated aligned core gene sequences from the *H. influenzae* cgMLST scheme, was generated through the Genome Comparator plug-in on the PubMLST website. From this alignment, a maximum-likelihood (ML) tree was constructed with RAxML v8 and ClonalFrameML to account for recombination events [100, 151]. The resulting tree was visualised **through** iTOL and annotated with supporting metadata, including megalineages or superlineages, isolate source (i.e. carriage or disease), and AST results [189]. Core genome alignment and phylogeny reconstruction and visualisation were also conducted separately for *H. influenzae* isolates from invasive disease cases using the same methods and tools. Both annotated phylogenies would be available for online viewing.

5.2.5. Global collection of *H. influenzae* genomes

To evaluate the diversity of *Haemophilus influenzae* isolates from Indonesia in the context of globally circulating strains, a second dataset was curated from publicly available genomes in the PubMLST and NCBI Sequence Read Archive databases. This dataset was initially compiled from those used in Chapters 2, 3, and 4, which included only high-quality draft genomes selected through a systematic filtering process. In the current study, an additional filtering criterion was applied: draft genomes lacking geographical information or source of isolation were excluded. A minimum spanning tree (MST) of the global and Indonesian *H. influenzae* genomes was then constructed and visualised using GrapeTree with the MSTreeV2 model [153].

5.3. Result

5.3.1. High positive rate of *H. influenzae* among carriage and invasive cases were detected at the extremes of age.

The overall workflow and results from each step of process from archived isolates selection until post-assembly quality control are depicted in **Figure 5.19**. A total of 256 NP swabs were selected from the three regions (**Table 5.9**) for *H. influenzae* identification resulting in 107 isolates H. influenzae positive samples. A total of 70 (65.42%) were obtained from children aged 5 years and under, with 22 (20.56%) from 5–9-year-old children and 15 (14.02%) 10–18-year-old children. The age group distribution reflected the differences in the scope of previous projects from which the archived isolates originated.

In contrast, 7 out of 10 (70%) *H. influenzae* isolates from clinical samples were obtained from adult patients over 45 years of age, including 4 geriatric patients > 60 [311]. Among the

three paediatric patients, one was a neonate less than one month old. All of these isolates were recovered from sterile body sites, confirming their invasive nature. There was also a slight difference in sex distribution between carriage and invasive cases: in the carriage group, most isolates were from male subjects (58/107, 54.2%), whereas among invasive cases, 6 out of 10 were from female patients (6/10, 60%).

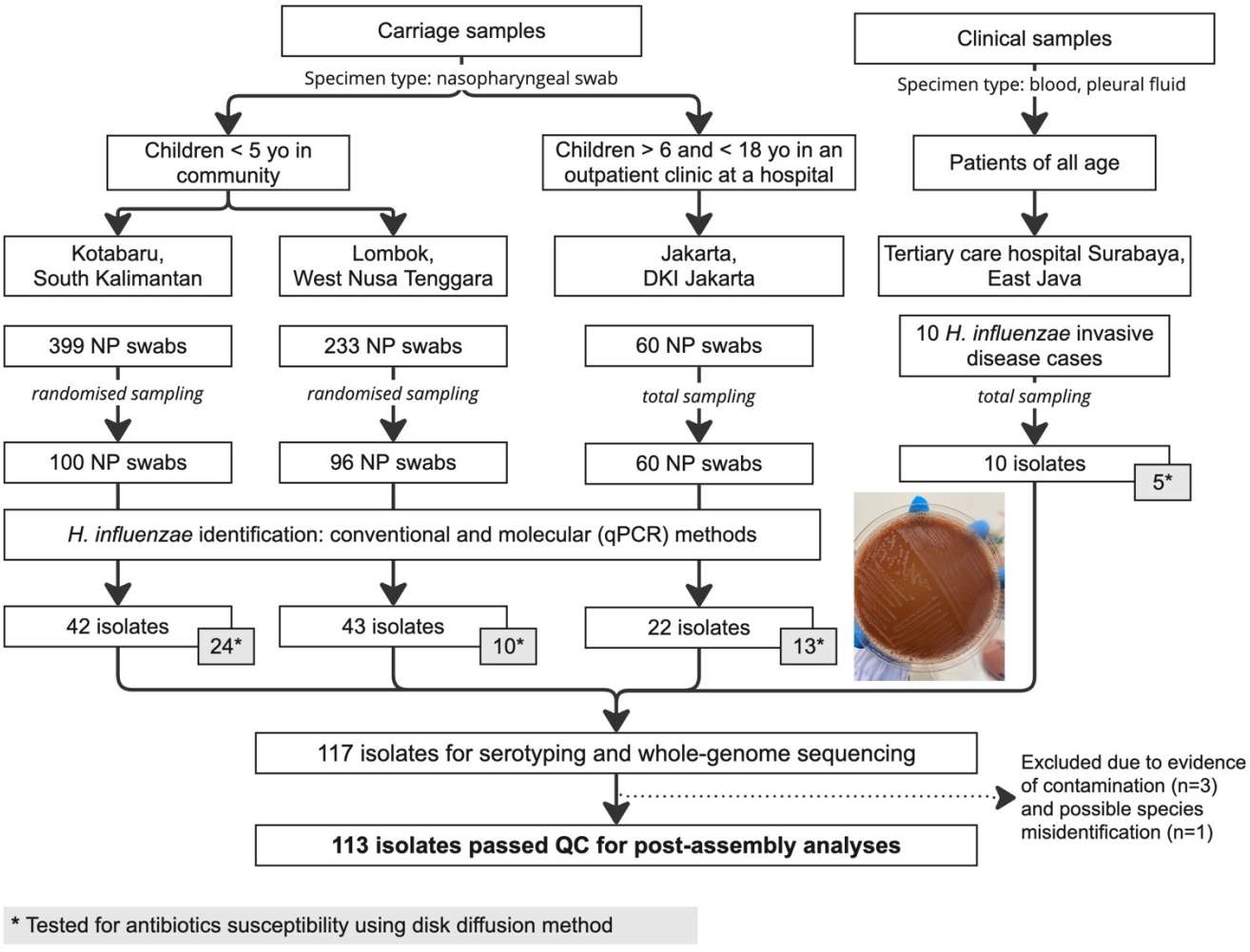


Figure 5.19 Description and results at each stage of the study workflow from sample selection, *H. influenzae* identification, serotyping, and whole-genome sequencing. Each nasopharyngeal (NP) swab originated from one unique subject and each isolate represented one identified single *H. influenzae* colony, ensuring de-duplication. *qPCR: quantitative polymerase-chain reaction, QC: quality control.*

All *H. influenzae* from invasive cases were collected in 2019 at a single tertiary care hospital. Two isolates were identified in January and two in April, while one isolate was identified in each of the following months: March, May, June, July, August, and December. All patients were hospitalised in inpatient wards. Eight isolates were obtained from blood, and two from pleural fluid specimens. One pleural fluid sample was suppurative, supporting the clinical diagnosis of empyema. These isolates were collected as part of routine clinical microbiology diagnostic procedures.

The carriage rate across isolates sampled from different projects and regions ranged from 36.67 (22/60) to 44.8% (43/96). All *H. influenzae* isolates, identified based on the colony characteristics and *hpd* gene detection, were serotyped. A total of 104 (94.69%) of the isolates were non-typeable, 4 (3.54%) were Hib, and 2 (0.77%) were Hif. All encapsulated *H. influenzae* were isolated from carriage, with Hib carriage rate of 4/256 (1.56%). Following WGS and genome assembly, three isolates were excluded due to obvious evidence of contamination and one isolate was designated as *H. haemolyticus* based on the ribosomal gene typing (i.e. rST) (Appendix 5.2) For 113 isolates which passed the QC step, there was 100% congruence between serotyping by qRT-PCR and the capsule genotyping based on the presence/absence of the capsule operon using the hicap tool.

5.3.2. *H. influenzae* isolates from Indonesia were highly diverse, belonging to 21 different megalineages based on cgLIN code.

All 113 *H. influenzae* draft genomes were deposited in the PubMLST database for assignment of cgSTs, LIN codes, and MLST sequence types (STs) and clonal complexes (CCs) (Supplementary Table 5.1). A minimum of 1,012 core genes is required for cgST and LIN code assignment, and all genomes in this study met this criterion. Isolates were distributed across

21 unique megalineages, meaning that each group differed from the others by at least 900 core gene alleles, reflecting a highly diverse population. A total of 65 out of 113 (57.52%) *H. influenzae* were in the megalineage 0, with most of them belonging to superlineage 0_0 (n = 21) and 0_1 (n = 23). There were 9 megalineages associated with single *H. influenzae* isolates. This high level of diversity was confirmed after the reconstruction of the phylogeny based on the core genome alignment using the maximum-likelihood method (Figure 5.20). The LIN code mega- and superlineages assignment was in agreement with the tree topology.

At a more magnified level of the LIN code, *H. influenzae* isolates were divided into 76 clades at the 110 allelic mismatches threshold. Even though there were also 76 STs identified in the dataset, as shown in Chapter 3, the two clustering approaches did not entirely coincide. For example, within ST-407, seven *H. influenzae* isolates belonged to three different LIN code superlineages, and within ST-137, four isolates were assigned to two different megalineages (Supplementary Table 5.1). Since LIN code hierarchical clustering is based on 1,037 core genes, its delineation of population structure is more precise. However, MLST ST and CC were reported for comparison to previous studies as it was the most universally used scheme for depicting *H. influenzae* population structure.

The three largest CCs were the ST-3, ST-393, and ST-422 complexes, comprising 11, 8, and 5, out of 113, isolates, respectively. All isolates within these CCs were phenotypically and genotypically non-typeable, except for one Hif. Similarly, the fourth largest CC in the dataset, the ST-6 complex, was primarily associated with Hib; however, one NTHi isolate also belonged to this CC. This underscores the extent of recombination affecting the seven housekeeping genes used in the *H. influenzae* MLST scheme, highlighting its limitations in accurately describing population structure. Nonetheless, the reconstructed phylogeny showed

clear clustering of the four Hib isolates within superlineage 29 and the two Hif isolates within superlineage 40, with no NTHi isolates found in these clusters. Notably, three of the four Hib isolates differed from each other by only one core gene, and the two Hif isolates differed by ten core genes (Supplementary Table 5.1).

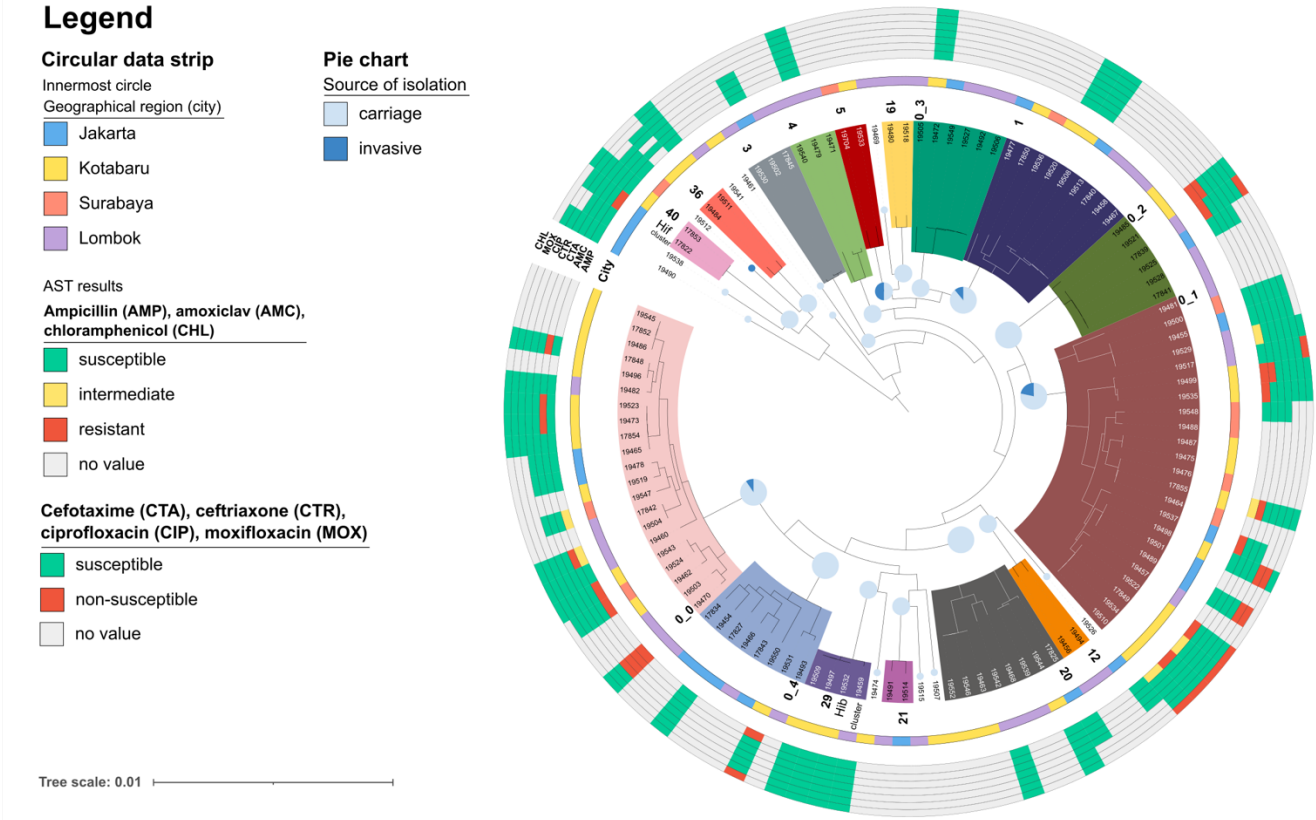


Figure 5.20. The maximum-likelihood phylogeny of *H. influenzae* isolates from Indonesia based on the core genome alignment using the *H. influenzae* cgMLST scheme. The tree clades were coloured by megalineage and superlineage (applicable only to megalineage 0) information as derived from the cgLIN code. Two capsulated *H. influenzae* clusters were identified: Hif and Hib, corresponding to megalineages 40 and 29, respectively. Megalineages represented by single isolates (singletons) were not annotated on the tree. Pie charts represented the proportion of isolates from carriage and invasive cases within the megalineage, superlineage, or terminal branch. *AMP: ampicillin, AMC: amoxicillin-clavulanic acid, CTA: cefotaxime, CTR: ceftriaxone, CIP: ciprofloxacin, MOX: moxifloxacin, CHL: chloramphenicol.*

The reconstructed phylogeny of *H. influenzae* from Indonesia depicted in **Figure 5.20** was annotated with the geographic origin and source of each isolate. No distinct geographical clustering was observed, as isolates from different regions were often found within the same lineage. All invasive isolates originated from a single city, Surabaya; therefore, their geographic data was not informative and could be omitted. Notably, these invasive isolates did not form a distinct lineage or cluster in the phylogeny, although 5 out of 10 belonged to superlineage 0_1, the most common superlineage in the dataset.

The genetic relatedness among the 10 invasive isolates was assessed by constructing a ML phylogeny (**Figure 5.21**). Each isolate had a unique ST, including four novel STs not previously reported in PubMLST. Based on the cgLIN nomenclature, isolates were distributed across six distinct lineages: four belonged to lineage 0_1_0, two to lineage 0_0_0, and the remaining four to separate lineages. All lineage 0_1_0 isolates were part of the ST-107 clonal complex. These four isolates were obtained from patients diagnosed with septicaemia, with or without a clearly identified focus of infection. No carriage isolates in the dataset belonged to this CC, although 12 carriage isolates were assigned to lineage 0_1_0. Two of the four invasive isolates within this lineage were highly similar, differing by no more than 40 core gene alleles (i.e. they shared the first seven characters of the LIN code prefix), and were recovered from patients with septicaemia and pneumonia.

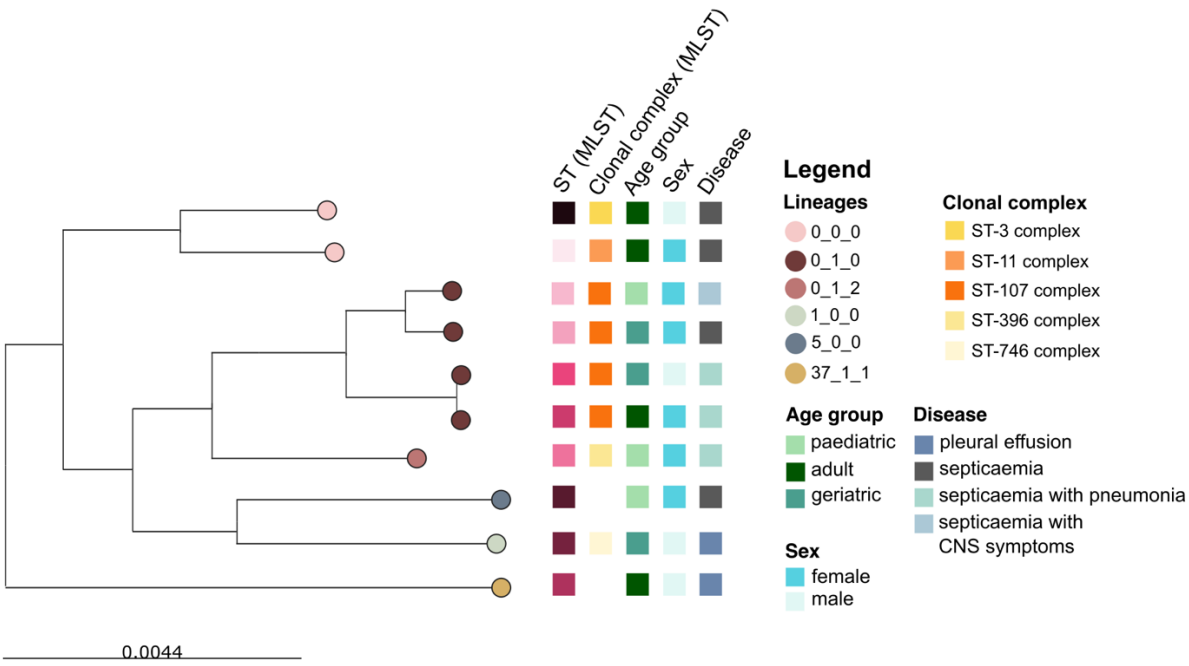


Figure 5.21. ML phylogeny of *H. influenzae* isolates from invasive cases. The tree nodes were coloured with lineages grouping in accordance with the assigned LIN code barcodes. The MLST clonal complex (CC) showed congruence with the lineage designation; however, not all isolates had CC assigned. *CNS: central nervous system.* This figure was adapted from Krisna et al [297].

The two isolates most distantly related to lineage 0_1_0 were obtained from patients with pleural effusion. Although both shared the same clinical diagnosis, they belonged to different megalineages, suggesting substantial variation in core gene content. Similarly, no clear clustering was observed among *H. influenzae* isolates based on patient age group or sex. Due to the limited sample size, no definitive conclusions could be drawn regarding associations between demographic or clinical characteristics and genetic lineage.

5.3.3. All phenotypically resistant *H. influenzae* isolates were explained by previously described AMR determinants

AST was performed on 52 out of 113 *H. influenzae* isolates from Indonesia using seven antibiotics. Nearly half (24/52, 46%) were resistant to at least one antibiotic, with most (21/24,

88%) showing resistance to a beta-lactam (**Table 5.10.** Molecular characteristics of *H. influenzae* isolates non-sensitive to beta-lactam antibiotics.**Table 5.10**). Both acquired and chromosomal AMR determinants associated with beta-lactam resistance were investigated in these isolates. Of the 21 beta-lactam-resistant isolates, 18 showed non-susceptibility to AMP, while 3 were resistant to AMC but not AMP. Notably, 13 of the 18 AMP-non-susceptible isolates carried the beta-lactamase gene *bla*TEM-1, whereas none of the AMC-only resistant isolates did. A BLAST search for another beta-lactamase gene, *bla*ROB-1, yielded no matches in any of the isolates tested.

A BLAST search against a curated database of *H. influenzae* plasmid and transposons, collectively referred as mobile genetic elements (MGEs), revealed that all isolates carrying the *bla*TEM-1 gene matched at least one MGE sequence with >95% nucleotide identity (**Table 5.10**). The most common MGE identified was ICEhin1056, originally described by Juhas et al. [312], found in six isolates. Of note, four of these six isolates also carried the *sul2* gene, which confers resistance to sulphonamide antibiotics although phenotypic testing for this antibiotic class was not conducted in this study. Five additional *bla*TEM-1-positive isolates harboured the transposon Tn6685, distinguished by the presence of a complete Tn3 element encoding a transposase. Lastly, two isolates carried the *bla*TEM-1 gene on plasmids.

Table 5.10. Molecular characteristics of *H. influenzae* isolates non-sensitive to beta-lactam antibiotics.

Isolate	Phenotypic resistance				Presence of TEM-1 beta lactamase	ftsI allele[a]	ftsI group[b]	Mobile genetic element present	Resistance to other antibiotics
	AMP	AMC	CTA	CTR					
17822	R	S	S	S	Yes	6	WT	pA1209,pLFH49,pLFS5,pA1606,pLFH64,pJ612	-
17839	R	S	S	S	Yes	193	WT	ICEhin1056	CHL
17849	I	S	S	S	Yes	15	WT	ICEhin1056*	CHL
19499	R	R	S	S	Yes	1	IIB	Transposon Tn6685	-
19510	I	S	S	S	Yes	15	WT	ICEhin1056*	CHL
19521	R	R	S	S	Yes	2	IIB	pPN223	-
19522	R	S	S	S	Yes	15	WT	ICEhin1056*	CHL
19524	R	S	S	S	Yes	27	WT	Transposon Tn6685	-
19531	R	S	S	S	Yes	37	WT	ICEhin1056	CHL
19534	R	S	S	S	Yes	15	WT	ICEhin1056*	CHL
19535	R	S	S	S	Yes	1	IIB	Transposon Tn6685	-
19537	R	S	S	S	Yes	15	WT	Transposon Tn6685	-
19543	R	S	S	S	Yes	27	WT	Transposon Tn6685	-
17848	S	R	S	S	No	43	IIA	-	-
17855	I	R	S	S	No	151	IIA	-	-
19470	R	R	NS	NS	No	26	III-like+	-	-
19473	S	R	S	S	No	23	IIB	-	-
19504	I	R	S	S	No	2	IIB	-	-
19523	S	R	S	S	No	23	IIB	-	-
19529	I	S	S	S	No	5	IID	-	-
19547	I	S	S	S	No	43	IIA	-	-

*Similar to ICEhin1056 with 88.57-97.14% genes present in these isolates but with one additional sul2 gene conferring resistance to sulphonamide
#Isolates had A502V and N526K mutations characteristics of ftsI group IIB but with an extra M377I mutation.
[a]Allele number as curated on the PubMLST H. influenzae sequence typing database, accessed January 2025.
[b]According to grouping by Nurnberg et al [59]
AMP: ampicillin, AMC: amoxicillin-clavulanic acid, CTA: cefotaxime, CTR: ceftriaxone, CHL: chloramphenicol; WT: Wild type

Five AMP-non-susceptible but *bla*TEM-1-negative isolates, along with all three AMC-only resistant isolates, carried mutations in the *ftsI* gene associated with reduced susceptibility to beta-lactam antibiotics. The *ftsI* mutations were classified based on a previously published scheme [59, 313, 314] (**Table 5.10**), which groups them according to the type and location of amino acid substitutions. No *ftsI* mutation group was found to be exclusive to either resistance category. One isolate (ID 19470) was assigned to *ftsI* group III-like+, carrying a combination of mutations previously associated with elevated CTR minimum inhibitory concentrations (MICs). Phenotypically, this isolate was resistant to all beta-lactam antibiotics tested, including AMP, AMC, CTA, and CTR.

Most of the *bla*TEM-1-positive isolates had wild-type *ftsI*; however, three carried *ftsI* mutations at three positions: M377I, A502V, and N526K. In total, six beta-lactam-resistant *H. influenzae* isolates in the dataset possessed this mutation pattern. Although this combination does not align with any previously defined *ftsI* group, it most closely resembles group IIB. The *ftsI* grouping scheme provides a framework for interpreting beta-lactam resistance associated with *ftsI* mutations; however, additional mutations described in the literature have not yet been incorporated into this classification system (Chapter 1, Table 1.4). To provide a more detailed overview, a complete list of mutations and their respective positions for each *ftsI* allele is included in Appendix 5.3.

Seven out of 52 (13%) *H. influenzae* isolates from Indonesia tested for antibiotic susceptibility were resistant to chloramphenicol. All of them carried the *catII* gene, which encodes chloramphenicol acetyltransferase, an enzyme that inactivates chloramphenicol. Similar to *bla*TEM-1, this gene is typically carried on ICEhin1056; however, one isolate did not exhibit concurrent resistance to any beta-lactam antibiotics. Upon evaluation and visualization of the

draft genome assembly, this isolate was found to contain 90% of the coding sequences (CDSs) associated with ICEhin1056 but lacked the region where the *bla*TEM-1 gene is usually located. This deletion also included the neighbouring gene encoding a recombinase family protein.

Lastly, resistance to fluoroquinolone antibiotics, CIP and MOX, was identified in two isolates (IDs 19489 and 19498). No resistance determinants were detected through BLAST search against the CARD database, likely due to the limited curation of fluoroquinolone-resistant *H. influenzae* at the time of analysis. However, three genes, *gyrA*, *parC*, and *par*E, have been reported in the literature to confer fluoroquinolone resistance when mutated. In the *H. influenzae* PubMLST sequence definition database, these genes are annotated as HAEM1394, HAEM1650, and HAEM1649, respectively. Both fluoroquinolone-resistant isolates shared identical alleles for these three loci. Critical mutations identified included S84L and D88N in HAEM1394 (*gyrA*) and S84I in HAEM1650 (*parC*). These mutations were determined by aligning the allele sequences with those of two *H. influenzae* reference genomes, FDAARGOS1560 and Rd KW20 (NCBI accession GCF_020736045.1 and GCF_037094705.1). A complete list of point mutations in the quinolone resistance-determining regions of HAEM1394, HAEM1650, and HAEM1649 is provided in Appendix 5.4.

One important limitation of this study was the incompleteness of AST data, even for the largest lineages, which hindered conclusive interpretation of the relationship between antibiotic resistance and the population structure of *H. influenzae*. Nonetheless, one clonal group within superlineage 0_1 (LIN code prefix 0_1_0_0) included seven isolates, all resistant to more than one antibiotic: five to AMP and CHL, and two to CIP and MOX. The five *H. influenzae* isolates had nearly identical LIN codes and differed from each other by no more than five core gene allelic mismatches. A search of the *H. influenzae* isolate database on PubMLST (accessed on 4

April 2025) revealed no other genomes with this LIN code prefix within five allelic mismatches (0_1_0_0_1_0_0_0_0_0). Similarly, the two fluoroquinolone-resistant isolates shared the same LIN prefix up to this threshold. A database search identified two additional public *H. influenzae* genomes with identical LIN code prefixes (0_1_0_0_0_1_0_0_0_0). All four genomes carried the same alleles for HAEM1394 (*gyrA*), HAEM1650 (*parC*), and HAEM1649 (*parE*). These isolates would be predicted to be resistant to both CIP and MOX.

5.3.4. Known virulence factors-encoding genes were distributed across *H. influenzae* isolates from Indonesia.

From over 30,000 VF-related genes in the VFDB full dataset, only 122 genes were detected via BLAST search in at least one of the *H. influenzae* genomes in this study. After further filtering using the chewBBACA Schema Evaluator module, 103 of these identified genes were confirmed as coding sequences (CDSs) and subsequently processed in the allele calling step. Based on this process, the presence/absence of each VFs-related genes was summarised in **Error! Reference source not found.** (A) and the full data was available from Supplementary Table 4.2. Most of genes encoding VFs involved as endotoxins and in iron uptake mechanism were present in more than 90% of total isolates in the dataset. In contrast, adherence-related genes were less invariably present.

Within the invasive group, 8 VF genes were observed at a higher frequency, with differences ranging from 22 to 30 percentage points compared to the carriage group, while 5 VF genes were less common, with differences of up to 34 percentage points. However, due to the small sample size of invasive isolates, these comparisons should be interpreted with caution and cannot be considered conclusive.

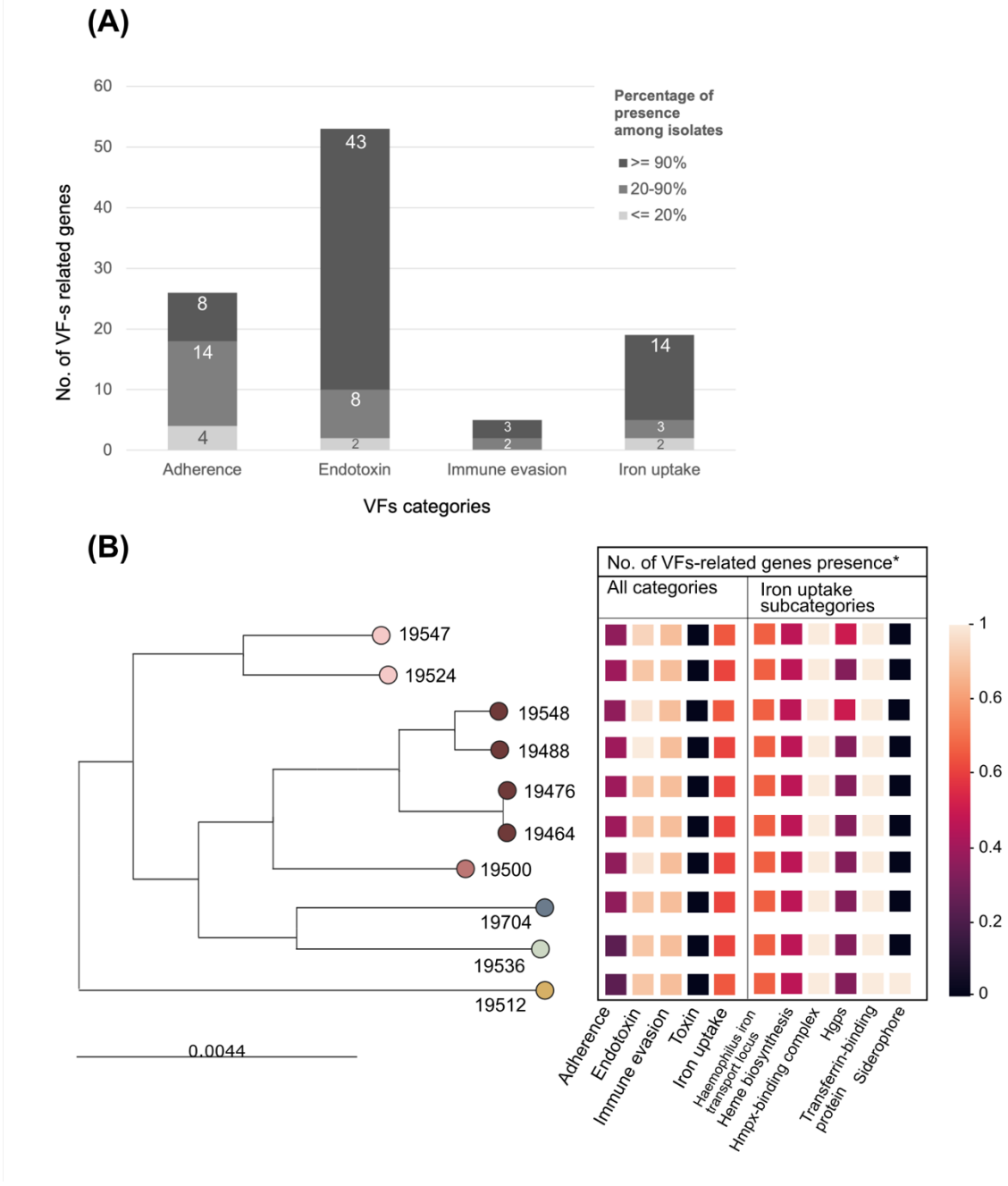


Figure 5.22. Virulence factors (VFs)-related genes present in (A) the whole collection of *H. influenzae* genomes from Indonesia (N = 113) and in (B) 10 genomes from invasive cases. **(A)** VFs-related genes were grouped according to their main function. None of the genes belonging to the "Toxin" group were present in the dataset therefore not shown in the bar. **(B)** Proportion of VFs genes within each functional group present in each invasive isolate. The tree is identical to Figure 5.3 and nodes were coloured according to the LIN code lineages. This figure was adapted from Krisna et al [297].

The presence of VF-related genes for each functional category was also assessed further amongst the invasive group, in the context of the reconstructed phylogeny (**Error! Reference source not found.** (B)). There were a consistent number of genes present for each VF category across lineages; however, there were 2 isolates (isolate ID 19536 and 19512) with a lower number of total adherence-associated genes. These isolates were detected in two patients with pleural effusion but without septicaemia. One isolate obtained from a patient with septicaemia (isolate ID 19704) had 3 additional adherence-related genes, *hifA/C/D,* which were not present in other invasive isolates.

In addition to the 103 VF-related genes commonly identified in *Haemophilus* sp., a virulence gene originating from *Acinetobacter baumannii*, *basG*, was detected in one genome (isolate ID 19512). This gene encodes histidine decarboxylase, an enzyme responsible for acinetobactin biosynthesis, which functions as a siderophore [315].

5.3.5. Everything is everywhere*: H. influenzae* from Indonesia were spread across the phylogeny of global collection of *H. influenzae*.

A total of 3,776 unique, publicly available *H. influenzae* genomes were compiled from two sources: 1) the dataset curated for the *H. influenzae* cgMLST scheme and cgLIN code (Chapter 2 and 3), and 2) the dataset used for initial GWAS for identifying genetic variants associated with invasive *H. influenze* disease (Chapter 4). After exclusion of genomes without geographical data or source of isolation (n = 1,020), there were 2,869 high-quality *H. influenzae* genomes in the final global dataset for comparison in this study (Figure 5.23).

The megalineage 0 was the largest megalineage in both global and Indonesian dataset, with most isolates belonging to the superlineage 0_0 and 0_1. However, the size of superlineage 0_0 was nearly twice as big as the superlineage 0_1 in the global dataset, while the size of each of

these superlineages was almost identical among circulating *H. influenzae* from Indonesia. Similarly, there were only a total of 18 non-Indonesian isolates in the superlineage 0_3, while 6 Indonesian isolates clustered within this superlineage. The non-Indonesian isolates mostly (n = 14) originated from countries in the Europe (Supplementary Table 4.3).

Another cluster consisting of mostly *H. influenzae* from Indonesia was the megalineage 20 with only 1 out of 9 isolates from another country, Finland (Supplementary Table 4.3). Interestingly, this isolate formed its own clonal group while the Indonesian isolates were split into 2 clonal groups. Furthermore, isolates in the clonal group 20_0_0_0 (n = 3) shared a highly similar core gene contents, with only 5 allelic mismatches maximum and they were all isolated from the same district through the same project [291].

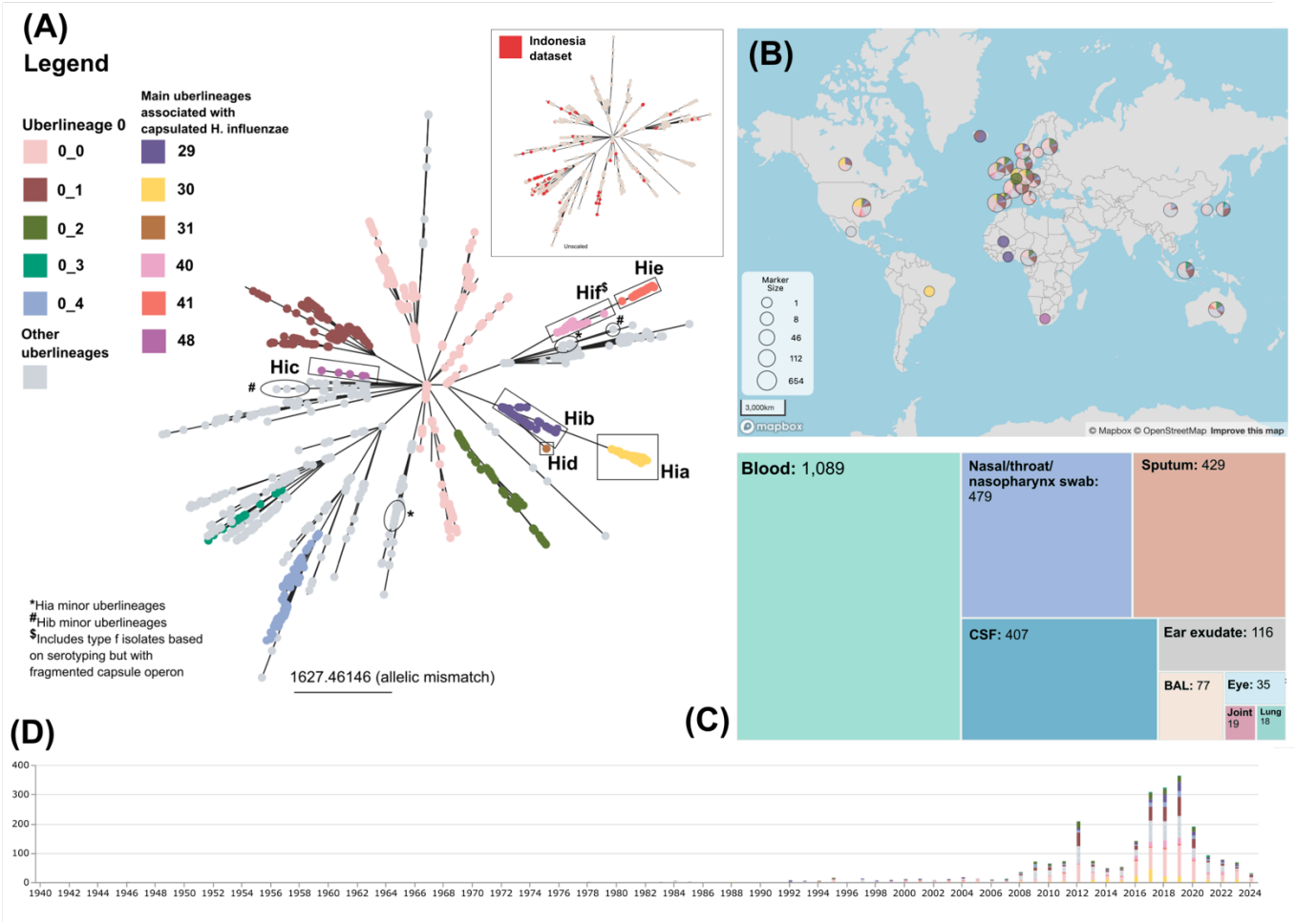


Figure 5.23. Minimum-spanning tree of 2,869 *H. influenzae* isolates from the global collection and Indonesia isolates annotated with **(A)** LIN code megalineages, superlineages, and serotype, **(B)** their geographic distribution, **(C)** source attribute, and **(D)** year of isolation. Isolates belonged to Megalineage 0 were annotated according to their superlineages within the megalineage in panel A, B, and D. Pie chart in panel B and D showed distribution of mega- and superlineages within each country and year, respectively. Not shown in panel C: *H. influenzae* genomes isolated from brain tissue (n=7), (n=5), vaginal swab (n=2), urine (n=1), lymph node (n=1), peritoneal fluid (n=1), wound swab (n=1), and those with unspecified source of isolations (n=185). *CSF: cerebrospinal fluid, BAL: bronchoalveolar lavage.*

Despite the specific observations mentioned above, *H. influenzae* isolates from Indonesia were generally distributed throughout the MST, showing no clear evidence of geographical clustering. This pattern was observed for both non-typeable and capsulated isolates: all Hib and Hif isolates from Indonesia clustered within superlineages 29 and 40, respectively, together with

global isolates of the same capsule type. As shown in Figure 5.23 (B), most major megalineages and superlineages are globally distributed, although certain countries may exhibit a higher prevalence of specific capsule-associated megalineages, for example, Hia in Canada and the U.S.A., where invasive Hia cases have been disproportionately reported among American Indian and Alaska Native (AI/AN) [33]. Lastly, most global isolates originated from normally sterile body site specimens, such as blood and CSF (Figure 5.23 (C)), consistent with public health policies in many high-income countries where routine genomic surveillance of invasive *H. influenzae* is implemented.

5.4. Discussion

H. influenzae remains a significant pathogen responsible for invasive diseases worldwide. The introduction of the polysaccharide-conjugate vaccine targeting Hib has effectively eliminated invasive disease caused by this capsule type, particularly meningitis in children under five years of age. However, invasive disease due to other strains, especially NTHi, has been increasing, predominantly affecting individuals at the extremes of age (i.e., neonates and the elderly) [20]. As of April 2025, data on *H. influenzae* from LMICs remain limited, particularly studies that incorporate genomic analyses. This study addresses that gap by focusing on circulating *H. influenzae* strains in Indonesia, archived from previous projects, with the aim of characterising their population structure through WGS.

Most of the specimens processed for *H. influenzae* identification in this study came from healthy children, with the positivity rate reaching nearly 50% in one region. A previous study on NP carriage among Indonesian children aged 12 to 24 months reported a carriage rate of 27.5% [38]. Recent data from Uganda and a meta-analysis from China reported carriage rates of 21% in

healthy children [316, 317], while a study in Italy found an even lower prevalence of 14% [318]. Several factors could account for these differences. One possible explanation is sampling bias, as this study analysed archived isolates collected from different projects, leading to variability in study populations and methodologies [319]. Another explanation is the influence of unmeasured external factors, such as household crowding, home size, attendance at day care, and the number of children in day care, all of which may increase the likelihood of upper respiratory tract carriage [320].

Although the overall carriage rate was high, this study also reflected the global trend of declining *Hib* carriage following the introduction of the *Hib* vaccine. Gessner et al. conducted the only population-based *Hib* carriage prevalence study in Indonesia before the vaccine was introduced, reporting a prevalence of 4.6% (95% CI 3.7–5.5) [289]. In Nepal, *Hib* carriage among children was 6.4% during the pre-vaccine era, whereas a follow-up study in 2018 found no *Hib* carriage among 666 healthy children aged 6 to 60 months [321]. Conversely, in countries where *Hib* vaccine has not been incorporated into the national immunisation programme, such as China, *Hib* carriage remains relatively high at 5.87% (95% CI 3.42–8.33) [322].

The introduction of the Hib vaccine has been effective in reducing Hib-related invasive disease globally. However, alongside this success, NTHi has emerged as the predominant cause of invasive *H. influenzae* infections. Reports from reference laboratories in Ontario (Canada), Paraguay, and Argentina show that NTHi accounts for 62.3%, 57.9%, and 44.5% of invasive *H. influenzae* cases, respectively [24, 323, 324]. In addition, in certain settings, other encapsulated types such as Hia have also become major causes of invasive disease [24, 325]. Despite these findings, most of the data come from high-income countries or regions outside the WHO South-East Asia Region (SEARO), and information from LMICs, including Indonesia, remains limited.

In the present study, all 10 *H. influenzae* isolates recovered from sterile clinical sites were NTHi. However, due to the absence of prior reports, it is not possible to determine whether this finding reflects a shift in *H. influenzae* invasive disease trend in Indonesia following Hib vaccine implementation.

In this study, the majority of invasive *H. influenzae* cases occurred in adults (7/10), with four classified as elderly (>60 years old, as defined by Indonesia's Presidential Legislation for National Strategy for Older People) [311], consistent with global patterns observed in the post-Hib vaccine era where NTHi disease is more common among older adults. Surveillance data from Europe (2007–2014) reported a median age of 58 years for invasive NTHi infections [207], while other studies from Spain and Queensland reported similarly high proportions of older adult cases [60, 326]. Meanwhile, two of the patients in this study were children under five years old, including a newborn. This mirrors reports from South Africa [20], Queensland [327], and Japan [6] that describe a high incidence of NTHi in infants, with many cases presenting as neonatal sepsis, the same clinical diagnosis observed in the newborn case from this study

As discussed in previous chapters, the NTHi population is highly diverse, which is evident from the long branches in the reconstructed phylogeny, even after adjusting for recombination. This diversity was observed regardless of the source of isolation (carriage or invasive disease). Two of the most frequent CCs in the dataset were associated with NTHi: the ST-3 and ST-393 complexes. A search of the *H. influenzae* PubMLST database (accessed on 1 March 2025) identified 1,192 isolate records assigned to the ST-3 complex, making it the largest clonal complex in the database. In contrast, only 78 isolates were recorded as part of the ST-393 complex. Most (8 out of 9) Indonesian isolates assigned to these two CCs were also part of superlineage 0_3, indicating that this superlineage disproportionately contained Indonesian

isolates, a pattern not visible when looking at CCs alone. Furthermore, 39 isolates in this study did not have a CC assignment. In the *H. influenzae* PubMLST database, a large number of STs (2,014 out of 2,859) are singleton STs, each represented by only one isolate. Even STs with more than one isolate often had more than four locus differences from the defined CCs, preventing their clustering into any existing CC. The implementation of the cgMLST scheme alongside the LIN code barcoding system in PubMLST allows for detailed, systematic comparisons of isolates across regions and projects, supported by a stable nomenclature. Therefore, public health authorities and researchers working on *H. influenzae* molecular epidemiology are encouraged to adopt the cgLIN coding system.

Consistent with the observations in Chapter 4, *H. influenzae* isolates from invasive cases did not form a single cluster in the reconstructed phylogeny. Based on their LIN codes, the invasive isolates in this study grouped into two categories: the first group consisted of seven isolates sharing the same megalineage LIN code, five of which also belonged to the same superlineage; the second group included three more distantly related isolates. This distribution pattern has been observed in previous studies investigating the population structure of invasive NTHi. A core genome SNP-based phylogenetic analysis of 58 invasive NTHi isolates from Spain showed that 30 of them clustered around a most recent common ancestor with short branch lengths, while the rest were more genetically diverse [60]. Similarly, data from Denmark's National Surveillance Laboratory based on 503 isolates revealed a comparable structure [23]. These findings suggest that invasive NTHi may fall into two distinct categories based on genome diversity. In contrast, an investigation of invasive NTHi isolates from Queensland, Australia, did not show this distinction [326]. These inconsistencies may be partly caused by the use of core-SNP-based phylogenetic analyses, which rely on the use of an appropriate reference genome, rather than employing de novo assembled core genome alignments [114]. Nevertheless, H.

influenzae, particularly NTHi, is known for its high capacity for horizontal gene transfer and recombination, which contributes significantly to its genomic diversification [216].

In addition to the lack of clear clustering of invasive isolates in the phylogeny, most previously known VF-related genes were consistently present across *H. influenzae* genomes in this study, regardless of the source of isolation. As this was a descriptive study with a limited number of invasive isolates, it was not possible to formally assess the relationship between VF gene presence and clinical outcomes through statistical testing. Nonetheless, two key points should be highlighted. First, as demonstrated in Chapter 4, the potential for invasiveness appears to be associated with allelic variation and changes in intergenic regions rather than the mere presence or absence of virulence genes. Therefore, a general "virulence factor" classification, such as that provided by the VFDB, is less informative for interpretative purposes (i.e. understanding why specific *H. influenzae* strains cause disease). Second, the invasive phenotype in *H. influenzae*, particularly among non-typeable strains, is likely multifactorial and not solely driven by genetic background. Instead, it is shaped by other factors such as recombination events [249, 328], epistatic interaction [329], and host-related factors [330-332].

Unlike virulence and invasive phenotype prediction, AMR determinants in *H. influenzae* appeared more consistent and could be explained by known drivers of resistance. In this study, susceptibility testing against seven antibiotics was conducted for 52 *H. influenzae* isolates using the disk diffusion method. According to a systematic review, ampicillin, ceftriaxone, cefotaxime, and moxifloxacin were among the most frequently prescribed antibiotics in inpatient settings in Indonesia, chloramphenicol was common in outpatient settings, and ciprofloxacin was widely used in both [300]. Preliminary studies in specific settings, such as a community pharmacy and a dental surgery clinic, also identified co-amoxiclav as one of the five most commonly dispensed

and prescribed antibiotics, respectively [8, 333]. In 2022, antibiotic consumption in inpatient settings reached its highest level nationally, while outpatient usage continued to rise. Additionally, pooled estimates for overall appropriate prescribing were low, at 33.5% in hospitals and 49.4% in primary health care facilities [300]. These findings suggest that *H. influenzae*, an opportunistic pathogen residing in the nasopharynx as part of the normal flora, is regularly exposed to antibiotics consumed by the human host, contributing to the selection and emergence of resistant strains.

The finding that nearly half of the isolates tested for antibiotic susceptibility were non-susceptible to at least one antibiotic, most commonly to beta-lactams, was expected, given the patterns of antibiotic use and prescribing practices described above. The rate of AMP resistance reported in this study (18/52 or 34.6%) was consistent with the pooled estimate of 36% from a meta-analysis [74]. Individual studies have reported a wide range of AMP resistance prevalence in *H. influenzae*, from as low as 8.7% [86] to as high as 51% [87]. Previous studies from Indonesia assessing the antibiotic susceptibility profile of *H. influenzae* found AMP resistance rates of 14.1% in a community setting and 38.2% in an outpatient hospital clinic [39, 288, 334]. Additionally, in this study, the majority of AMP-resistant *H. influenzae* isolates (13/18 or 72.2%) carried a beta-lactamase gene, whereas globally, the prevalence of beta-lactamase-producing *H. influenzae* has been estimated at 34.9% [74]. However, considerable variability exists across studies, with one reporting that 96% of AMP-resistant *H. influenzae* produced this enzyme [58].

All isolates lacking the *bla*TEM-1 gene, except for one, had mutated ***ftsI*** alleles classified as group II, which has been reported as the dominant group in previous studies [76]. One isolate carrying ***ftsI*** mutations consistent with the group III-like+ classification was resistant to third-generation cephalosporins tested in this study. This ***ftsI*** group has been linked to elevated MICs

for AMP and CTA [76]. The finding of only a single *H. influenzae* isolate exhibiting resistance to third-generation cephalosporins agrees with previous reports showing low prevalence of resistance to this antibiotic class [88, 288]. Nevertheless, these finding warrants attention, as this antibiotic class remains a first-line treatment option for suspected cases of bacterial meningitis [335].

Seven of the 52 *H. influenzae* isolates in this study were resistant to CHL, six of which also showed concurrent resistance to AMP. The *catII* gene, the only known determinant of CHL resistance in *H. influenzae* [336], was detected in all seven isolates and was carried on the MGE ICEhin1056. Most reports on CHL resistance in *H. influenzae* predate 2000; one large-scale study of over 8,000 isolates reported a prevalence of just 1.9% [337], significantly lower than both the rate observed in this study and a more recent pooled estimate of 17.2%. This emerging evidence of increasing CHL resistance is cause for concern, as CHL continues to be the first-line treatment option for bacterial meningitis in patients with severe beta-lactam allergies [335].

In contrast to beta-lactam and CHL, fluoroquinolone resistance was relatively rare. Only two isolates in this study were resistant to fluoroquinolones. Published resistance rates range from 2% in Germany to 3.7% in Indonesia [58, 338], while pooled estimates indicate lower resistance to CIP (1.7%) and higher resistance to MOX (7.5%) [74]. Resistance to fluoroquinolones was conferred by mutations in genes encoding the target enzymes of these antibiotics.

The main limitation of this study in interpreting the AST data was the incomplete testing, as not all isolates underwent susceptibility analysis. This limited the ability to observe AMR patterns in relation to the bacterial population structure. However, it was noted that seven resistant *H. influenzae* isolates clustered within the clonal group 0_1_0_0, with five carrying resistance associated with the MGE ICEhin1056 and two due to point mutations. This suggests

the presence of non-random associations between specific genetic backgrounds and resistance mechanisms, whether plasmid-borne or mutation-driven, which may be shaped by global epistatic interactions that facilitate or restrict the acquisition and persistence of AMR determinants, as has been proposed in other bacterial pathogens [339, 340].

5.5. Conclusion and future directions

This study presented the first whole-genome-based investigation of the population structure of *H. influenzae* circulating in Indonesia from both carriage and invasive disease cases. Employing the recently developed cgMLST scheme and LIN code hierarchical clustering, high-resolution genetic lineages to 113 *H. influenzae* isolates were successfully assigned. The cgLIN barcoding system, already integrated into the PubMLST platform, enabled systematic delineation of clonal relationships among isolates and allowed comparison across studies with stable nomenclature. Data from this study showed that, similar to global observations, Indonesian *H. influenzae* populations were dominated by highly diverse non-typeable strains (NTHi), with encapsulated types (Hib, Hif) appearing only sporadically and exclusively in carriage samples.

Despite the overall similarity to global lineages, several lineage-level features potentially unique to the Indonesian population were identified. Notably, isolates belonging to superlineage 0_3 were overrepresented in Indonesia compared to global data. Furthermore, within superlineage 0_1, there was one clonal group (0_1_0_0) containing multiple resistant isolates, all of which shared closely related cgLIN profiles. This observation underscored the utility of cgLIN in identifying and tracking emerging resistant lineages, which may otherwise go unrecognised with conventional 7-locus MLST alone.

Our findings highlight the growing burden of AMR among *H. influenzae* isolates in Indonesia, with nearly half of tested strains showing non-susceptibility to at least one antibiotic, particularly beta-lactams. Resistance was mainly associated with either the presence of MGE carrying *bla*TEM-1 or chromosomal mutations such as those in *ftsI*. In contrast, VFs appeared broadly conserved across isolates from both carriage and disease, with no distinct profiles separating invasive from non-invasive strains. These observations suggest that while AMR was largely driven by identifiable genetic mechanisms, the invasive potential of NTHi may be shaped by more complex genetic and host-related factors.

A key limitation of this study was the incomplete antibiotic susceptibility data, which limited the ability to conduct more robust genotype-phenotype association analyses across the full dataset. Nevertheless, the observation that genetically similar isolates shared resistance profiles suggests the possible influence of epistatic interactions between genetic background and resistance determinants, an emerging concept in bacterial evolution and AMR biology. Future studies using larger, prospectively collected datasets would be valuable to confirm these associations.

Moving forward, several directions are critical. First, the implementation of cgMLST and cgLIN-based typing for *H. influenzae* in public health laboratories should be encouraged, enabling real-time surveillance and outbreak detection. Second, comprehensive longitudinal studies incorporating host data, clinical outcomes, and environmental context will be crucial to dissecting the multifactorial basis of invasiveness in NTHi. Third, expanded AMR monitoring using both phenotypic and genomic approaches is essential to inform empirical treatment guidelines, especially as resistance to first-line agents like ampicillin and chloramphenicol continues to rise. Lastly, inclusion of carriage isolates in future genomic surveillance is necessary

to capture the full spectrum of circulating *H. influenzae* diversity and pre-emptively identify lineages with epidemic or resistance potential. Collectively, this study established a critical foundation for genomic epidemiology of *H. influenzae* in Indonesia and contributes to broader efforts toward evidence-based interventions in LMICs.

Chapter 6

General Discussion and Conclusion

This chapter provides a comprehensive overview of key findings from the thesis and explains how these results have been interpreted. It details inferences drawn and discusses implications for advancing our understanding of H. influenzae molecular epidemiology, genome biology, and AMR. The overview was organised according to the thesis objectives, followed by a discussion on the implications for public health policy and future H. influenzae research. Finally, this chapter outlines the limitations of the project and suggests potential directions for future research.

6.1. Key findings

The primary objective of this thesis was to address critical knowledge gaps in the molecular epidemiology and genomic diversity of *H. influenzae*. To achieve this, three key aspects were focussed on: (1) establishing a stable nomenclature system for high-resolution strain characterization and population structure delineation in *H. influenzae*, (2) identifying genetic variants associated with the invasive-disease phenotype in NTHi , and (3) investigating the population genomics of *H. influenzae* circulating in Indonesia, where genomic surveillance data are scarce.

6.1.1. Establishing core genome LIN code nomenclature as the standard in *H. influenzae* population genetics.

A cgMLST scheme extends the existing MLST framework by characterizing bacterial strains and capturing their population diversity [47, 48], typically within a specific species [174]. In Chapter 2, this scheme was developed and validated for *H. influenzae*, encompassing 1,037 core genes. Adopting a cgMLST scheme for *H. influenzae* offers two fundamental advantages. First, as a gene-by-gene approach, it does not require a reference genome, an important benefit for highly recombining species [103, 117]. Second, it treats each allelic change as a single event, a feature that is also useful for organisms with extensive recombination [63].

Another major benefit of this approach is that the scheme was implemented in a PubMLST database powered by BIGSdb, which supports the automatic annotation of user-submitted genomes [49]. In developing the *H. influenzae* scheme, particular attention was paid to ensuring that the 1,037 core genes could be consistently annotated by this automated process, a goal achieved through a dedicated validation step.

Due to its inclusion of a large number of loci, the cgMLST scheme provides high granular characterisation of bacterial isolates. In a genetically diverse species such as *H. influenzae*, this granularity results in each genome receiving a distinct cgST. As of 15 April 2025, the PubMLST *H. influenzae* database contained 16,913 defined core genome profiles. In public health and research applications, key objectives- often include quantifying genomic similarity among collections of *H. influenzae* isolates and identifying those associated with previously characterised high-risk clusters. Consequently, the development of a clustering strategy for such a highly granular scheme is as essential as the scheme itself [173]. To address this requirement,

a LIN code nomenclature system [175] based on core genome profiles (cgLIN) was proposed for *H. influenzae* [174].

Initially implemented in *K. pneumoniae*, the cgLIN approach offers four advantages for cluster delineation within bacterial species [174]. First, and most importantly, compared to the single-linkage clustering approach that assigns isolates to CGCs, as described in Chapter 2, the cgLIN approach prevents the merging of distinct CGCs when new genomes are added [172, 173]. This is possible because the barcode generated by cgLIN embeds stable clustering information at multiple thresholds [174, 175]. Thus, new genomes are assigned barcodes consistent with previously established clusters across these thresholds. However, it is crucial to begin cgLIN implementation with a representative collection of genomes [174]. Although selection bias cannot be entirely eliminated (e.g., isolates from invasive infections are often overrepresented in existing literature), in Chapters 2 and 3, the dataset curation process and its outcomes were presented, highlighting the efforts made to ensure the dataset was as representative and genetically diverse as possible. The dataset characterisation and visualisation provided clear evidence that these goals were achieved.

Second, the cgLIN approach is hierarchical, allowing users to select appropriate clustering levels for their specific research interests. In Chapter 3, this flexibility was demonstrated using published datasets, employing higher-level clustering to identify groups of clinically significant *H. influenzae* strains and lower-level clustering to confirm or exclude potential outbreak scenarios. Furthermore, in Chapter 5, cgLIN clustering facilitated the straightforward and consistent characterisation of a novel set of *H. influenzae* genomes from Indonesia, aligning closely with the reconstructed phylogeny. This approach significantly enhanced the reliability, reproducibility, and interpretability of associations between isolate characteristics (e.g., source

of isolation or phenotypic traits such as resistance) and population structure, and simplified comparisons with global datasets.

Third, the cgLIN approach is integrated into the PubMLST database, ensuring that each *H. influenzae* genome uploaded to the database automatically receives both cgLIN and cgMLST assignments. This integration guarantees the reproducibility, accuracy, and continuity of the scheme and its associated nomenclature. These attributes were lacking in the original LIN code concept, which was developed using the ANI metric [174].

Finally, because the LIN code is derived from a cgMLST scheme encompassing a large number of loci, the impact of recombination on cluster assignments is minimised [120]. As demonstrated in Chapters 3 and 5, isolates sharing identical STs or CCs formed polyphyletic groups in reconstructed phylogenies [174]. cgLIN uberlineage and superlineage assignments accurately captured these phylogenetic distances despite extensive recombination. These proved that by incorporating a large number of genes, cgST and cgLIN clusters integrate signals from recombined and non-recombined loci, yielding robust lineage delineation.

Given its clear advantages and successful implementation in other major human pathogens, including *S. pneumoniae* [341] and *N. gonorrhoeae* [342], the cgLIN approach is recommended be adopted as the standard for defining lineages in *H. influenzae* and other clinically important bacterial species.

6.1.2. Complex genetic architecture of invasive NTHi: Variants in core virulence genes and noncoding regions.

In the post-Hib vaccine era, NTHi has become the leading cause of invasive *H. influenzae* disease worldwide, yet the genetic basis for its invasiveness remains unclear. In Chapter 4, the

application of the cgLIN framework, corroborated by phylogenetic analysis, demonstrated that no specific NTHi lineages were preferentially associated with invasive disease [343], in contrast to pathogens such as *N. meningitidis*. GWAS was then conducted to identify genetic determinants linked to invasive NTHi infection. In this analysis, cgLIN-based lineage definitions were used to control for lineage effects, underscoring the utility of the cgLIN approach in microbial genomics research.

Chapter 4 employed a k-mer–based GWAS capable of detecting all genetic variants, both gene presence/absence and nucleotide-level allelic changes, associated with invasive phenotype [225]. Of the four protein-coding loci identified, three encoded established *H. influenzae* virulence factors: a porin, a hemoglobin/haptoglobin receptor, and a TonB-dependent receptor. These surface-exposed proteins are subject to constant immune selection and therefore exhibit- high allelic diversity. Phase variation [221] and recombination [96, 244] are the primary drivers of diversity, consistent with the presence of recombination hotspots at these loci. As a result, simple presence or absence of virulence factor genes cannot reliably distinguish invasive from non-invasive variants or predict the invasive potential of individual NTHi isolates.

Most k-mers associated with invasive NTHi mapped to non–protein coding or intergenic regions. This aspect of *H. influenzae* genome biology remains largely uncharacterised. However, studies in other organisms suggest that such noncoding sequences may serve regulatory functions. Furthermore, additional factors, such as epistatic interactions and host determinants, were not considered in this analysis.

Overall, Chapter 4 reaffirmed that no novel virulence determinants linked to invasive NTHi infection were identified, nor is there a single factor that fully explains this complex phenotype. However, these results help to focus future vaccine development efforts, particularly

since Chapters 4 and 5 demonstrated that all variant-associated coding sequences were core genes. Moving forward, the critical task is to map regions of these antigens that are both conserved and strongly immunogenic.

6.1.3. High prevalence of NTHi carriage and beta-lactam resistance in Indonesia: Genomic insights via cgLIN scheme

In Chapter 5, a high NP carriage rate of *H. influenzae* in Indonesia was reported [38]. Because the majority of isolates were NTHi, this elevated rate is unlikely to reflect low Hib vaccine coverage or vaccine failure. Unexamined factors, such as household and day care overcrowding, may contribute to increased transmission [320]. Comparative genomic analysis, using cgLIN assignment against a curated global dataset, revealed no evidence of clonal expansion underlying the higher carriage rate. Moreover, the previously observed lack of phylogeographic signal in *H. influenzae* persists, despite the overrepresentation of superlineages 0_1 and 0_3 in the Indonesian cohort. Definitive assessment of these associations will require a large-scale, prospective carriage surveillance study with WGS to control for selection bias.

Beta-lactam resistance was common among the *H. influenzae* isolates tested. Although the observed resistance rate was similar to the cumulative estimate from a recent meta-analysis, it is concerning that most resistant strains carried beta-lactamase genes on MGEs. This pattern likely reflects strong selection pressure from inappropriate antibiotic prescribing and use, which has allowed MGE bearing strains to proliferate. In conjunction with the high carriage rate, these MGEs may therefore spread readily between *H. influenzae* isolates and from one host to another.

Once again, application of the cgLIN scheme enabled rapid assignment of each resistant *H. influenzae* isolate to a defined lineage. However, inherent limitations of this study (see Limitations) precluded determining whether resistance emergence was linked to specific lineages or clonal groups.

6.2. Limitations

In Chapters 2, 3, and 4, a large collection of publicly available *H. influenzae* genomes were drawn from diverse regions. Although this dataset exceeded typical sample sizes for population-genomic analyses [254], it remains vulnerable to sampling bias: isolates of clinical interest, particularly those from invasive cases or exhibiting antibiotic resistance, are preferentially sequenced, potentially underrepresenting the full breadth of population diversity. Furthermore, the cgLIN nomenclature depends on a thoroughly validated, species-specific cgMLST scheme and on clustering thresholds calibrated to the underlying population structure. While *H. influenzae* exhibits clear genetic discontinuities that guided threshold selection, other species may present more ambiguous patterns, and cgLIN cannot be applied universally across genera.

Second, the GWAS in Chapter 4 was constrained by the limitations of a cross-sectional design, offering only a single timepoint snapshot and omitting host factor covariates that could modulate invasive potential. Longitudinal sampling, coupled with functional validation, will be necessary to dissect the interplay between genetic variation, recombination dynamics, and host susceptibility.

Lastly, in Chapter 5, the investigation of NP carriage and beta-lactam resistance in Indonesia faced several constraints. Archived isolates were drawn from multiple projects with variable collection protocols, and although NP swab selection was randomised, the cohort may

still not fully represent the broader population. Furthermore, the small number of invasive isolates made formal statistical testing unfeasible, thereby limiting direct comparisons between carriage and invasive groups. Finally, antibiotic susceptibility testing, performed by disk diffusion on only a subset of strains, precluded determination of MIC and prevented comprehensive antimicrobial resistance profiling across the entire dataset.

6.3. Implications for public health and research and future directions

The cgMLST scheme and accompanying cgLIN code developed in this thesis establish a robust, hierarchical nomenclature for defining *H. influenzae* lineages, thereby enhancing our understanding of species population dynamics. Integrated into the PubMLST database, these tools allow users, whether in public health, clinical, or research settings, to assign newly sequenced genomes to well-characterised clusters. In an era of rapidly advancing sequencing technology, accurate lineage delineation is indispensable for outbreak detection, antimicrobial-resistance monitoring, and rational vaccine design.

Building on this stable genomic framework, the GWAS presented here reaffirmed the central role of known VFs in invasive *H. influenzae* disease. These core proteins arguably emerge as prime candidates for a universal vaccine that extends protection beyond serotype b. To move from discovery to deployment, *in vitro* assays must validate the immunogenicity of conserved antigenic regions, followed by clinical trials to assess safety and efficacy across diverse *H. influenzae* lineages.

Translating these advances into public health impact requires sustained diagnostic capacity and genomic surveillance. In Indonesia, where *H. influenzae* data have been confined to discrete research projects, limited laboratory resources have constrained both culture-based

isolation and whole genome sequencing. Our findings, some of which diverge from the established molecular epidemiology of the species, underscore the urgency of investing in routine, nationwide surveillance programs. Such infrastructure will be essential to track lineage emergence, monitor resistance trends, and guide vaccine implementation on a global scale.

6.4. Conclusion

In summary, this thesis establishes a robust genomic framework for *H. influenzae* by developing a 1,037-locus cgMLST scheme and hierarchical cgLIN nomenclature within PubMLST, enabling stable, reproducible, high resolution lineage assignments essential for outbreak investigation, AMR monitoring, and global comparisons. Applying this framework in a k-mer–based GWAS confirmed that core virulence factors remain central to invasive NTHi and identified conserved antigenic regions as promising universal vaccine targets. Extension of these methods to *H. influenzae* isolates in Indonesia revealed high colonisation rates and widespread beta lactam resistance driven by MGEs and point mutations, underscoring critical gaps in diagnostic capacity and the urgency of routine, nationwide genomic surveillance. Together, these findings not only deepen our understanding of *H. influenzae* population dynamics but also provide the tools and direction needed for rational vaccine design, longitudinal studies of host–pathogen interactions, and strengthened laboratory infrastructure to meet future public health-challenges.

References

1. Carrol KC, Funke G, Landry ML, Richter SS, Warnock DW: *Manual of Clinical Microbiology*. Washington, DC: ASM Press; 2019.
2. Brooks GF, Jawetz E, Melnick JL, Adelberg EA: *Jawetz, Melnick, & Adelberg's medical microbiology*. New York: McGraw Hill Medical; 2019.
3. Potts CC, Topaz N, Rodriguez-Rivera LD, Hu F, Chang HY, Whaley MJ, Schmink S, Retchless AC, Chen A, Ramos E, et al: **Genomic characterization of Haemophilus influenzae: a focus on the capsule locus.** *BMC Genomics* 2019, **20:**733.
4. Satola SW, Schirmer PL, Farley MM: **Complete sequence of the cap locus of Haemophilus influenzae serotype b and nonencapsulated b capsule-negative variants.** *Infect Immun* 2003, **71:**3639-3644.
5. Soeters HM, Blain A, Pondo T, Doman B, Farley MM, Harrison LH, Lynfield R, Miller L, Petit S, Reingold A, et al: **Current Epidemiology and Trends in Invasive Haemophilus influenzae Disease-United States, 2009-2015.** *Clin Infect Dis* 2018, **67:**881-889.
6. Suga S, Ishiwada N, Sasaki Y, Akeda H, Nishi J, Okada K, Fujieda M, Oda M, Asada K, Nakano T, et al: **A nationwide population-based surveillance of invasive Haemophilus influenzae diseases in children after the introduction of the Haemophilus influenzae type b vaccine in Japan.** *Vaccine* 2018, **36:**5678-5684.
7. Procop GW, Church DL, Hall GS, Janda WM, Koneman EW, Schreckenberger PC, Woods GL: *Koneman's Color Atlas & Textbook of Diagnostic Microbiology*. Philadelphia: Wolters Kluwer; 2017.
8. Bakaletz LO, Novotny LA: **Nontypeable Haemophilus influenzae (NTHi).** *Trends Microbiol* 2018, **26:**727-728.
9. Cifuente JO, Schulze J, Bethe A, Di Domenico V, Litschko C, Budde I, Eidenberger L, Thiesler H, Ramon Roth I, Berger M, et al: **A multi-enzyme machine polymerizes the Haemophilus influenzae type b capsule.** *Nat Chem Biol* 2023, **19:**865-877.
10. World Health O, Centers for Disease C, Prevention: **Laboratory methods for the diagnosis of meningitis caused by neisseria meningitidis, streptococcus pneumoniae, and haemophilus influenzae: WHO manual.** 2nd ed edition. Geneva: World Health Organization; 2011.
11. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q: **VFDB: a reference database for bacterial virulence factors.** *Nucleic Acids Res* 2005, **33:**D325-328.
12. Liu B, Zheng D, Zhou S, Chen L, Yang J: **VFDB 2022: a general classification scheme for bacterial virulence factors.** *Nucleic Acids Res* 2022, **50:**D912-D917.
13. Erwin AL, Smith AL: **Nontypeable Haemophilus influenzae: understanding virulence and commensal behavior.** *Trends Microbiol* 2007, **15:**355-362.
14. Langereis JD, de Jonge MI: **Unraveling Haemophilus influenzae virulence mechanisms enable discovery of new targets for antimicrobials and vaccines.** *Curr Opin Infect Dis* 2020, **33:**231-237.
15. Wen S, Mai Y, Chen X, Xiao K, Lin Y, Xu Z, Yang L: **Molecular Epidemiology and Antibiotic Resistance Analysis of Non-Typeable Haemophilus influenzae (NTHi) in Guangzhou: A Representative City of Southern China.** *Antibiotics (Basel)* 2023, **12**.

16. Bailey KL, LeVan TD, Yanov DA, Pavlik JA, DeVasure JM, Sisson JH, Wyatt TA: **Non-typeable Haemophilus influenzae decreases cilia beating via protein kinase Cepsilon.** *Respir Res* 2012, **13:**49.
17. Jalalvand F, Riesbeck K: **Haemophilus influenzae: recent advances in the understanding of molecular pathogenesis and polymicrobial infections.** *Curr Opin Infect Dis* 2014, **27:**268-274.
18. Duell BL, Su YC, Riesbeck K: **Host-pathogen interactions of nontypeable Haemophilus influenzae: from commensal to pathogen.** *FEBS Lett* 2016, **590:**3840-3853.
19. Gilsdorf JR: **Hib Vaccines: Their Impact on Haemophilus influenzae Type b Disease.** *J Infect Dis* 2021, **224:**S321-S330.
20. Slack MPE, Cripps AW, Grimwood K, Mackenzie GA, Ulanova M: **Invasive Haemophilus influenzae Infections after 3 Decades of Hib Protein Conjugate Vaccine Use.** *Clin Microbiol Rev* 2021, **34:**e0002821.
21. Hani E, Abdullahi F, Bertran M, Eletu S, D'Aeth J, Litt DJ, Fry NK, Ladhani SN: **Trends in invasive Haemophilus influenzae serotype b (Hib) disease in England: 2012/13 to 2022/23.** *J Infect* 2024, **89:**106247.
22. McTaggart LR, Cronin K, Seo CY, Wilson S, Patel SN, Kus JV: **Increased Incidence of Invasive Haemophilus influenzae Disease Driven by Non-Type B Isolates in Ontario, Canada, 2014 to 2018.** *Microbiol Spectr* 2021, **9:**e0080321.
23. Slotved HC, Johannesen TB, Stegger M, Fuursted K: **Evaluation of molecular typing for national surveillance of invasive clinical Haemophilus influenzae isolates from Denmark.** *Front Microbiol* 2022, **13:**1030242.
24. Efron A, Napoli D, Neyro S, Juarez MDV, Moscoloni M, Eluchans NS, Regueira M, Lavayen S, Argentinean HiWG, Faccone D, Santos M: **Laboratory surveillance of invasive Haemophilus influenzae disease in Argentina, 2011-2019.** *Rev Argent Microbiol* 2023, **55:**133-142.
25. Zanella RC, Bokermann S, Galhardo M, Gava C, Almeida SCG, Pereira GA, de Lemos APS: **Trends in serotype distribution and antimicrobial susceptibility pattern of invasive Haemophilus influenzae isolates from Brazil, 2009-2021.** *Int Microbiol* 2025, **28:**157-163.
26. Intusoma U, Thewamit R, Thamcharoenvipas T, Khantee P: **Epidemiology and burden of Haemophilus influenzae disease in Thai children before implementation of the routine immunisation programme: A National Health Data Analysis.** *Trop Med Int Health* 2022, **27:**546-552.
27. Hefele L, Lai J, Vilivong K, Bounkhoun T, Chanthaluanglath V, Chanthongthip A, Balloch A, Black AP, Hubschen JM, Russell FM, Muller CP: **Haemophilus influenzae serotype b seroprevalence in central Lao PDR before and after vaccine introduction.** *PLoS One* 2022, **17:**e0274558.
28. Hong E, Terrade A, Denizon M, Aouiti-Trabelsi M, Falguieres M, Taha MK, Deghmane AE: **Haemophilus influenzae type b (Hib) seroprevalence in France: impact of vaccination schedules.** *BMC Infect Dis* 2021, **21:**715.
29. Marques JG, Inacio Cunha FM, Bajanca-Lavado MP, Portuguese Study Group on Haemophilus influenzae Invasive Disease in C: **Haemophilus influenzae Type b Vaccine Failure in Portugal: A Nationwide Multicenter Pediatric Survey.** *Pediatr Infect Dis J* 2023, **42:**824-828.

30. Zhang H, Garcia C, Yu W, Knoll MD, Lai X, Xu T, Jing R, Qin Y, Yin Z, Wahl B, Fang H: **National and provincial impact and cost-effectiveness of Haemophilus influenzae type b conjugate vaccine in China: a modeling analysis.** *BMC Med* 2021, **19:**181.
31. Weinberg MM, Akel K, Akinyemi O, Balasubramanian T, Blankenship HM, Collins JP, Collins J, Henderson T, Johnson S, Lai J, et al: **Invasive Nontypeable Haemophilus influenzae Disease Outbreak at an Elementary School - Michigan, May 2023.** *MMWR Morb Mortal Wkly Rep* 2024, **73:**691-695.
32. Bertran M, D'Aeth JC, Hani E, Amin-Chowdhury Z, Fry NK, Ramsay ME, Litt DJ, Ladhani SN: **Trends in invasive Haemophilus influenzae serotype a disease in England from 2008-09 to 2021-22: a prospective national surveillance study.** *Lancet Infect Dis* 2023, **23:**1197-1206.
33. Brown NE, Blain AE, Burzlaff K, Harrison LH, Petit S, Schaffner W, Smelser C, Thomas A, Triden L, Watt JP, et al: **Racial Disparities in Invasive Haemophilus influenzae Disease-United States, 2008-2017.** *Clin Infect Dis* 2021, **73:**1617-1624.
34. Frankel C, Robinson J, Khan S, Alghounaim M, McDonald J, Lopez A, Fanella S, Gunawan J, Wong J, Comeau J, et al: **A Pediatric Investigators Collaborative Network on Infections in Children (PICNIC) multi-centre Canadian descriptive analysis of Haemophilus influenzae bacteremia in children: Emerging serotypes.** *Can Commun Dis Rep* 2023, **49:**368-374.
35. Reilly AS, McElligott M, Mac Dermott Casement C, Drew RJ: **Haemophilus influenzae type f in the post-Haemophilus influenzae type b vaccination era: a systematic review.** *J Med Microbiol* 2022, **71**.
36. Salvador EC, Buddha N, Bhola A, Sinha SK, Kato M, Wijesinghe PR, Samuel R, Naidoo D, Singh SK, Perera WLS, Singh PK: **Health Emergency Risk Management in World Health Organization – South-East Asia Region during 2014–2023: synthesis of experiences.** *The Lancet Regional Health - Southeast Asia* 2023, **18**.
37. Rahman M, Hossain S, Baqui AH, Shoma S, Rashid H, Nahar N, Zaman MK, Khatun F: **Haemophilus influenzae type-b and non-b-type invasive diseases in urban children (<5years) of Bangladesh: implications for therapy and vaccination.** *J Infect* 2008, **56:**191-196.
38. Dunne EM, Murad C, Sudigdoadi S, Fadlyana E, Tarigan R, Indriyani SAK, Pell CL, Watts E, Satzke C, Hinds J, et al: **Carriage of Streptococcus pneumoniae, Haemophilus influenzae, Moraxella catarrhalis, and Staphylococcus aureus in Indonesian children: A cross-sectional study.** *PLoS One* 2018, **13:**e0195098.
39. Safari D, Wahyono DJ, Tafroji W, Darmawan AB, Winarti Y, Kusdaryanto WD, Paramaiswari WT, Pramono H, Pratiwi M, Chamadi MR: **Serotype Distribution and Antimicrobial Resistance Profile of Haemophilus influenzae Isolated from School Children with Acute Otitis Media.** *Int J Microbiol* 2022, **2022:**5391291.
40. Lokida D, Farida H, Triasih R, Mardian Y, Kosasih H, Naysilla AM, Budiman A, Hayuningsih C, Anam MS, Wastoro D, et al: **Epidemiology of community-acquired pneumonia among hospitalised children in Indonesia: a multicentre, prospective study.** *BMJ Open* 2022, **12:**e057957.
41. Norskov-Lauritsen N: **Classification, identification, and clinical significance of Haemophilus and Aggregatibacter species with host specificity for humans.** *Clin Microbiol Rev* 2014, **27:**214-240.

42. Simar SR, Hanson BM, Arias CA: **Techniques in bacterial strain typing: past, present, and future.** *Curr Opin Infect Dis* 2021, **34:**339-345.
43. Uelze L, Grutzke J, Borowiak M, Hammerl JA, Juraschek K, Deneke C, Tausch SH, Malorny B: **Typing methods based on whole genome sequencing data.** *One Health Outlook* 2020, **2:**3.
44. Francisco AP, Bugalho M, Ramirez M, Carrico JA: **Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach.** *BMC Bioinformatics* 2009, **10:**152.
45. Sabat AJ, Budimir A, Nashev D, Sa-Leao R, van Dijl J, Laurent F, Grundmann H, Friedrich AW, Markers ESGoE: **Overview of molecular typing methods for outbreak detection and epidemiological surveillance.** *Euro Surveill* 2013, **18:**20380.
46. Nutman A, Marchaim D: **How to: molecular investigation of a hospital outbreak.** *Clin Microbiol Infect* 2019, **25:**688-695.
47. Maiden MC: **Multilocus sequence typing of bacteria.** *Annu Rev Microbiol* 2006, **60:**561-588.
48. Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND: **MLST revisited: the gene-by-gene approach to bacterial genomics.** *Nat Rev Microbiol* 2013, **11:**728-736.
49. Jolley KA, Bray JE, Maiden MCJ: **Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications.** *Wellcome Open Res* 2018, **3:**124.
50. Jolley KA, Maiden MC: **BIGSdb: Scalable analysis of bacterial genome variation at the population level.** *BMC Bioinformatics* 2010, **11:**595.
51. Meats E, Feil EJ, Stringer S, Cody AJ, Goldstein R, Kroll JS, Popovic T, Spratt BG: **Characterization of encapsulated and noncapsulated Haemophilus influenzae and determination of phylogenetic relationships by multilocus sequence typing.** *J Clin Microbiol* 2003, **41:**1623-1636.
52. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al.: **Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.** *Science* 1995, **269:**496-512.
53. Iskander M: **Development and Evaluation of Core Genome MLST Schema for Haemophilus influenzae.** University of Manitoba, Department of Medical Microbiology and Infectious Disease; 2017.
54. Pinto M, Gonzalez-Diaz A, Machado MP, Duarte S, Vieira L, Carrico JA, Marti S, Bajanca-Lavado MP, Gomes JP: **Insights into the population structure and pan-genome of Haemophilus influenzae.** *Infect Genet Evol* 2019, **67:**126-135.
55. Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carrico JA: **chewBBACA: A complete suite for gene-by-gene schema creation and strain identification.** *Microb Genom* 2018, **4**.
56. Zallot R, Harrison KJ, Kolaczkowski B, de Crecy-Lagard V: **Functional Annotations of Paralogs: A Blessing and a Curse.** *Life (Basel)* 2016, **6**.
57. Zhang J, Halkilahti J, Hanninen ML, Rossi M: **Refinement of whole-genome multilocus sequence typing analysis by addressing gene paralogy.** *J Clin Microbiol* 2015, **53:**1765-1767.
58. Frank T, Wohlfarth E, Claus H, Krone M, Lam TT, Kresken M, Study Group 'Antimicrobial Resistance' of the Paul Ehrlich Society for Infection T: **Antibiotic resistance and molecular characterization of non-invasive clinical Haemophilus**

influenzae isolates in Germany 2019 and 2020. *JAC Antimicrob Resist* 2024, **6:**dlae197.
59. Nurnberg S, Claus H, Krone M, Vogel U, Lam TT: **Cefotaxime resistance in invasive Haemophilus influenzae isolates in Germany 2016-19: prevalence, epidemiology and relevance of PBP3 substitutions.** *J Antimicrob Chemother* 2021, **76:**920-929.
60. Carrera-Salinas A, Gonzalez-Diaz A, Calatayud L, Mercado-Maza J, Puig C, Berbel D, Camara J, Tubau F, Grau I, Dominguez MA, et al: **Epidemiology and population structure of Haemophilus influenzae causing invasive disease.** *Microb Genom* 2021, **7**.
61. De Chiara M, Hood D, Muzzi A, Pickard DJ, Perkins T, Pizza M, Dougan G, Rappuoli R, Moxon ER, Soriani M, Donati C: **Genome sequencing of disease and carriage isolates of nontypeable Haemophilus influenzae identifies discrete population structure.** *Proc Natl Acad Sci U S A* 2014, **111:**5439-5444.
62. Wailan AM, Coll F, Heinz E, Tonkin-Hill G, Corander J, Feasey NA, Thomson NR: **rPinecone: Define sub-lineages of a clonal expansion via a phylogenetic tree.** *Microb Genom* 2019, **5**.
63. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, et al: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.** *Proc Natl Acad Sci U S A* 1998, **95:**3140-3145.
64. Topaz N, Tsang R, Deghmane AE, Claus H, Lam TT, Litt D, Bajanca-Lavado MP, Perez-Vazquez M, Vestrheim D, Giufre M, et al: **Phylogenetic Structure and Comparative Genomics of Multi-National Invasive Haemophilus influenzae Serotype a Isolates.** *Front Microbiol* 2022, **13:**856884.
65. Gonzalez-Diaz A, Carrera-Salinas A, Pinto M, Cubero M, van der Ende A, Langereis JD, Dominguez MA, Ardanuy C, Bajanca-Lavado P, Marti S: **Comparative pangenome analysis of capsulated Haemophilus influenzae serotype f highlights their high genomic stability.** *Sci Rep* 2022, **12:**3189.
66. Wall EC, Taha MK: **Haemophilus influenzae is fighting back: is serotype a an emerging threat?** *Lancet Infect Dis* 2023, **23:**1106-1108.
67. Slotved HC, Johannesen TB, Stegger M, Dalby T, Fuursted K: **National Danish surveillance of invasive clinical Haemophilus influenzae isolates and their resistance profile.** *Front Microbiol* 2023, **14:**1307261.
68. Tonnessen R, Garcia I, Debech N, Lindstrom JC, Wester AL, Skaare D: **Molecular epidemiology and antibiotic resistance profiles of invasive Haemophilus influenzae from Norway 2017-2021.** *Front Microbiol* 2022, **13:**973257.
69. Sakamoto N, Hitomi S: **Clinical and microbiological characteristics of invasive diseases due to Haemophilus influenzae in the Minami Ibaraki Area, Japan.** *J Infect Chemother* 2025, **31:**102633.
70. Collins LF, Havers FP, Tunali A, Thomas S, Clennon JA, Wiley Z, Tobin-D'Angelo M, Parrott T, Read TD, Satola SW, et al: **Invasive Nontypeable Haemophilus influenzae Infection Among Adults With HIV in Metropolitan Atlanta, Georgia, 2008-2018.** *JAMA* 2019, **322:**2399-2410.
71. **Global Antimicrobial Resistance and Use Surveillance System (GLASS) Report 2022.** World Health Organisation; 2022.
72. **WHO Bacterial Priority Pathogens List, 2024.** World Health Organisation; 2024.

73. Johnson CN, Wilde S, Tuomanen E, Rosch JW: **Convergent impact of vaccination and antibiotic pressures on pneumococcal populations.** *Cell Chem Biol* 2024, **31:**195-206.
74. Abavisani M, Keikha M, Karbalaei M: **First global report about the prevalence of multi-drug resistant Haemophilus influenzae: a systematic review and meta-analysis.** *BMC Infect Dis* 2024, **24:**90.
75. de Andrade AL, Brandileone MC, Di Fabio JL, Oliveira RM, Silva SA, Baiocchi SS, Martelli CM: **Haemophilus influenzae resistance in Latin America: systematic review of surveillance data.** *Microb Drug Resist* 2001, **7:**403-411.
76. Diricks M, Petersen S, Bartels L, Lam TT, Claus H, Bajanca-Lavado MP, Hauswaldt S, Stolze R, Vazquez OJ, Utpatel C, et al: **Revisiting mutational resistance to ampicillin and cefotaxime in Haemophilus influenzae.** *Genome Med* 2024, **16:**140.
77. Su PY, Huang AH, Lai CH, Lin HF, Lin TM, Ho CH: **Extensively drug-resistant Haemophilus influenzae - emergence, epidemiology, risk factors, and regimen.** *BMC Microbiol* 2020, **20:**102.
78. Zhou M, Fu P, Fang C, Shang S, Hua C, Jing C, Xu H, Chen Y, Deng J, Zhang H, et al: **Antimicrobial resistance of Haemophilus influenzae isolates from pediatric hospitals in Mainland China: Report from the ISPED program, 2017-2019.** *Indian J Med Microbiol* 2021, **39:**434-438.
79. Yamada S, Seyama S, Wajima T, Yuzawa Y, Saito M, Tanaka E, Noguchi N: **beta-Lactamase-non-producing ampicillin-resistant Haemophilus influenzae is acquiring multidrug resistance.** *J Infect Public Health* 2020, **13:**497-501.
80. Sun J, Li Y: **Long-term, low-dose macrolide antibiotic treatment in pediatric chronic airway diseases.** *Pediatr Res* 2022, **91:**1036-1042.
81. Cherkaoui A, Gaia N, Baud D, Leo S, Fischer A, Ruppe E, Francois P, Schrenzel J: **Molecular characterization of fluoroquinolones, macrolides, and imipenem resistance in Haemophilus influenzae: analysis of the mutations in QRDRs and assessment of the extent of the AcrAB-TolC-mediated resistance.** *Eur J Clin Microbiol Infect Dis* 2018, **37:**2201-2210.
82. Church JA, Fitzgerald F, Walker AS, Gibb DM, Prendergast AJ: **The expanding role of co-trimoxazole in developing countries.** *Lancet Infect Dis* 2015, **15:**327-339.
83. Pouwels KB, Batra R, Patel A, Edgeworth JD, Robotham JV, Smieszek T: **Will co-trimoxazole resistance rates ever go down? Resistance rates remain high despite decades of reduced co-trimoxazole consumption.** *J Glob Antimicrob Resist* 2017, **11:**71-74.
84. Denizon M, Hong E, Terrade A, Taha MK, Deghmane AE: **A Hunt for the Resistance of Haemophilus influnezae to Beta-Lactams.** *Antibiotics (Basel)* 2024, **13**.
85. Wajima T, Tanaka E, Uchiya KI: **Unique and Ingenious Mechanisms Underlying Antimicrobial Resistance and Spread of Haemophilus influenzae.** *Biol Pharm Bull* 2025, **48:**205-212.
86. Jakubu V, Malisova L, Musilek M, Pomorska K, Zemlickova H: **Characterization of Haemophilus influenzae Strains with Non-Enzymatic Resistance to beta-Lactam Antibiotics Caused by Mutations in the PBP3 Gene in the Czech Republic in 2010-2018.** *Life (Basel)* 2021, **11**.
87. Deghmane AE, Hong E, Chehboub S, Terrade A, Falguieres M, Sort M, Harrison O, Jolley KA, Taha MK: **High diversity of invasive Haemophilus influenzae isolates in France and the emergence of resistance to third generation cephalosporins by alteration of ftsI gene.** *J Infect* 2019, **79:**7-14.

88. Lee S, Kim G, Kim JH, Kim MN, Lee J: **Characterization of Ceftriaxone-Resistant Haemophilus influenzae Among Korean Children.** *J Korean Med Sci* 2024, **39:**e136.
89. Tanaka E, Wajima T, Hirano S, Seyama S, Nakaminami H, Uchiya KI: **Genomic characterization of Haemophilus influenzae harbouring an exogenous resistance gene.** *J Med Microbiol* 2024, **73**.
90. Seyama S, Wajima T, Nakaminami H, Noguchi N: **Amino Acid Substitution in the Major Multidrug Efflux Transporter Protein AcrB Contributes to Low Susceptibility to Azithromycin in Haemophilus influenzae.** *Antimicrob Agents Chemother* 2017, **61**.
91. Cadenas-Jimenez I, Saiz-Escobedo L, Carrera-Salinas A, Camprubi-Marquez X, Calvo-Silveria S, Camps-Massa P, Berbel D, Tubau F, Santos S, Dominguez MA, et al: **Molecular characterization of macrolide resistance in Haemophilus influenzae and Haemophilus parainfluenzae strains (2018-21).** *J Antimicrob Chemother* 2024, **79:**2194-2203.
92. Mohd-Zain Z, Kamsani NH, Ahmad N: **Molecular insights of co-trimoxazole resistance genes in Haemophilus influenzae isolated in Malaysia.** *Trop Biomed* 2013, **30:**584-590.
93. Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R, Post JC, Ehrlich GD: **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* 2007, **8:**R103.
94. Eutsey RA, Hiller NL, Earl JP, Janto BA, Dahlgren ME, Ahmed A, Powell E, Schultz MP, Gilsdorf JR, Zhang L, et al: **Design and validation of a supragenome array for determination of the genomic content of Haemophilus influenzae isolates.** *BMC Genomics* 2013, **14:**484.
95. Power PM, Bentley SD, Parkhill J, Moxon ER, Hood DW: **Investigations into genome diversity of Haemophilus influenzae using whole genome sequencing of clinical isolates and laboratory transformants.** *BMC Microbiol* 2012, **12:**273.
96. Mell JC, Shumilina S, Hall IM, Redfield RJ: **Transformation of natural genetic variation into Haemophilus influenzae genomes.** *PLoS Pathog* 2011, **7:**e1002151.
97. Bobay LM, Traverse CC, Ochman H: **Impermanence of bacterial clones.** *Proc Natl Acad Sci U S A* 2015, **112:**8893-8900.
98. Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV: **Coalescent methods for estimating phylogenetic trees.** *Mol Phylogenet Evol* 2009, **53:**320-328.
99. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR: **Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins.** *Nucleic Acids Res* 2015, **43:**e15.
100. Didelot X, Wilson DJ: **ClonalFrameML: efficient inference of recombination in whole bacterial genomes.** *PLoS Comput Biol* 2015, **11:**e1004041.
101. Michel C, Argudin MA, Wautier M, Echahidi F, Prevost B, Vandenberg O, Martiny D, Hallin M: **Multiple interspecies recombination events documented by whole-genome sequencing in multidrug-resistant Haemophilus influenzae clinical isolates.** *Access Microbiol* 2024, **6**.
102. Tanaka E, Wajima T, Nakaminami H, Uchiya KI: **Alternative quinolone-resistance pathway caused by simultaneous horizontal gene transfer in Haemophilus influenzae.** *J Antimicrob Chemother* 2022, **77:**3270-3274.

103. Vos M, Didelot X: **A comparison of homologous recombination rates in bacteria and archaea.** *ISME J* 2009, **3:**199-208.
104. Torrance EL, Burton C, Diop A, Bobay LM: **Evolution of homologous recombination rates across bacteria.** *Proc Natl Acad Sci U S A* 2024, **121:**e2316302121.
105. Bertran M, D'Aeth JC, Hani E, Amin-Chowdhury Z, Fry NK, Ramsay ME, Litt DJ, Ladhani SN: **Trends in invasive Haemophilus influenzae serotype a disease in England from 2008-09 to 2021-22: a prospective national surveillance study.** *Lancet Infect Dis* 2023.
106. Su PY, Cheng WH, Ho CH: **Molecular characterization of multidrug-resistant non-typeable Haemophilus influenzae with high-level resistance to cefuroxime, levofloxacin, and trimethoprim-sulfamethoxazole.** *BMC Microbiol* 2023, **23:**178.
107. Zhou Y, Wang Y, Cheng J, Zhao X, Liang Y, Wu J: **Molecular epidemiology and antimicrobial resistance of Haemophilus influenzae in Guiyang, Guizhou, China.** *Front Public Health* 2022, **10:**947051.
108. Cox AD, Williams D, Cairns C, St Michael F, Fleming P, Vinogradov E, Arbour M, Masson L, Zou W: **Investigating the candidacy of a capsular polysaccharide-based glycoconjugate as a vaccine to combat Haemophilus influenzae type a disease: A solution for an unmet public health need.** *Vaccine* 2017, **35:**6129-6136.
109. Shoukat A, Van Exan R, Moghadas SM: **Cost-effectiveness of a potential vaccine candidate for Haemophilus influenzae serotype 'a'.** *Vaccine* 2018, **36:**1681-1688.
110. Tsang RSW, Ulanova M: **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* 2017, **35:**4270-4275.
111. Wilkinson TMA, Schembri S, Brightling C, Bakerly ND, Lewis K, MacNee W, Rombo L, Hedner J, Allen M, Walker PP, et al: **Non-typeable Haemophilus influenzae protein vaccine in adults with COPD: A phase 2 clinical trial.** *Vaccine* 2019, **37:**6102-6111.
112. **Haemophilus influenzae: Surveillance standard.** pp. 1-14. Geneva: World Health Organization; 2018:1-14.
113. McInerney JO, McNally A, O'Connell MJ: **Why prokaryotes have pangenomes.** *Nat Microbiol* 2017, **2:**17040.
114. Schurch AC, Arredondo-Alonso S, Willems RJL, Goering RV: **Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches.** *Clin Microbiol Infect* 2018, **24:**350-354.
115. Preska Steinberg A, Lin M, Kussell E: **Core genes can have higher recombination rates than accessory genes within global microbial populations.** *Elife* 2022, **11**.
116. de Been M, Pinholt M, Top J, Bletz S, Mellmann A, van Schaik W, Brouwer E, Rogers M, Kraat Y, Bonten M, et al: **Core Genome Multilocus Sequence Typing Scheme for High- Resolution Typing of Enterococcus faecium.** *J Clin Microbiol* 2015, **53:**3788-3797.
117. Blanc DS, Magalhaes B, Koenig I, Senn L, Grandbastien B: **Comparison of Whole Genome (wg-) and Core Genome (cg-) MLST (BioNumerics(TM)) Versus SNP Variant Calling for Epidemiological Investigation of Pseudomonas aeruginosa.** *Front Microbiol* 2020, **11:**1729.
118. Palma F, Mangone I, Janowicz A, Moura A, Chiaverini A, Torresi M, Garofolo G, Criscuolo A, Brisse S, Di Pasquale A, et al: **In vitro and in silico parameters for precise cgMLST typing of Listeria monocytogenes.** *BMC Genomics* 2022, **23:**235.

119. Bratcher HB, Corton C, Jolley KA, Parkhill J, Maiden MC: **A gene-by-gene population genomics platform: de novo assembly, annotation and genealogical analysis of 108 representative Neisseria meningitidis genomes.** *BMC Genomics* 2014, **15:**1138.
120. Harrison OB, Cehovin A, Skett J, Jolley KA, Massari P, Genco CA, Tang CM, Maiden MCJ: **Neisseria gonorrhoeae Population Genomics: Use of the Gonococcal Core Genome to Improve Surveillance of Antimicrobial Resistance.** *J Infect Dis* 2020, **222:**1816-1825.
121. Cody AJ, Bray JE, Jolley KA, McCarthy ND, Maiden MCJ: **Core Genome Multilocus Sequence Typing Scheme for Stable, Comparative Analyses of Campylobacter jejuni and C. coli Human Disease Isolates.** *J Clin Microbiol* 2017, **55:**2086-2097.
122. Rensburg MJJv, Berger DJ, Fohrmann A, Bray JE, Jolley KA, Maiden MCJ, Brueggemann AB: **Development of the Pneumococcal Genome Library, a core genome multilocus sequence typing scheme, and a taxonomic life identification number barcoding system to investigate and define pneumococcal population structure.** *bioRxiv* 2023**:**2023.2012.2019.571883.
123. Whiley D, Jolley K, Blanchard A, Coffey T, Leigh J: **A core genome multi-locus sequence typing scheme for Streptococcus uberis: an evolution in typing a genetically diverse pathogen.** *Microb Genom* 2024, **10**.
124. Liang KYH, Orata FD, Islam MT, Nasreen T, Alam M, Tarr CL, Boucher YF: **A Vibrio cholerae Core Genome Multilocus Sequence Typing Scheme To Facilitate the Epidemiological Study of Cholera.** *J Bacteriol* 2020, **202**.
125. Gonzalez-Escalona N, Jolley KA, Reed E, Martinez-Urtaza J: **Defining a Core Genome Multilocus Sequence Typing Scheme for the Global Epidemiology of Vibrio parahaemolyticus.** *J Clin Microbiol* 2017, **55:**1682-1697.
126. Abdel-Glil MY, Chiaverini A, Garofolo G, Fasanella A, Parisi A, Harmsen D, Jolley KA, Elschner MC, Tomaso H, Linde J, Galante D: **A Whole-Genome-Based Gene-by-Gene Typing System for Standardized High-Resolution Strain Typing of Bacillus anthracis.** *J Clin Microbiol* 2021, **59:**e0288920.
127. Tourasse NJ, Jolley KA, Kolsto AB, Okstad OA: **Core genome multilocus sequence typing scheme for Bacillus cereus group bacteria.** *Res Microbiol* 2023, **174:**104050.
128. Appelt S, Rohleder AM, Jacob D, von Buttlar H, Georgi E, Mueller K, Wernery U, Kinne J, Joseph M, Jose SV, Scholz HC: **Genetic diversity and spatial distribution of Burkholderia mallei by core genome-based multilocus sequence typing analysis.** *PLoS One* 2022, **17:**e0270499.
129. Moreno-Manjón J, Jolley KA, Maiden MC: **Acinetobacter baumannii core genome multilocus sequence typing.** Universidad Nacional Autónoma de México; University of Oxford; 2022.
130. Abdel-Glil MY, Thomas P, Linde J, Jolley KA, Harmsen D, Wieler LH, Neubauer H, Seyboldt C: **Establishment of a Publicly Available Core Genome Multilocus Sequence Typing Scheme for Clostridium perfringens.** *Microbiol Spectr* 2021, **9:**e0053321.
131. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J: **Roary: rapid large-scale prokaryote pan genome analysis.** *Bioinformatics* 2015, **31:**3691-3693.
132. Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ: **PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria.** *Gigascience* 2019, **8**.

133. Ding W, Baumdicker F, Neher RA: **panX: pan-genome analysis and exploration.** *Nucleic Acids Res* 2018, **46:**e5.
134. Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J: **PGAP: pan-genomes analysis pipeline.** *Bioinformatics* 2012, **28:**416-418.
135. Gautreau G, Bazin A, Gachet M, Planel R, Burlot L, Dubois M, Perrin A, Medigue C, Calteau A, Cruveiller S, et al: **PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph.** *PLoS Comput Biol* 2020, **16:**e1007732.
136. Peng Y, Tang S, Wang D, Zhong H, Jia H, Cai X, Zhang Z, Xiao M, Yang H, Wang J, et al: **MetaPGN: a pipeline for construction and graphical visualization of annotated pangenome networks.** *Gigascience* 2018, **7**.
137. Zhou Z, Charlesworth J, Achtman M: **Accurate reconstruction of bacterial pan- and core genomes with PEPPAN.** *Genome Res* 2020, **30:**1667-1679.
138. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, Gladstone RA, Lo S, Beaudoin C, Floto RA, et al: **Producing polished prokaryotic pangenomes with the Panaroo pipeline.** *Genome Biol* 2020, **21:**180.
139. Boubour A: **Genomic Characterisation of Haemophilus influenzae Capsular Locus.** University of Oxford, Nuffield Department of Population Health; 2021.
140. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, Maiden MCJ: **Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain.** *Microbiology (Reading)* 2012, **158:**1005-1015.
141. Hood DW: **The genome sequence of Haemophilus influenzae.** *Methods Mol Med* 2003, **71:**147-159.
142. Pettigrew MM, Ahearn CP, Gent JF, Kong Y, Gallo MC, Munro JB, D'Mello A, Sethi S, Tettelin H, Murphy TF: **Haemophilus influenzae genome evolution during persistence in the human airways in chronic obstructive pulmonary disease.** *Proc Natl Acad Sci U S A* 2018, **115:**E3256-E3265.
143. May BJ, Zhang Q, Li LL, Paustian ML, Whittam TS, Kapur V: **Complete genomic sequence of Pasteurella multocida, Pm70.** *Proc Natl Acad Sci U S A* 2001, **98:**3460-3465.
144. Smits THM: **The importance of genome sequence quality to microbial comparative genomics.** *BMC Genomics* 2019, **20:**662.
145. Seemann T: **Prokka: rapid prokaryotic genome annotation.** *Bioinformatics* 2014, **30:**2068-2069.
146. Cantalapiedra CP, Hernandez-Plaza A, Letunic I, Bork P, Huerta-Cepas J: **eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale.** *Mol Biol Evol* 2021, **38:**5825-5829.
147. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV: **COG database update: focus on microbial diversity, model organisms, and widespread pathogens.** *Nucleic Acids Res* 2021, **49:**D274-D281.
148. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K: **KEGG: new perspectives on genomes, pathways, diseases and drugs.** *Nucleic Acids Res* 2017, **45:**D353-D361.
149. Bruen TC, Philippe H, Bryant D: **A simple and robust statistical test for detecting the presence of recombination.** *Genetics* 2006, **172:**2665-2681.
150. Lai YP, Ioerger TR: **A statistical method to identify recombination in bacterial genomes based on SNP incompatibility.** *BMC Bioinformatics* 2018, **19:**450.

151. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics* 2014, **30:**1312-1313.
152. Watts SC, Holt KE: **hicap: In Silico Serotyping of the Haemophilus influenzae Capsule Locus.** *J Clin Microbiol* 2019, **57**.
153. Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, Carrico JA, Achtman M: **GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens.** *Genome Res* 2018, **28:**1395-1404.
154. Atack JM, Murphy TF, Bakaletz LO, Seib KL, Jennings MP: **Closed Complete Genome Sequences of Two Nontypeable Haemophilus influenzae Strains Containing Novel modA Alleles from the Sputum of Patients with Chronic Obstructive Pulmonary Disease.** *Microbiol Resour Announc* 2018, **7**.
155. Harrison A, Dyer DW, Gillaspy A, Ray WC, Mungur R, Carson MB, Zhong H, Gipson J, Gipson M, Johnson LS, 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:**4627-4636.
156. Krisna MA, Jolley KA, Monteith W, Boubour A, Hamers RL, Brueggemann AB, Harrison OB, Maiden MCJ: **Development and implementation of a core genome multilocus sequence typing scheme for Haemophilus influenzae.** *Microb Genom* 2024, **10**.
157. Loman NJ, Pallen MJ: **Twenty years of bacterial genome sequencing.** *Nat Rev Microbiol* 2015, **13:**787-794.
158. Kc R, Leong KWC, Harkness NM, Lachowicz J, Gautam SS, Cooley LA, McEwan B, Petrovski S, Karupiah G, O'Toole RF: **Whole-genome analyses reveal gene content differences between nontypeable Haemophilus influenzae isolates from chronic obstructive pulmonary disease compared to other clinical phenotypes.** *Microb Genom* 2020, **6**.
159. Meyler K, Meehan M, Bennett D, Mulhall R, Harrison O, Gavin P, Drew RJ, Cunney R: **Spontaneous capsule loss in Haemophilus influenzae serotype b associated with Hib conjugate vaccine failure and invasive disease.** *Clin Microbiol Infect* 2019, **25:**390-391.
160. Kilian M: **A taxonomic study of the genus Haemophilus, with the proposal of a new species.** *J Gen Microbiol* 1976, **93:**9-62.
161. Vinatzer BA, Tian L, Heath LS: **A proposal for a portal to make earth's microbial diversity easily accessible and searchable.** *Antonie Van Leeuwenhoek* 2017, **110:**1271-1279.
162. Contreras A, Posada R: **Haemophilus influenzae Infections.** *Pediatr Rev* 2023, **44:**422-424.
163. Oliver SE, Rubis AB, Soeters HM, Reingold A, Barnes M, Petit S, Farley MM, Harrison LH, Como-Sabetti K, Khanlian SA, et al: **Epidemiology of Invasive Nontypeable Haemophilus influenzae Disease-United States, 2008-2019.** *Clin Infect Dis* 2023, **76:**1889-1895.
164. Ahearn CP, Kirkham C, Chaves LD, Kong Y, Pettigrew MM, Murphy TF: **Discovery and Contribution of Nontypeable Haemophilus influenzae NTHI1441 to Human Respiratory Epithelial Cell Invasion.** *Infect Immun* 2019, **87**.
165. Clementi CF, Murphy TF: **Non-typeable Haemophilus influenzae invasion and persistence in the human respiratory tract.** *Front Cell Infect Microbiol* 2011, **1:**1.
166. Naghavi M, Vollset SE, Ikuta KS, Swetschinski LR, Gray AP, Wool EE, Robles Aguilar G, Mestrovic T, Smith G, Han C, et al: **Global burden of bacterial antimicrobial**

resistance 1990–2021: a systematic analysis with forecasts to 2050. *The Lancet* 2024, **404:**1199-1226.
167. Wan TW, Huang YT, Lai JH, Chao QT, Yeo HH, Lee TF, Chang YC, Chiu HC: **The emergence of transposon-driven multidrug resistance in invasive nontypeable Haemophilus influenzae over the last decade.** *Int J Antimicrob Agents* 2024**:**107319.
168. Control ECfDPa: **Haemophilus influenzae disease.** In *Annual Epidemiological Report for 2021*. pp. 8. Stockholm: European Centre for Disease Prevention and Control; 2024:8.
169. Prevention UCfDCa: **Active Bacterial Core Surveillance Report, Emerging Infections Program Network, Haemophilus influenzae.** Atlanta; 2021.
170. Tsang RS, Shuel M, Wylie J, Lefebvre B, Hoang L, Law DK: **Population genetics of Haemophilus influenzae serotype a in three Canadian provinces.** *Can J Microbiol* 2013, **59:**362-364.
171. Cardoso B, Fontana H, Esposito F, Cerdeira L, Santos SR, Yoshioka CRM, da Silveira IR, Cassettari V, Lincopan N: **Genomic insights of international clones of Haemophilus influenzae causing invasive infections in vaccinated and unvaccinated infants.** *Microb Pathog* 2021, **150:**104644.
172. Murtagh F, Contreras P: **Algorithms for hierarchical clustering: an overview, II.** *WIREs Data Mining and Knowledge Discovery* 2017, **7:**e1219.
173. Zhou Z, Charlesworth J, Achtman M: **HierCC: a multi-level clustering scheme for population assignments based on core genome MLST.** *Bioinformatics* 2021, **37:**3645-3646.
174. Hennart M, Guglielmini J, Bridel S, Maiden MCJ, Jolley KA, Criscuolo A, Brisse S: **A Dual Barcoding Approach to Bacterial Strain Nomenclature: Genomic Taxonomy of Klebsiella pneumoniae Strains.** *Mol Biol Evol* 2022, **39**.
175. Marakeby H, Badr E, Torkey H, Song Y, Leman S, Monteil CL, Heath LS, Vinatzer BA: **A system to automatically classify and name any individual genome-sequenced organism independently of current biological classification and nomenclature.** *PLoS One* 2014, **9:**e89142.
176. Tian L, Huang C, Mazloom R, Heath LS, Vinatzer BA: **LINbase: a web server for genome-based identification of prokaryotes as members of crowdsourced taxa.** *Nucleic Acids Res* 2020, **48:**W529-W537.
177. Weisberg AJ, Elmarakeby HA, Heath LS, Vinatzer BA: **Similarity-based codes sequentially assigned to ebolavirus genomes are informative of species membership, associated outbreaks, and transmission chains.** *Open Forum Infect Dis* 2015, **2:**ofv024.
178. Jansen van Rensburg MJ, Berger DJ, Fohrmann A, Bray JE, Jolley KA, Maiden MCJ, Brueggemann AB: **Development of the Pneumococcal Genome Library, a core genome multilocus sequence typing scheme, and a taxonomic life identification number barcoding system to investigate and define pneumococcal population structure.** *bioRxiv* 2023**:**2023.2012.2019.571883.
179. Palma F, Hennart M, Jolley KA, Crestani C, Wyres KL, Bridel S, Yeats CA, Brancotte B, Raffestin B, David S, et al: **Bacterial strain nomenclature in the genomic era: Life Identification Numbers using a gene-by-gene approach.** *bioRxiv* 2024**:**2024.2003.2011.584534.
180. Nolen LD, DeByle C, Topaz N, Simons BC, Tiffany A, Reasonover A, Castrodale L, McLaughlin J, Klejka J, Wang X, Bruce M: **Genomic Diversity of Haemophilus**

influenzae Serotype a in an Outbreak Community-Alaska, 2018. *J Infect Dis* 2022, **225:**520-524.
181. Tsang RSW, Shuel M, Ahmad T, Hayden K, Knox N, Van Domselaar G, Hoang L, Tyrrell GJ, Minion J, Van Caeseele P, et al: **Whole genome sequencing to study the phylogenetic structure of serotype a Haemophilus influenzae recovered from patients in Canada.** *Can J Microbiol* 2020, **66:**99-110.
182. Rousseeuw PJ: **Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.** *Journal of Computational and Applied Mathematics* 1987, **20:**53-65.
183. Shutaywi M, Kachouie NN: **Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering.** *Entropy* 2021, **23:**759.
184. Bouchez V, Guglielmini J, Dazas M, Landier A, Toubiana J, Guillot S, Criscuolo A, Brisse S: **Genomic Sequencing of Bordetella pertussis for Epidemiology and Global Surveillance of Whooping Cough.** *Emerg Infect Dis* 2018, **24:**988-994.
185. Severiano A, Pinto FR, Ramirez M, Carrico JA: **Adjusted Wallace coefficient as a measure of congruence between typing methods.** *J Clin Microbiol* 2011, **49:**3997-4000.
186. Koopman R, Wang S: **Mutual information based labelling and comparing clusters.** *Scientometrics* 2017, **111:**1157-1167.
187. Zhou Z, Alikhan NF, Mohamed K, Fan Y, Agama Study G, Achtman M: **The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity.** *Genome Res* 2020, **30:**138-152.
188. Breurec S, Criscuolo A, Diancourt L, Rendueles O, Vandenbogaert M, Passet V, Caro V, Rocha EP, Touchon M, Brisse S: **Genomic epidemiology and global diversity of the emerging bacterial pathogen Elizabethkingia anophelis.** *Sci Rep* 2016, **6:**30379.
189. Letunic I, Bork P: **Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool.** *Nucleic Acids Res* 2024, **52:**W78-W82.
190. Bray JE, Correia A, Varga M, Jolley KA, Maiden MCJ, Rodrigues CMC: **Ribosomal MLST nucleotide identity (rMLST-NI), a rapid bacterial species identification method: application to Klebsiella and Raoultella genomic species validation.** *Microb Genom* 2022, **8**.
191. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD, Croucher NJ: **Fast and flexible bacterial genomic epidemiology with PopPUNK.** *Genome Res* 2019, **29:**304-316.
192. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S: **High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.** *Nat Commun* 2018, **9:**5114.
193. Musser JM, Kroll JS, Granoff DM, Moxon ER, Brodeur BR, Campos J, Dabernat H, Frederiksen W, Hamel J, Hammond G, et al.: **Global genetic structure and molecular epidemiology of encapsulated Haemophilus influenzae.** *Rev Infect Dis* 1990, **12:**75-111.
194. Shuel M, Knox N, Tsang RSW: **Global population structure of Haemophilus influenzae serotype a (Hia) and emergence of invasive Hia disease: capsule switching or capsule replacement?** *Can J Microbiol* 2021, **67:**875-884.
195. Fong W, Martinez E, Timms V, Ginn A, Nguyen T, Rahman H, Sintchenko V: **Increase in invasive Haemophilus influenzae serotype A infections during the COVID-19 pandemic in New South Wales, Australia.** *Pathology* 2024, **56:**696-701.

196. Terrat Y, Farnaes L, Bradley J, Tromas N, Shapiro BJ: **Two cases of type-a Haemophilus influenzae meningitis within the same week in the same hospital are phylogenetically unrelated but recently exchanged capsule genes.** *Microb Genom* 2020, **6**.
197. Boisvert AA, Moore D: **Invasive disease due to Haemophilus influenzae type A in children in Canada's north: A priority for prevention.** *Can J Infect Dis Med Microbiol* 2015, **26:**291-292.
198. Quebec MoHaSS: **Haemophilus influenzae au Nunavik.** (Services MoHaS ed., vol. 8. pp. 1-3. Quebec: Ministry of Health and Social Services Quebec; 2013:1-3.
199. Connor TR, Corander J, Hanage WP: **Population subdivision and the detection of recombination in non-typable Haemophilus influenzae.** *Microbiology* 2012, **158:**2958-2964.
200. **Haemophilus influenzae Infection** [https://www.statpearls.com/ArticleLibrary/viewarticle/91515]
201. Murphy TF, Faden H, Bakaletz LO, Kyd JM, Forsgren A, Campos J, Virji M, Pelton SI: **Nontypeable Haemophilus influenzae as a pathogen in children.** *Pediatr Infect Dis J* 2009, **28:**43-48.
202. Wunrow HY, Bender RG, Vongpradith A, Sirota SB, Swetschinski LR, Novotney A, Gray AP, Ikuta KS, Sharara F, Wool EE, et al: **Global, regional, and national burden of meningitis and its aetiologies, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019.** *The Lancet Neurology* 2023, **22:**685-711.
203. Slack M, Esposito S, Haas H, Mihalyi A, Nissen M, Mukherjee P, Harrington L: **Haemophilus influenzae type b disease in the era of conjugate vaccines: critical factors for successful eradication.** *Expert Rev Vaccines* 2020, **19:**903-917.
204. Hasegawa Y, Arinuma Y, Tanaka S, Tono T, Tanaka T, Muramatsu T, Kondo J, Matsueda Y, Hoshiyama T, Wada T, et al: **Haemophilus influenzae Non-type b Infection in an Adult Patient with Systemic Lupus Erythematosus.** *Intern Med* 2020, **59:**3097-3101.
205. Langereis JD, de Jonge MI: **Invasive Disease Caused by Nontypeable Haemophilus influenzae.** *Emerg Infect Dis* 2015, **21:**1711-1718.
206. Ladhani S, Slack MP, Heath PT, von Gottberg A, Chandra M, Ramsay ME, European Union Invasive Bacterial Infection Surveillance p: **Invasive Haemophilus influenzae Disease, Europe, 1996-2006.** *Emerg Infect Dis* 2010, **16:**455-463.
207. Whittaker R, Economopoulou A, Dias JG, Bancroft E, Ramliden M, Celentano LP, European Centre for Disease P, Control Country Experts for Invasive Haemophilus influenzae D: **Epidemiology of Invasive Haemophilus influenzae Disease, Europe, 2007-2014.** *Emerg Infect Dis* 2017, **23:**396-404.
208. NICD: **GERMS South Africa: Annual Surveillance Review 2018.** (Lebaka T, Quan V eds.). pp. 50: National Institute for Communicable Diseases; 2018:50.
209. Gallo MC, Kirkham C, Eng S, Bebawee RS, Kong Y, Pettigrew MM, Tettelin H, Murphy TF: **Changes in IgA Protease Expression Are Conferred by Changes in Genomes during Persistent Infection by Nontypeable Haemophilus influenzae in Chronic Obstructive Pulmonary Disease.** *Infect Immun* 2018, **86**.
210. Murphy TF, Kirkham C, D'Mello A, Sethi S, Pettigrew MM, Tettelin H: **Adaptation of Nontypeable Haemophilus influenzae in Human Airways in COPD: Genome Rearrangements and Modulation of Expression of HMW1 and HMW2.** *mBio* 2023, **14:**e0014023.

211. Osman KL, Jefferies JM, Woelk CH, Cleary DW, Clarke SC: **The adhesins of non-typeable Haemophilus influenzae.** *Expert Rev Anti Infect Ther* 2018, **16:**187-196.
212. Fluit AC, Bayjanov JR, Benaissa-Trouw BJ, Rogers MRC, Diez-Aguilar M, Canton R, Tunney MM, Elborn JS, Ekkelenkamp MB: **Whole-genome analysis of Haemophilus influenzae strains isolated from persons with cystic fibrosis.** *J Med Microbiol* 2022, **71**.
213. Zhang L, Xie J, Patel M, Bakhtyar A, Ehrlich GD, Ahmed A, Earl J, Marrs CF, Clemans D, Murphy TF, Gilsdorf JR: **Nontypeable Haemophilus influenzae genetic islands associated with chronic pulmonary infection.** *PLoS One* 2012, **7:**e44730.
214. Ribet D, Cossart P: **How bacterial pathogens colonize their hosts and invade deeper tissues.** *Microbes Infect* 2015, **17:**173-183.
215. Fernandez-Calvet A, Euba B, Gil-Campillo C, Catalan-Moreno A, Moleres J, Marti S, Merlos A, Langereis JD, Garcia-Del Portillo F, Bakaletz LO, et al: **Phase Variation in HMW1A Controls a Phenotypic Switch in Haemophilus influenzae Associated with Pathoadaptation during Persistent Infection.** *mBio* 2021, **12:**e0078921.
216. Maughan H, Redfield RJ: **Tracing the evolution of competence in Haemophilus influenzae.** *PLoS One* 2009, **4:**e5854.
217. Xu Q, Almudervar A, Casey J, Pichichero M: **Nasopharyngeal Bacterial Interactions in Children.** *Emerging Infectious Disease journal* 2012, **18:**1738.
218. Flynn M, Lyall Z, Shepherd G, Lee ONY, Marianna Da Fonseca I, Dong Y, Chalmers S, Hare J, Thomson J, Millar F: **Interactions of the bacteriome, virome, and immune system in the nose.** *FEMS Microbes* 2022, **3:**xtac020.
219. Da Costa RM, Rooke JL, Wells TJ, Cunningham AF, Henderson IR: **Type 5 secretion system antigens as vaccines against Gram-negative bacterial infections.** *NPJ Vaccines* 2024, **9:**159.
220. Delany I, Rappuoli R, Seib KL: **Vaccines, reverse vaccinology, and bacterial pathogenesis.** *Cold Spring Harb Perspect Med* 2013, **3:**a012476.
221. Phillips ZN, Jennison AV, Whitby PW, Stull TL, Staples M, Atack JM: **Examination of phase-variable haemoglobin-haptoglobin binding proteins in non-typeable Haemophilus influenzae reveals a diverse distribution of multiple variants.** *FEMS Microbiol Lett* 2022, **369**.
222. Galgani I, Poder A, Jogi R, Anttila VJ, Milleri S, Borobia AM, Launay O, Testa M, Casula D, Grassano L, et al: **Immunogenicity and safety of the non-typable Haemophilus influenzae-Moraxella catarrhalis (NTHi-Mcat) vaccine administered following the recombinant zoster vaccine versus administration alone: Results from a randomized, phase 2a, non-inferiority trial.** *Hum Vaccin Immunother* 2023, **19:**2187194.
223. Van Damme P, Leroux-Roels G, Vandermeulen C, De Ryck I, Tasciotti A, Dozot M, Moraschini L, Testa M, Arora AK: **Safety and immunogenicity of non-typeable Haemophilus influenzae-Moraxella catarrhalis vaccine.** *Vaccine* 2019, **37:**3113-3122.
224. Power RA, Parkhill J, de Oliveira T: **Microbial genome-wide association studies: lessons from human GWAS.** *Nat Rev Genet* 2017, **18:**41-50.
225. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J: **pyseer: a comprehensive tool for microbial pangenome-wide association studies.** *Bioinformatics* 2018, **34:**4310-4312.

226. Lee JH, Cho HK, Kim KH, Kim CH, Kim DS, Kim KN, Cha SH, Oh SH, Hur JK, Kang JH, et al: **Etiology of invasive bacterial infections in immunocompetent children in Korea (1996-2005): a retrospective multicenter study.** *J Korean Med Sci* 2011, **26:**174-183.
227. Nemoto K, Yatera K, Akata K, Ikegami H, Yamasaki K, Hata R, Naito K, Noguchi S, Kawanami T, Fukuda K, Mukae H: **Comparative study of bacterial flora in bronchoalveolar lavage fluid of pneumonia patients based on their pneumonia subtypes and comorbidities using 16S ribosomal RNA gene analysis.** *J Infect Chemother* 2022, **28:**1402-1409.
228. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A: **Using SPAdes De Novo Assembler.** *Curr Protoc Bioinformatics* 2020, **70:**e102.
229. Gurevich A, Saveliev V, Vyahhi N, Tesler G: **QUAST: quality assessment tool for genome assemblies.** *Bioinformatics* 2013, **29:**1072-1075.
230. Derelle R, von Wachsmann J, Maklin T, Hellewell J, Russell T, Lalvani A, Chindelevitch L, Croucher NJ, Harris SR, Lees JA: **Seamless, rapid, and accurate analyses of outbreak genomic data using split k-mer analysis.** *Genome Res* 2024, **34:**1661-1673.
231. Jaillard M, Lima L, Tournoud M, Mahe P, van Belkum A, Lacroix V, Jacob L: **A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events.** *PLoS Genet* 2018, **14:**e1007758.
232. Moxon ER, Kroll JS: **Type b capsular polysaccharide as a virulence factor of Haemophilus influenzae.** *Vaccine* 1988, **6:**113-115.
233. Dobson SR, Kroll JS, Moxon ER: **Insertion sequence IS1016 and absence of Haemophilus capsulation genes in the Brazilian purpuric fever clone of Haemophilus influenzae biogroup aegyptius.** *Infect Immun* 1992, **60:**618-622.
234. Strouts F, Power P, Croucher N, Corton N, van Tonder A, Quail M, Langford P, Hudson M, Parkhill J, Kroll JS, Bentley S: **Lineage-specific Virulence Determinants of Haemophilus influenzae Biogroup aegyptius.** *Emerging Infectious Disease journal* 2012, **18:**449.
235. Zhang C, Dong SS, Xu JY, He WM, Yang TL: **PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files.** *Bioinformatics* 2019, **35:**1786-1788.
236. Earle SG, Lobanovska M, Lavender H, Tang C, Exley RM, Ramos-Sevillano E, Browning DF, Kostiou V, Harrison OB, Bratcher HB, et al: **Genome-wide association studies reveal the role of polymorphisms affecting factor H binding protein expression in host invasion by Neisseria meningitidis.** *PLoS Pathog* 2021, **17:**e1009992.
237. Hosmer DW, Jr. Lemeshow, S. Sturdivant, R.X.: **Model-Building Strategies and Methods for Logistic Regression.** In *Applied Logistic Regression.* 2013: 89-151
238. **Beware Default Random Forest Importances** [https://explained.ai/rf-importance/index.html]
239. Leitner W, Turner WR: **Measurement and Analysis of Biodiversity☆.** In *Reference Module in Life Sciences.* Elsevier; 2017
240. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, et al: **Accurate structure prediction of biomolecular interactions with AlphaFold 3.** *Nature* 2024, **630:**493-500.
241. Izydorczyk C, Waddell BJ, Weyant RB, Surette MG, Somayaji R, Rabin HR, Conly JM, Church DL, Parkins MD: **The natural history and genetic diversity of Haemophilus**

influenzae infecting the airways of adults with cystic fibrosis. *Sci Rep* 2022, **12:**15765.
242. McFadden D: **Conditional Logit Analysis of Qualitative Choice Behavior.** In *Economic Theory and Mathematical Economics.* Edited by Zarembka P. New York: Academic Press; 1974: 105-142
243. Wang J, Chitsaz F, Derbyshire MK, Gonzales NR, Gwadz M, Lu S, Marchler GH, Song JS, Thanki N, Yamashita RA, et al: **The conserved domain database in 2023.** *Nucleic Acids Res* 2023, **51:**D384-D388.
244. Innamorati KA, Earl JP, Aggarwal SD, Ehrlich GD, Hiller NL: **The Bacterial Guide to Designing a Diversified Gene Portfolio.** In *The Pangenome: Diversity, Dynamics and Evolution of Genomes.* Edited by Tettelin H, Medini D. Cham (CH); 2020: 51-87
245. Tantoso E, Eisenhaber B, Kirsch M, Shitov V, Zhao Z, Eisenhaber F: **To kill or to be killed: pangenome analysis of Escherichia coli strains reveals a tailocin specific for pandemic ST131.** *BMC Biol* 2022, **20:**146.
246. Sakoparnig T, Field C, van Nimwegen E: **Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species.** *Elife* 2021, **10**.
247. den Bakker HC, Didelot X, Fortes ED, Nightingale KK, Wiedmann M: **Lineage specific recombination rates and microevolution in Listeria monocytogenes.** *BMC Evol Biol* 2008, **8:**277.
248. Didelot X, Urwin R, Maiden MCJ, Falush D: **Genealogical typing of Neisseria meningitidis.** *Microbiology (Reading)* 2009, **155:**3176-3186.
249. Didelot X, Maiden MCJ: **Impact of recombination on bacterial evolution.** *Trends in Microbiology* 2010, **18:**315-322.
250. Chen J, Du W, Li Y, Zhou H, Ouyang D, Yao Z, Fu J, Ye X: **Genome-based model for differentiating between infection and carriage Staphylococcus aureus.** *Microbiol Spectr* 2024, **12:**e0049324.
251. Eriksson L, Johannesen TB, Stenmark B, Jacobsson S, Sall O, Hedberg ST, Fredlund H, Stegger M, Molling P: **Genetic variants linked to the phenotypic outcome of invasive disease and carriage of Neisseria meningitidis.** *Microb Genom* 2023, **9**.
252. Meinshausen N, Bühlmann P: **Stability Selection.** *Journal of the Royal Statistical Society Series B: Statistical Methodology* 2010, **72:**417-473.
253. Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, Spencer CCA, Iqbal Z, Clifton DA, Hopkins KL, et al: **Identifying lineage effects when controlling for population structure improves power in bacterial association studies.** *Nat Microbiol* 2016, **1:**16041.
254. Coll F, Gouliouris T, Bruchmann S, Phelan J, Raven KE, Clark TG, Parkhill J, Peacock SJ: **PowerBacGWAS: a computational pipeline to perform power calculations for bacterial genome-wide association studies.** *Commun Biol* 2022, **5:**266.
255. MacAlasdair N, Pöntinen AK, Ling C, Mallawaarachchi S, Thaipadungpanit J, Nosten FH, Turner C, Bentley SD, Croucher NJ, Turner P, Corander J: **The major pathogen Haemophilus influenzae experiences pervasive recombination and purifying selection at local and global scales.** *bioRxiv* 2024**:**2024.2010.2016.618562.
256. Mallawaarachchi S, Tonkin-Hill G, Pontinen AK, Calland JK, Gladstone RA, Arredondo-Alonso S, MacAlasdair N, Thorpe HA, Top J, Sheppard SK, et al: **Detecting co-selection through excess linkage disequilibrium in bacterial genomes.** *NAR Genom Bioinform* 2024, **6:**lqae061.

257. Short B, Carson S, Devlin AC, Reihill JA, Crilly A, MacKay W, Ramage G, Williams C, Lundy FT, McGarvey LP, et al: **Non-typeable Haemophilus influenzae chronic colonization in chronic obstructive pulmonary disease (COPD).** *Crit Rev Microbiol* 2021, **47:**192-205.
258. Chen Y, Jiang Q, Peierdun M, Takiff HE, Gao Q: **The mutational signatures of poor treatment outcomes on the drug-susceptible Mycobacterium tuberculosis genome.** *Elife* 2023, **12**.
259. Tuan VP, Yahara K, Dung HDQ, Binh TT, Huu Tung P, Tri TD, Thuan NPM, Khien VV, Trang TTH, Phuc BH, et al: **Genome-wide association study of gastric cancer- and duodenal ulcer-derived Helicobacter pylori strains reveals discriminatory genetic variations and novel oncoprotein candidates.** *Microb Genom* 2021, **7**.
260. Saxena D, Maitra R, Bormon R, Czekanska M, Meiers J, Titz A, Verma S, Chopra S: **Tackling the outer membrane: facilitating compound entry into Gram-negative bacterial pathogens.** *NPJ Antimicrob Resist* 2023, **1:**17.
261. Moleres J, Fernandez-Calvet A, Ehrlich RL, Marti S, Perez-Regidor L, Euba B, Rodriguez-Arce I, Balashov S, Cuevas E, Linares J, et al: **Antagonistic Pleiotropy in the Bifunctional Surface Protein FadL (OmpP1) during Adaptation of Haemophilus influenzae to Chronic Lung Infection Associated with Chronic Obstructive Pulmonary Disease.** *mBio* 2018, **9**.
262. Germany EM, Thewasano N, Imai K, Maruno Y, Bamert RS, Stubenrauch CJ, Dunstan RA, Ding Y, Nakajima Y, Lai X, et al: **Dual recognition of multiple signals in bacterial outer membrane proteins enhances assembly and maintains membrane integrity.** *Elife* 2024, **12**.
263. Niramitranon J, Sansom MS, Pongprayoon P: **Why do the outer membrane proteins OmpF from E. coli and OprP from P. aeruginosa prefer trimers? Simulation studies.** *J Mol Graph Model* 2016, **65:**1-7.
264. Zambolin S, Clantin B, Chami M, Hoos S, Haouz A, Villeret V, Delepelaire P: **Structural basis for haem piracy from host haemopexin by Haemophilus influenzae.** *Nat Commun* 2016, **7:**11590.
265. Whitby PW, Seale TW, VanWagoner TM, Morton DJ, Stull TL: **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.
266. Zhou K, Aertsen A, Michiels CW: **The role of variable DNA tandem repeats in bacterial adaptation.** *FEMS Microbiology Reviews* 2014, **38:**119-141.
267. Noinaj N, Guillier M, Barnard TJ, Buchanan SK: **TonB-dependent transporters: regulation, structure, and function.** *Annu Rev Microbiol* 2010, **64:**43-60.
268. Wang J, Xiong K, Pan Q, He W, Cong Y: **Application of TonB-Dependent Transporters in Vaccine Development of Gram-Negative Bacteria.** *Front Cell Infect Microbiol* 2020, **10:**589115.
269. Acevedo R, Fernandez S, Zayas C, Acosta A, Sarmiento ME, Ferro VA, Rosenqvist E, Campa C, Cardoso D, Garcia L, Perez JL: **Bacterial outer membrane vesicles and vaccine applications.** *Front Immunol* 2014, **5:**121.
270. Pfluger-Grau K, Gorke B: **Regulatory roles of the bacterial nitrogen-related phosphotransferase system.** *Trends Microbiol* 2010, **18:**205-214.
271. Schulte JE, Goulian M: **The Phosphohistidine Phosphatase SixA Targets a Phosphotransferase System.** *mBio* 2018, **9**.

272. Higgs PG, Ran W: **Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage.** *Mol Biol Evol* 2008, **25:**2279-2291.
273. Ayan GB, Park HJ, Gallie J: **The birth of a bacterial tRNA gene by large-scale, tandem duplication events.** *Elife* 2020, **9**.
274. Tremblay-Savard O, Benzaid B, Lang BF, El-Mabrouk N: **Evolution of tRNA Repertoires in Bacillus Inferred with OrthoAlign.** *Mol Biol Evol* 2015, **32:**1643-1656.
275. Park J, Lee D, Yi H, Yun CW, Kim HS: **Bacterial persistence to antibiotics activated by tRNA mutations.** *J Antimicrob Chemother* 2024, **79:**2923-2931.
276. Ehrlich R, Davyt M, Lopez I, Chalar C, Marin M: **On the Track of the Missing tRNA Genes: A Source of Non-Canonical Functions?** *Front Mol Biosci* 2021, **8:**643701.
277. Sigmund CD, Ettayebi M, Morgan EA: **Antibiotic resistance mutations in 16S and 23S ribosomal RNA genes of Escherichia coli.** *Nucleic Acids Res* 1984, **12:**4653-4663.
278. Miyazaki K, Kitahara K: **Functional metagenomic approach to identify overlooked antibiotic resistance mutations in bacterial rRNA.** *Sci Rep* 2018, **8:**5179.
279. Harrison OB, Clemence M, Dillard JP, Tang CM, Trees D, Grad YH, Maiden MC: **Genomic analyses of Neisseria gonorrhoeae reveal an association of the gonococcal genetic island with antimicrobial resistance.** *J Infect* 2016, **73:**578-587.
280. Ke W, Li D, Tso LS, Wei R, Lan Y, Chen Z, Zhang X, Wang L, Liang C, Liao Y, et al: **Macrolide and fluoroquinolone associated mutations in Mycoplasma genitalium in a retrospective study of male and female patients seeking care at a STI Clinic in Guangzhou, China, 2016-2018.** *BMC Infect Dis* 2020, **20:**950.
281. Waldner M, Kinnear A, Yacoub E, McAllister T, Register K, Li C, Jelinski M: **Genome-Wide Association Study of Nucleotide Variants Associated with Resistance to Nine Antimicrobials in Mycoplasma bovis.** *Microorganisms* 2022, **10**.
282. Prestinaci F, Pezzotti P, Pantosti A: **Antimicrobial resistance: a global multifaceted phenomenon.** *Pathog Glob Health* 2015, **109:**309-318.
283. **Haemophilus influenzae Infection** [https://www.statpearls.com/ArticleLibrary/viewarticle/91515]
284. **Menkes Luncurkan Vaksin Pentavalen dan Program Imunisasi Lanjutan bagi Batita** [https://kemkes.go.id/id/%20menkes-luncurkan-vaksin-pentavalen-dan-progran-imunisasi-lanjutan-bagi-batita]
285. Rusmil K, Gunardi H, Fadlyana E, Soedjatmiko, Dhamayanti M, Sekartini R, Satari HI, Risan NA, Prasetio D, Tarigan R, et al: **The immunogenicity, safety, and consistency of an Indonesia combined DTP-HB-Hib vaccine in expanded program on immunization schedule.** *BMC Pediatr* 2015, **15:**219.
286. **WHO Antibiotic Categorization: AWaRe** [https://aware.essentialmeds.org/groups]
287. Chang CM, Lauderdale TL, Lee HC, Lee NY, Wu CJ, Chen PL, Lee CC, Chen PC, Ko WC: **Colonisation of fluoroquinolone-resistant Haemophilus influenzae among nursing home residents in southern Taiwan.** *J Hosp Infect* 2010, **75:**304-308.
288. Safari D, Lestari AN, Khoeri MM, Tafroji W, Giri-Rachman EA, Harimurti K, Kurniati N: **Nasopharyngeal carriage and antimicrobial susceptibility profile of Haemophilus influenzae among patients infected with HIV in Jakarta, Indonesia.** *Access Microbiol* 2020, **2:**acmi000165.
289. Gessner BD, Sutanto A, Steinhoff M, Soewignjo S, Widjaya A, Nelson C, Arjoso S: **A population-based survey of Haemophilus influenzae type b nasopharyngeal carriage prevalence in Lombok Island, Indonesia.** *Pediatr Infect Dis J* 1998, **17:**S179-182.

290. Lees JA, Ferwerda B, Kremer PHC, Wheeler NE, Seron MV, Croucher NJ, Gladstone RA, Bootsma HJ, Rots NY, Wijmega-Monsuur AJ, et al: **Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis.** *Nat Commun* 2019, **10:**2176.
291. Salsabila K, Paramaiswari WT, Amalia H, Ruyani A, Tafroji W, Winarti Y, Khoeri MM, Safari D: **Nasopharyngeal carriage rate, serotype distribution, and antimicrobial susceptibility profile of Streptococcus pneumoniae isolated from children under five years old in Kotabaru, South Kalimantan, Indonesia.** *J Microbiol Immunol Infect* 2022, **55:**482-488.
292. Prayitno A, Supriyatno B, Munasir Z, Karuniawati A, Hadinegoro SRS, Prihartono J, Safari D, Sundoro J, Khoeri MM: **Pneumococcal nasopharyngeal carriage in Indonesia infants and toddlers post-PCV13 vaccination in a 2+1 schedule: A prospective cohort study.** *PLoS One* 2021, **16:**e0245789.
293. Putri ND, Salsabila K, Prayitno A, Aprianti SC, Paramaiswari WT, Krisna MA, Safari D: **Epidemiology of Haemophilus influenzae in children on Lombok Island, Indonesia.** *Access Microbiol* 2023, **5**.
294. **Summary of Risk-based Pneumococcal Vaccination Recommendations** [https://www.cdc.gov/pneumococcal/hcp/vaccine-recommendations/risk-indications.html]
295. **Sekilas Vaksin Pneumokokus** [https://www.idai.or.id/artikel/klinik/imunisasi/sekilas-vaksin-pneumokokus]
296. Paramaiswari WT, Muktiarti D, Safari D, Amalia R, Padma M, Winarti Y, Khoeri MM, Daningrat WOD, Tafroji W, Soebandrio A: **Nasopharyngeal carriage and serotype distribution of Streptococcus pneumoniae among HIV-infected children aged >6 years: before and after vaccination of 13-valent pneumococcal conjugate vaccine.** *Clin Exp Vaccine Res* 2025, **14:**127-137.
297. Krisna MA, Alimsardjono L, Salsabila K, Vermasari N, Daningrat WOD, Kuntaman K, Harrison OB, Maiden MCJ, Safari D: **Whole-genome sequencing of non-typeable Haemophilus influenzae isolated from a tertiary care hospital in Surabaya, Indonesia.** *BMC Infect Dis* 2024, **24:**1097.
298. Wang X, Mair R, Hatcher C, Theodore MJ, Edmond K, Wu HM, Harcourt BH, Carvalho Mda G, Pimenta F, Nymadawa P, et al: **Detection of bacterial pathogens in Mongolia meningitis surveillance with a new real-time PCR assay to detect Haemophilus influenzae.** *Int J Med Microbiol* 2011, **301:**303-309.
299. Pimenta FC, Moiane B, Lessa FC, Venero AL, Moura I, Larson S, Massora S, Chauque A, Tembe N, Mucavele H, et al: **Dried blood spots for Streptococcus pneumoniae and Haemophilus influenzae detection and serotyping among children < 5 years old in rural Mozambique.** *BMC Pediatr* 2020, **20:**326.
300. Limato R, Lazarus G, Dernison P, Mudia M, Alamanda M, Nelwan EJ, Sinto R, Karuniawati A, Rogier van Doorn H, Hamers RL: **Optimizing antibiotic use in Indonesia: A systematic review and evidence synthesis to inform opportunities for intervention.** *The Lancet Regional Health - Southeast Asia* 2022, **2**.
301. CLSI: **Performance Standards for Antimicrobial Susceptibility Testing.** USA: Clinical and Laboratory Standard Institute; 2022.
302. **Breakpoint tables for interpretation of MICs and zone diameters.** The European Committee on Antimicrobial Susceptibility Testing; 2022.

303. Doern GV, Jorgensen JH, Thornsberry C, Snapper H: **Disk diffusion susceptibility testing of Haemophilus influenzae using haemophilus test medium.** *Eur J Clin Microbiol Infect Dis* 1990, **9:**329-336.
304. Jorgensen JH, Redding JS, Maher LA, Howell AW: **Improved medium for antimicrobial susceptibility testing of Haemophilus influenzae.** *J Clin Microbiol* 1987, **25:**2105-2113.
305. Andrews S: **FastQC: A Quality Control Tool for High Throughput Sequence Data.** 2010.
306. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics* 2014, **30:**2114-2120.
307. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications.** *BMC Bioinformatics* 2009, **10:**421.
308. Alcock BP, Huynh W, Chalil R, Smith KW, Raphenya AR, Wlodarski MA, Edalatmand A, Petkau A, Syed SA, Tsang KK, et al: **CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database.** *Nucleic Acids Res* 2023, **51:**D690-D699.
309. Hegstad K, Mylvaganam H, Janice J, Josefsen E, Sivertsen A, Skaare D: **Role of Horizontal Gene Transfer in the Development of Multidrug Resistance in Haemophilus influenzae.** *mSphere* 2020, **5**.
310. Johannessen H, Anthonisen IL, Zecic N, Hegstad K, Ranheim TE, Skaare D: **Characterization and Fitness Cost of Tn7100, a Novel Integrative and Conjugative Element Conferring Multidrug Resistance in Haemophilus influenzae.** *Front Microbiol* 2022, **13:**945411.
311. **Peraturan Presiden (PERPRES) Nomor 88 Tahun 2021 tentang Strategi Nasional Kelanjutusiaan.** Jaringan Dokumentasi dan Informasi Hukum BPK Indonesia; 2021.
312. Juhas M, Power PM, Harding RM, Ferguson DJ, Dimopoulou ID, Elamin AR, Mohd-Zain Z, Hood DW, Adegbola R, Erwin A, et al: **Sequence and functional analyses of Haemophilus spp. genomic islands.** *Genome Biol* 2007, **8:**R237.
313. Lam TT, Claus H, Elias J, Frosch M, Vogel U: **Ampicillin resistance of invasive Haemophilus influenzae isolates in Germany 2009-2012.** *Int J Med Microbiol* 2015, **305:**748-755.
314. Ubukata K, Shibasaki Y, Yamamoto K, Chiba N, Hasegawa K, Takeuchi Y, Sunakawa K, Inoue M, Konno M: **Association of amino acid substitutions in penicillin-binding protein 3 with beta-lactam resistance in beta-lactamase-negative ampicillin-resistant Haemophilus influenzae.** *Antimicrob Agents Chemother* 2001, **45:**1693-1699.
315. Mihara K, Tanabe T, Yamakawa Y, Funahashi T, Nakao H, Narimatsu S, Yamamoto S: **Identification and transcriptional organization of a gene cluster involved in biosynthesis and transport of acinetobactin, a siderophore produced by Acinetobacter baumannii ATCC 19606T.** *Microbiology (Reading)* 2004, **150:**2587-2597.
316. Ma C, Zhang Y, Wang H: **Characteristics of Haemophilus influenzae carriage among healthy children in China: A meta-analysis.** *Medicine (Baltimore)* 2023, **102:**e35313.
317. Nshimiyimana T, Najjuka CF, Nalwanga W, Katende G, Kateete DP: **Nasopharyngeal carriage and antibiotic susceptibility patterns of streptococcus pneumoniae, haemophilus influenzae, moraxella catarrhalis and staphylococcus aureus among urban Ugandan children post-PCV10 introduction: a cross-sectional study.** *Afr Health Sci* 2023, **23:**216-229.

318. Giufre M, Daprai L, Cardines R, Bernaschi P, Rava L, Accogli M, Raponi M, Garlaschi ML, Ciofi degli Atti ML, Cerquetti M: **Carriage of Haemophilus influenzae in the oropharynx of young children and molecular epidemiology of the isolates after fifteen years of H. influenzae type b vaccination in Italy.** *Vaccine* 2015, **33:**6227-6234.
319. Chan J, Nguyen CD, Dunne EM, Kim Mulholland E, Mungun T, Pomat WS, Rafai E, Satzke C, Weinberger DM, Russell FM: **Using pneumococcal carriage studies to monitor vaccine impact in low- and middle-income countries.** *Vaccine* 2019, **37:**6299-6309.
320. Jacoby P, Carville KS, Hall G, Riley TV, Bowman J, Leach AJ, Lehmann D, Kalgoorlie Otitis Media Research Project T: **Crowding and other strong predictors of upper respiratory tract carriage of otitis media-related bacteria in Australian Aboriginal and non-Aboriginal children.** *Pediatr Infect Dis J* 2011, **30:**480-485.
321. Shrestha S, Stockdale LK, Gautam MC, Gurung M, Feng S, Maskey P, Kerridge S, Kelly S, Voysey M, Pokhrel B, et al: **Impact of Vaccination on Haemophilus influenzae Type b Carriage in Healthy Children Less Than 5 Years of Age in an Urban Population in Nepal.** *J Infect Dis* 2021, **224:**S267-S274.
322. Yang Y, Pan X, Cheng W, Yang Y, Scherpbier RW, Zhu X, Chen Y, Zhou Y, Jiang Q: **Haemophilus influenzae type b carriage and burden of its related diseases in Chinese children: Systematic review and meta-analysis.** *Vaccine* 2017, **35:**6275-6282.
323. Adam HJ, Richardson SE, Jamieson FB, Rawte P, Low DE, Fisman DN: **Changing epidemiology of invasive Haemophilus influenzae in Ontario, Canada: evidence for herd effects and strain replacement due to Hib vaccination.** *Vaccine* 2010, **28:**4073-4078.
324. Leon ME, Kawabata A, Nagai M, Rojas L, Chamorro G, Zarate N, Gomez G, Leguizamon M, Irala J, Ortellado J, et al: **Epidemiologic study of Haemophilus influenzae causing invasive and non-invasive disease in Paraguay (1999-2017).** *Enferm Infecc Microbiol Clin (Engl Ed)* 2021, **39:**59-64.
325. Eton V, Schroeter A, Kelly L, Kirlew M, Tsang RSW, Ulanova M: **Epidemiology of invasive pneumococcal and Haemophilus influenzae diseases in Northwestern Ontario, Canada, 2010-2015.** *Int J Infect Dis* 2017, **65:**27-33.
326. Staples M, Graham RMA, Jennison AV: **Characterisation of invasive clinical Haemophilus influenzae isolates in Queensland, Australia using whole-genome sequencing.** *Epidemiol Infect* 2017, **145:**1727-1736.
327. Cleland G, Leung C, Wan Sai Cheong J, Francis J, Heney C, Nourse C: **Paediatric invasive Haemophilus influenzae in Queensland, Australia, 2002-2011: Young Indigenous children remain at highest risk.** *J Paediatr Child Health* 2018, **54:**36-41.
328. Nishimoto AT, Dao TH, Jia Q, Ortiz-Marquez JC, Echlin H, Vogel P, van Opijnen T, Rosch JW: **Interspecies recombination, not de novo mutation, maintains virulence after beta-lactam resistance acquisition in Streptococcus pneumoniae.** *Cell Rep* 2022, **41:**111835.
329. Royer G, Clermont O, Marin J, Condamine B, Dion S, Blanquart F, Galardini M, Denamur E: **Epistatic interactions between the high pathogenicity island and other iron uptake systems shape Escherichia coli extra-intestinal virulence.** *Nat Commun* 2023, **14:**3667.

330. Elling CL, Ryan AF, Yarza TKL, Ghaffar A, Llanes E, Kofonow JM, Reyes-Quintos MRT, Riazuddin S, Robertson CE, Tantoco MLC, et al: **A Novel SLPI Splice Variant Confers Susceptibility to Otitis Media in Humans.** *Int J Mol Sci* 2025, **26**.
331. Mittal R, Sanchez-Luege SV, Wagner SM, Yan D, Liu XZ: **Recent Perspectives on Gene-Microbe Interactions Determining Predisposition to Otitis Media.** *Front Genet* 2019, **10:**1230.
332. Zheng K, He FB, Liu H, He Q: **Genetic variations of toll-like receptors: Impact on susceptibility, severity and prognosis of bacterial meningitis.** *Infect Genet Evol* 2021, **93:**104984.
333. Putra RL: **POLA PENGGUNAAN ANTIBIOTIK PADA PASIEN KOMUNITAS DI APOTEK "X" PALEMBANG.** *SOCIAL CLINICAL PHARMACY INDONESIA JOURNAL* 2023, **7:**9-14.
334. Gach MW, Lazarus G, Simadibrata DM, Sinto R, Saharman YR, Limato R, Nelwan EJ, van Doorn HR, Karuniawati A, Hamers RL: **Antimicrobial resistance among common bacterial pathogens in Indonesia: a systematic review.** *The Lancet Regional Health - Southeast Asia* 2024, **26**.
335. NICE: **Meningitis (bacterial) and meningococcal disease: recognition, diagnosis, and management.** pp. 43. London: National Institute for Health and Care Excellence; 2024:43.
336. Tristram S, Jacobs MR, Appelbaum PC: **Antimicrobial resistance in Haemophilus influenzae.** *Clin Microbiol Rev* 2007, **20:**368-389.
337. Belinda R, Subarnas A, Mutiara I: **RASIONALITAS PENGGUNAAN ANTIBIOTIKA MENGGUNAKAN METODE GYSSENS PADA PASIEN POLI BEDAH MULUT DI RUMAH SAKIT GIGI DAN MULUT UNIVERSITAS PADJADJARAN BANDUNG** *Farmaka* 2022, **20:**9-16.
338. Gach MW, Lazarus G, Simadibrata DM, Sinto R, Saharman YR, Limato R, Nelwan EJ, van Doorn HR, Karuniawati A, Hamers RL: **Antimicrobial resistance among common bacterial pathogens in Indonesia: a systematic review.** *Lancet Reg Health Southeast Asia* 2024, **26:**100414.
339. DelaFuente J, Diaz-Colunga J, Sanchez A, San Millan A: **Global epistasis in plasmid-mediated antimicrobial resistance.** *Mol Syst Biol* 2024, **20:**311-320.
340. Porse A, Jahn LJ, Ellabaan MMH, Sommer MOA: **Dominant resistance and negative epistasis can limit the co-selection of de novo resistance mutations and antibiotic resistance genes.** *Nat Commun* 2020, **11:**1199.
341. Jansen van Rensburg MJ, Berger DJ, Yassine I, Shaw D, Fohrmann A, Bray JE, Jolley KA, Maiden MCJ, Brueggemann AB: **Development of the Pneumococcal Genome Library, a core genome multilocus sequence typing scheme, and a taxonomic life identification number barcoding system to investigate and define pneumococcal population structure.** *Microb Genom* 2024, **10**.
342. Unitt A, Krisna M, Parfitt KM, Jolley KA, Maiden MCJ, Harrison OB: **Neisseria gonorrhoeae LIN codes: a Robust, Multi-Resolution Lineage Nomenclature.** *bioRxiv* 2025**:**2025.2003.2028.646058.
343. Zhang D, Wang W, Song C, Huang T, Chen H, Liu Z, Zhou Y, Wang H: **Comparative genomic study of non-typeable Haemophilus influenzae in children with pneumonia and healthy controls.** *iScience* 2024, **27:**111330.
344. Pickering J, Binks MJ, Beissbarth J, Hare KM, Kirkham LA, Smith-Vaughan H: **A PCR-high-resolution melt assay for rapid differentiation of nontypeable**

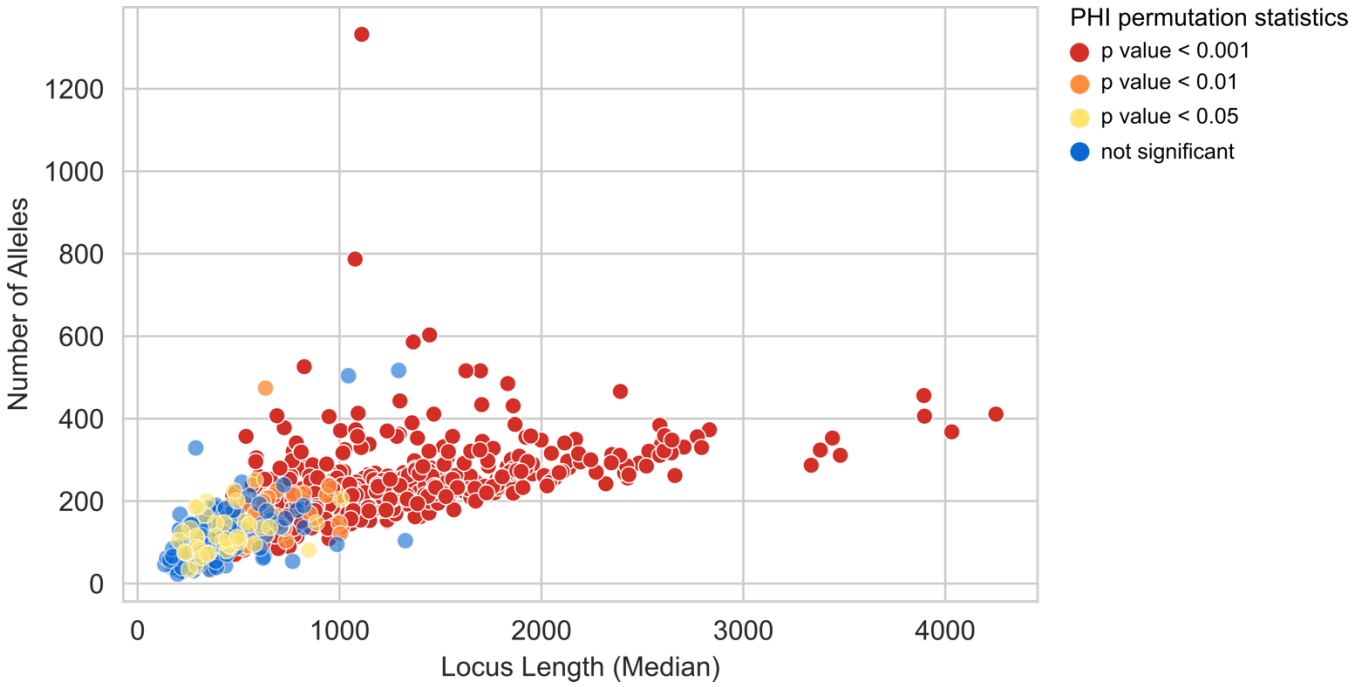
Haemophilus influenzae and Haemophilus haemolyticus. *J Clin Microbiol* 2014, **52:**663-667.

345. Zhang B, Kunde D, Tristram S: **Haemophilus haemolyticus is infrequently misidentified as Haemophilus influenzae in diagnostic specimens in Australia.** *Diagn Microbiol Infect Dis* 2014, **80:**272-273.

Appendix 2

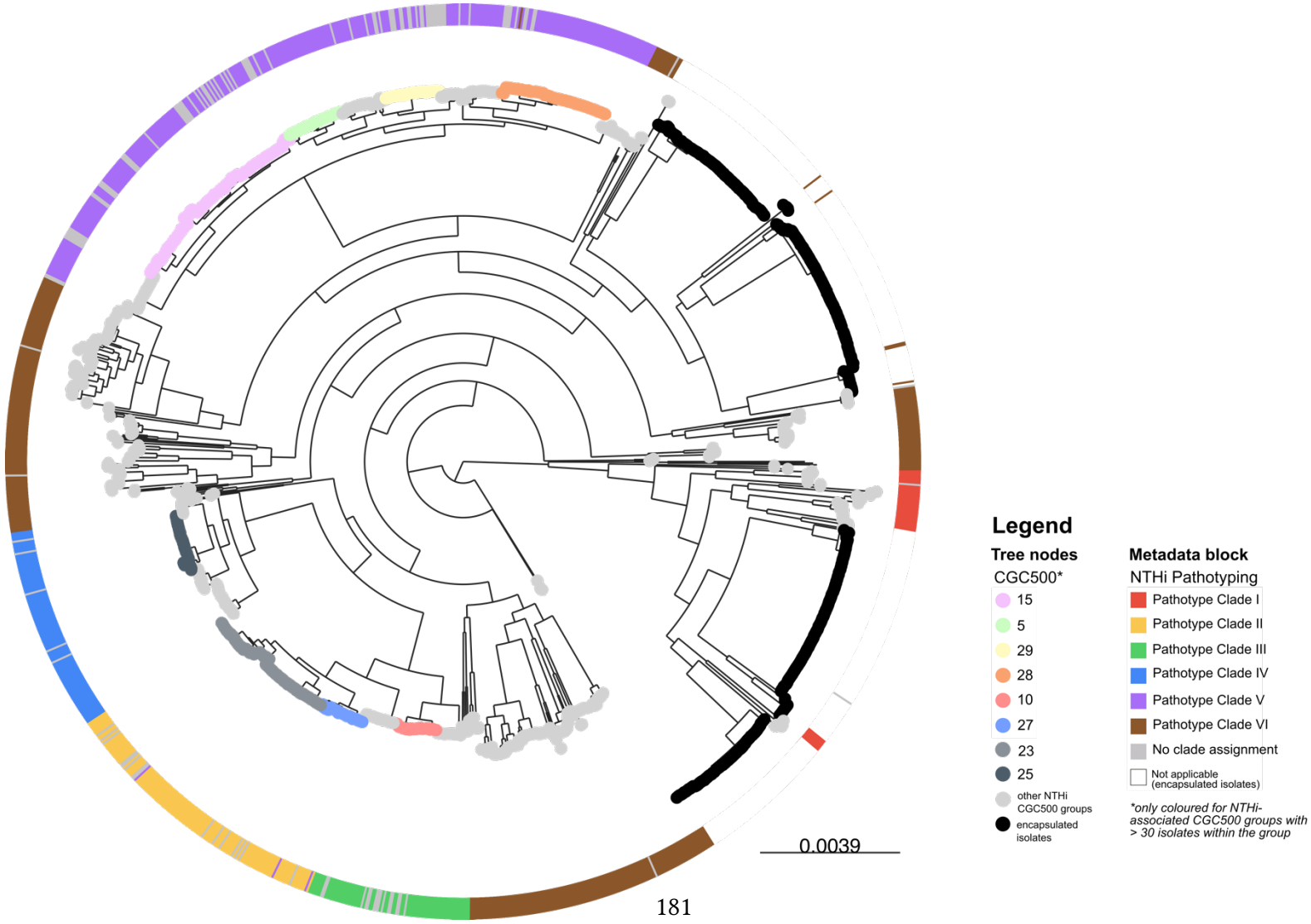
Appendix 2.2. Scatter plot of the number of alleles and length variation for each core gene in the cgMLST scheme, coloured by PHI permutation statistics.

The median locus length and the allele count were plotted, as shown on Figure 2.4 panel c; however, the plot was coloured based on the PHI permutation statistics p-value.

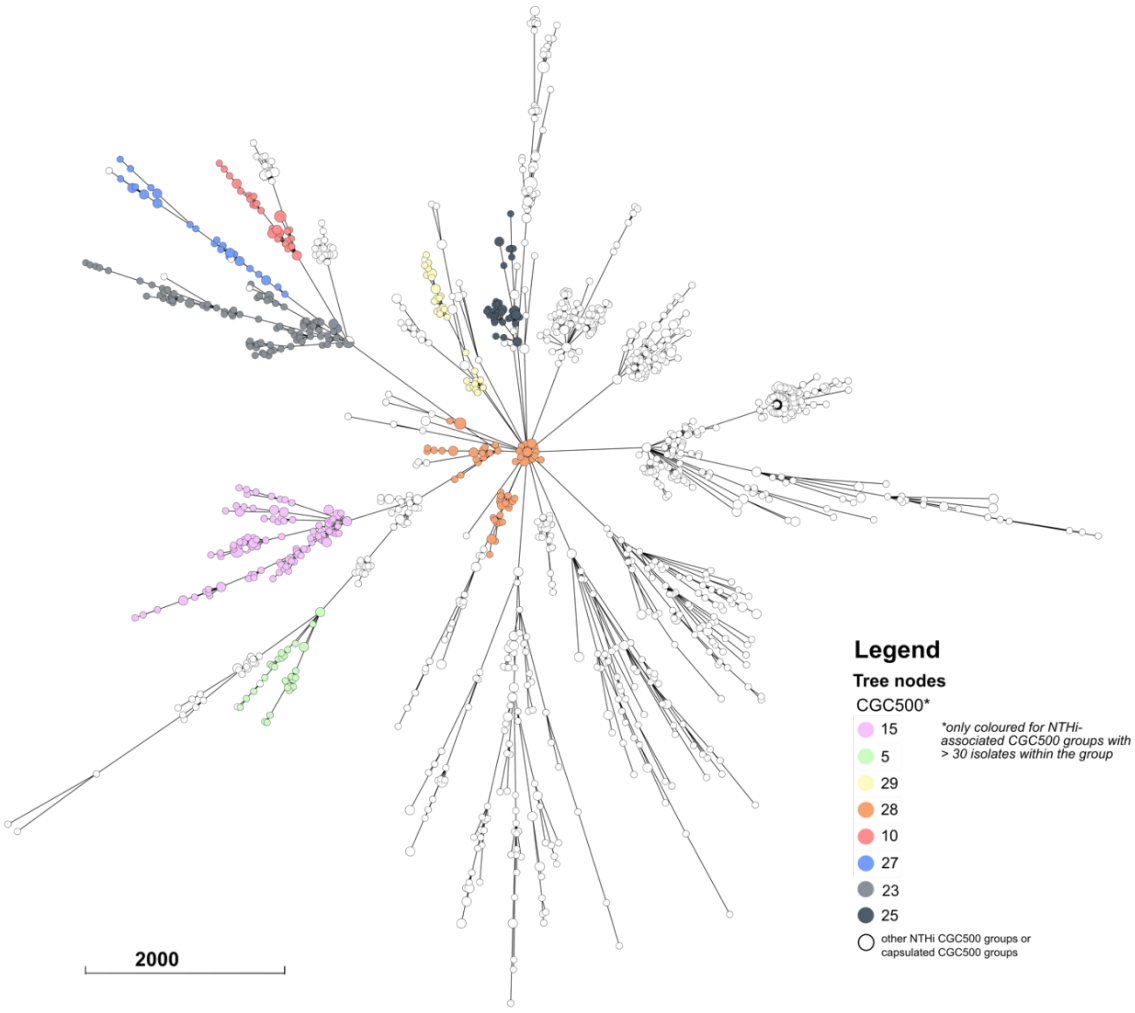


Appendix 2.3.1. The maximum-likelihood tree from core genome alignment of 1,376 *H. influenzae* genomes in the validation dataset.

Tree nodes with black colour were encapsulated isolates. For NTHi, tree nodes were coloured with the CGC500 group, if the group consisted of at least 30 isolates; otherwise, the nodes were coloured as grey. Metadata block shows different NTHi pathotype clades.



Appendix 2.3.2. A minimum-spanning tree based on core genome profile, showing different CGC500, CGC200, and CGC50 groups clustered together.

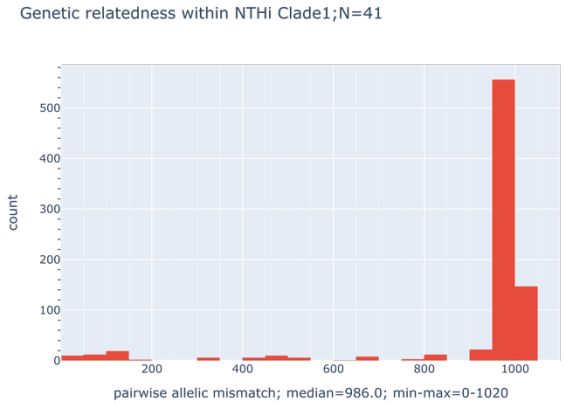


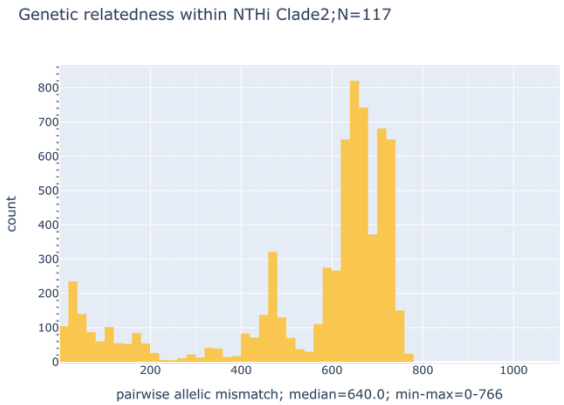
Although direct comparison based on the core genome allelic profile or its cgST is useful in the context of a possible outbreak for epidemiological tracking, almost all *H. influenzae* genomes in the PubMLST database (as accessed May 2023) has its own cgST. Therefore, to allow for a more straightforward comparison between the cgMLST scheme and pre-existing classification systems, we clustered cgSTs into core genome cluster (CGC) groups based on the number of core genome allelic mismatches among different cgSTs. The CGC thresholds were chosen arbitrarily, as reported by other highly recombining human pathogens such as *Klebsiella pneumoniae* species complex (Kpsc) [174], *Streptococcus pneumoniae* [122], and *Neisseria gonorrhoeae* [120]. Initially, as implemented on the PubMLST *H. influenzae* database, there were 7 allelic mismatch thresholds for cgST clustering: 500, 400, 300, 200, 100, 50, and 25. We mapped the CGC groups for each threshold to the phylogeny and compared the clustering results with each other. We found that between 500 and 400 allelic mismatch thresholds, the resulting CGC groups were very similar; as well as between 300 and 200, and 100 and 50. Therefore we chose only 500, 200, and 50 allelic mismatches thresholds for the purpose of comparing the core genome clustering to pre-existing classification system. The mapping of CGC group clustering at multiple thresholds to the phylogeny can

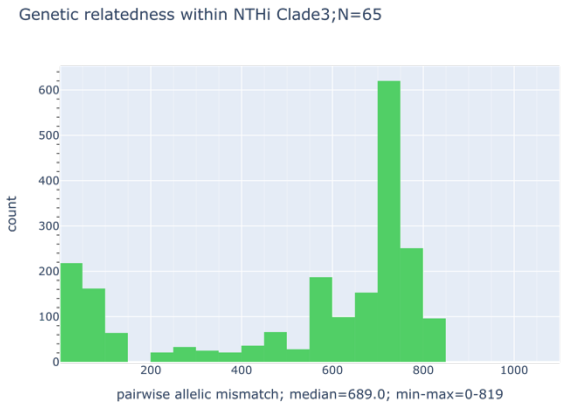
be accessed from: https://microreact.org/project/aBH3zddMifvweC2KJNEnN4-phylogenycgccgmlstproject.

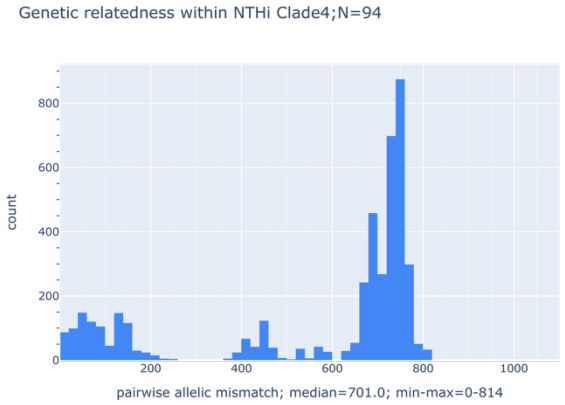
Appendix 2.4. Variation of genetic relatedness among NTHi isolates within the same pathotype clade.

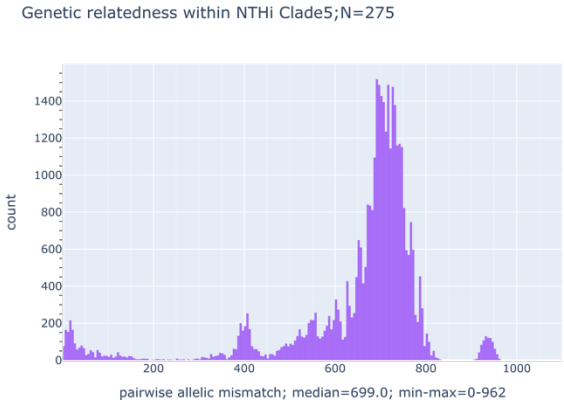
Genetic relatedness was reflected by the number of allelic mismatches for each possible combination of paired isolates (i.e. pairwise allelic mismatch).

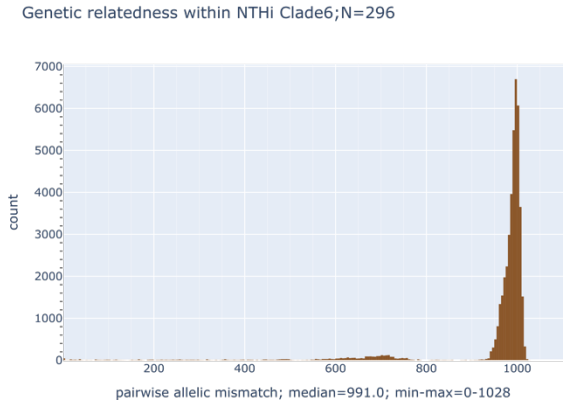






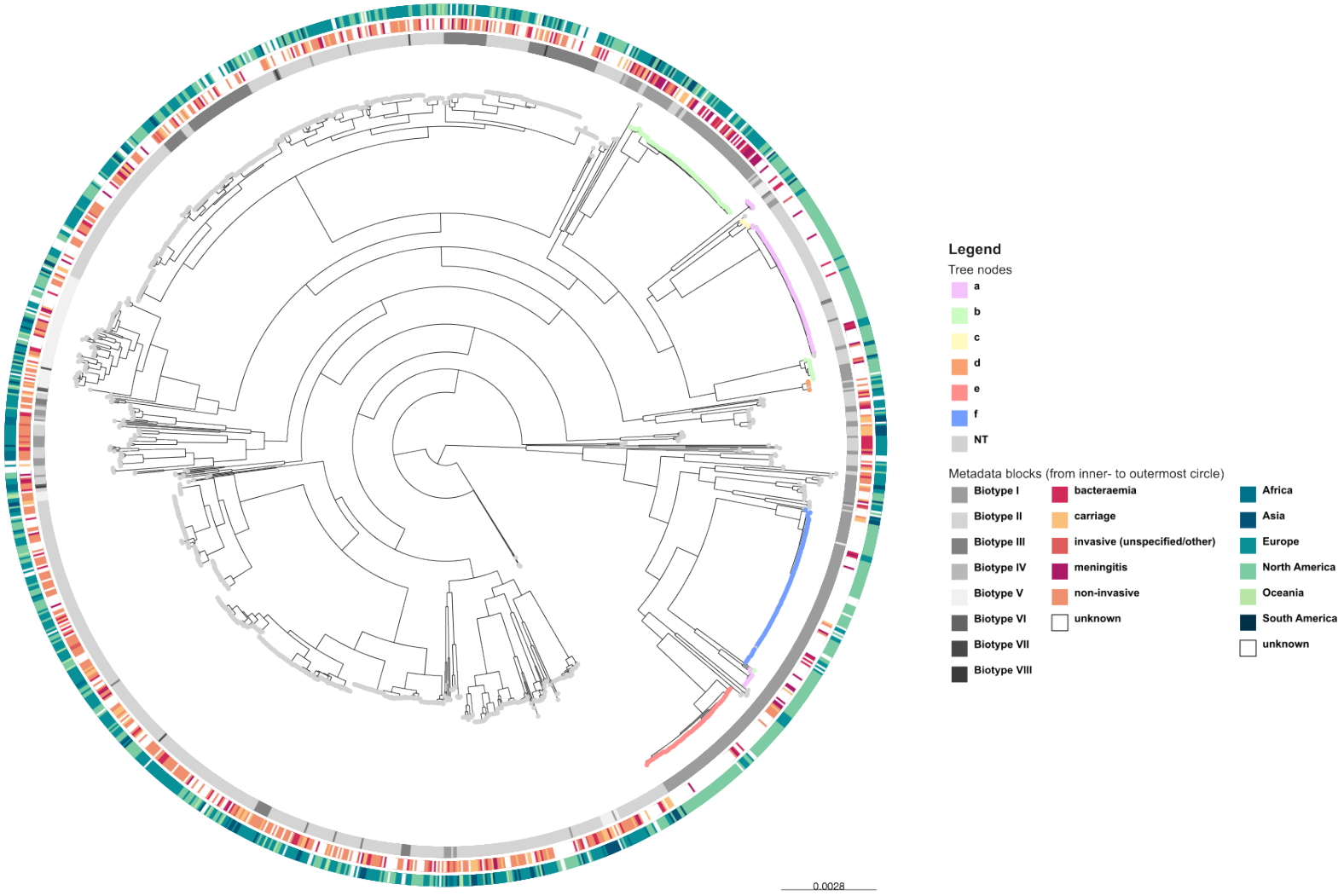






Appendix 2.5. The maximum-likelihood tree from core genome alignment of 1,376 *H. influenzae* genomes in the validation dataset.

Tree nodes were coloured based on the capsule type. The innermost metadata block showed biotype assignments, based on the presence/absence of genes encoding ornithine decarboxylase (ODC), urease, and tryptophanase. The middle and outermost metadata block showed disease associated with the isolates and the continent where the isolates originating, respectively.

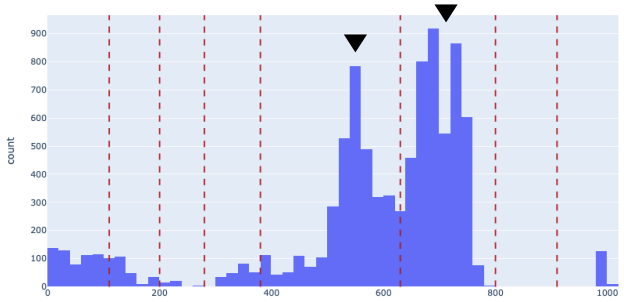


Appendix 3

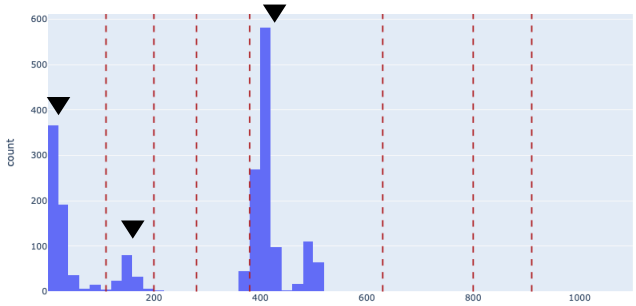
Appendix 3.1. Distribution curves of core genome allelic mismatches in H. influenzae subpopulation based on serotype and clonal complex.

Clonal complexes of NTHi.

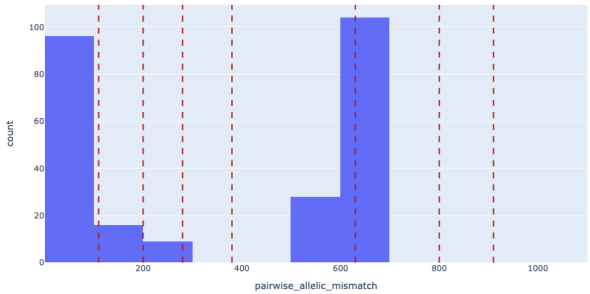
ST-3 complex (n = 135)



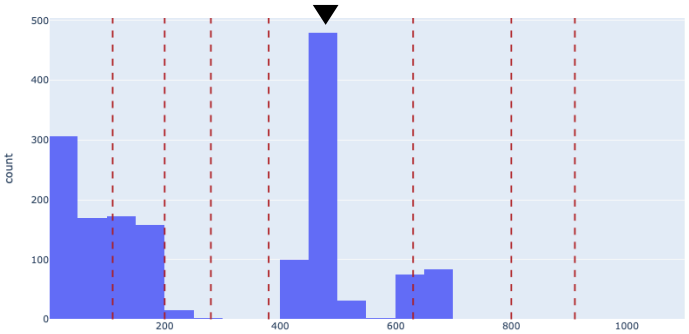
ST-11 complex (n = 63)



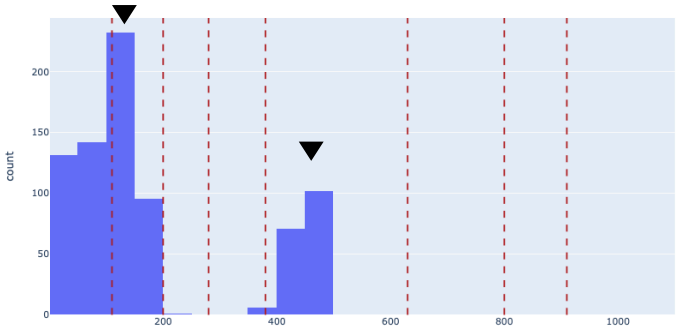
ST-139 complex (n = 23)



ST-107 complex (n = 57)

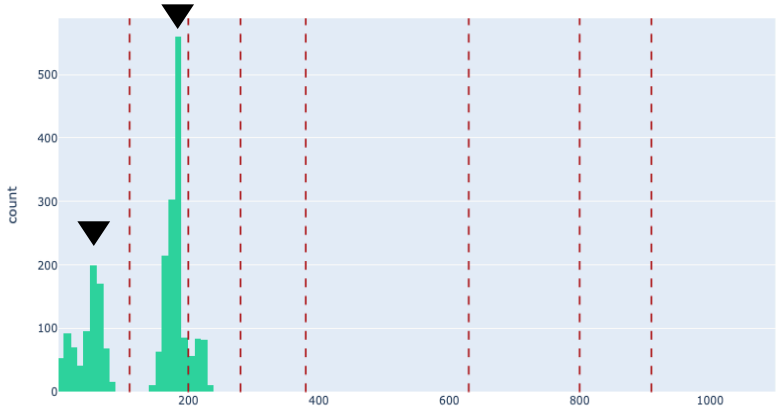


ST-57 complex (n = 40)

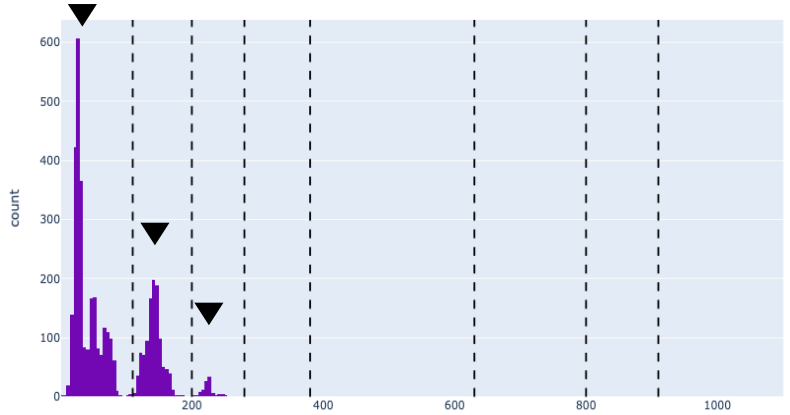


Clonal complexes of typeable *H. influenzae.*

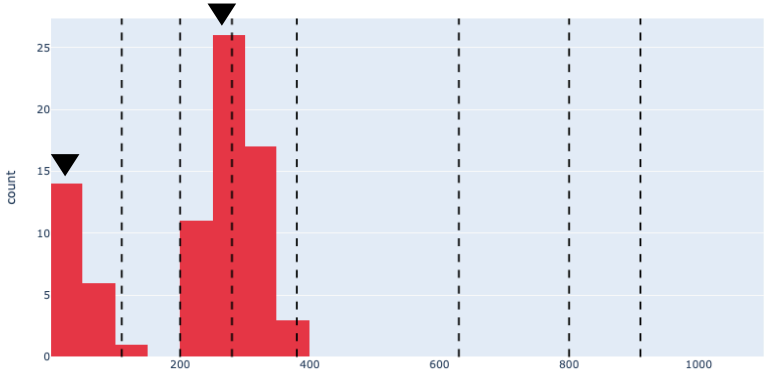
Type a ST23 complex (N = 68)



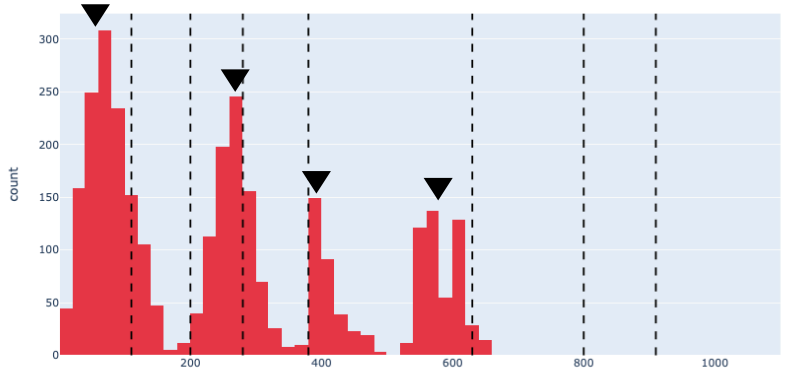
Type f ST124 complex (N = 78)



Type b ST222 complex (N = 13)

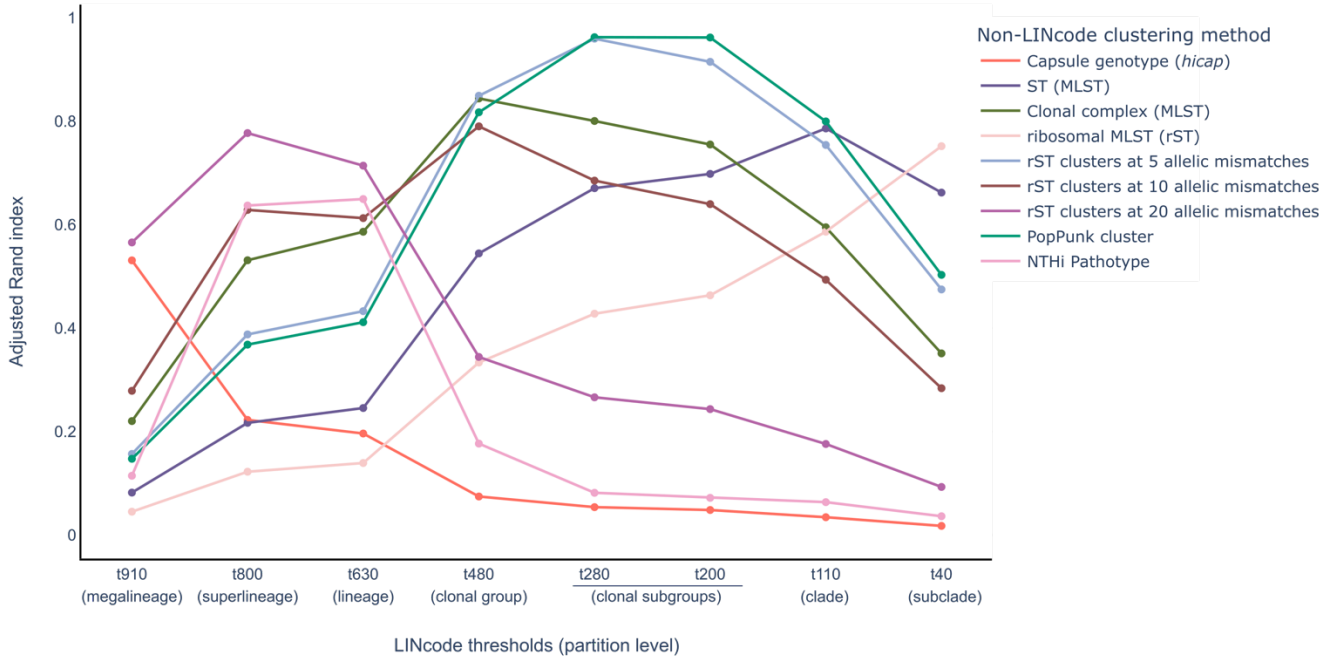


Type b ST6 complex (N = 78)



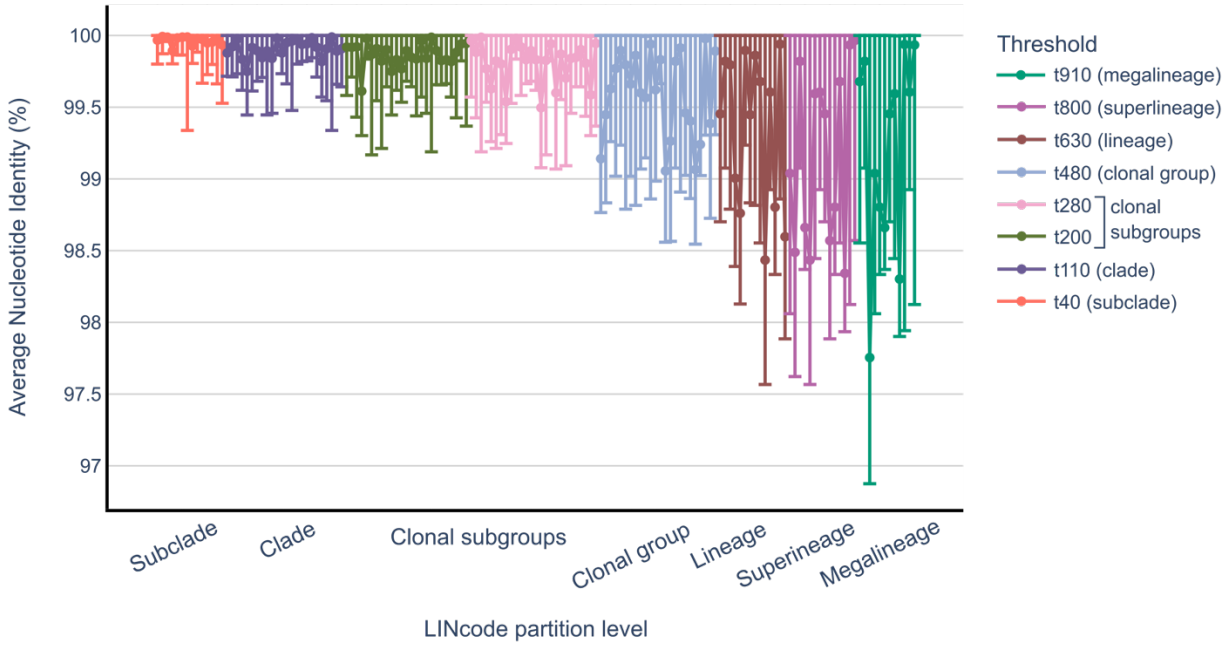
Appendix 3.2. Comparison between clustering results using LIN method at different threshold and pre-existing methods in the literature.

Clustering congruence was evaluated using adjusted Rand index with value close to 1 indicating high agreement. This assessment indicates that pre-existing method with less resolution such as MLST (ST and CC), or even capsule type, is still valuable for *H. influenzae* clustering when WGS is not possible. CC: Clonal complex; NTHi: non-typeable *H. influenzae*.The results showed that pre-existing method had high (> 0.5) concordance with at least one LINcode threshold. For instance, capsule grouping is similar to clustering at t910, clonal complex to t480, ST to t110, rST to t40, PopPunk cluster to t200, and NTHi pathotype to t630.



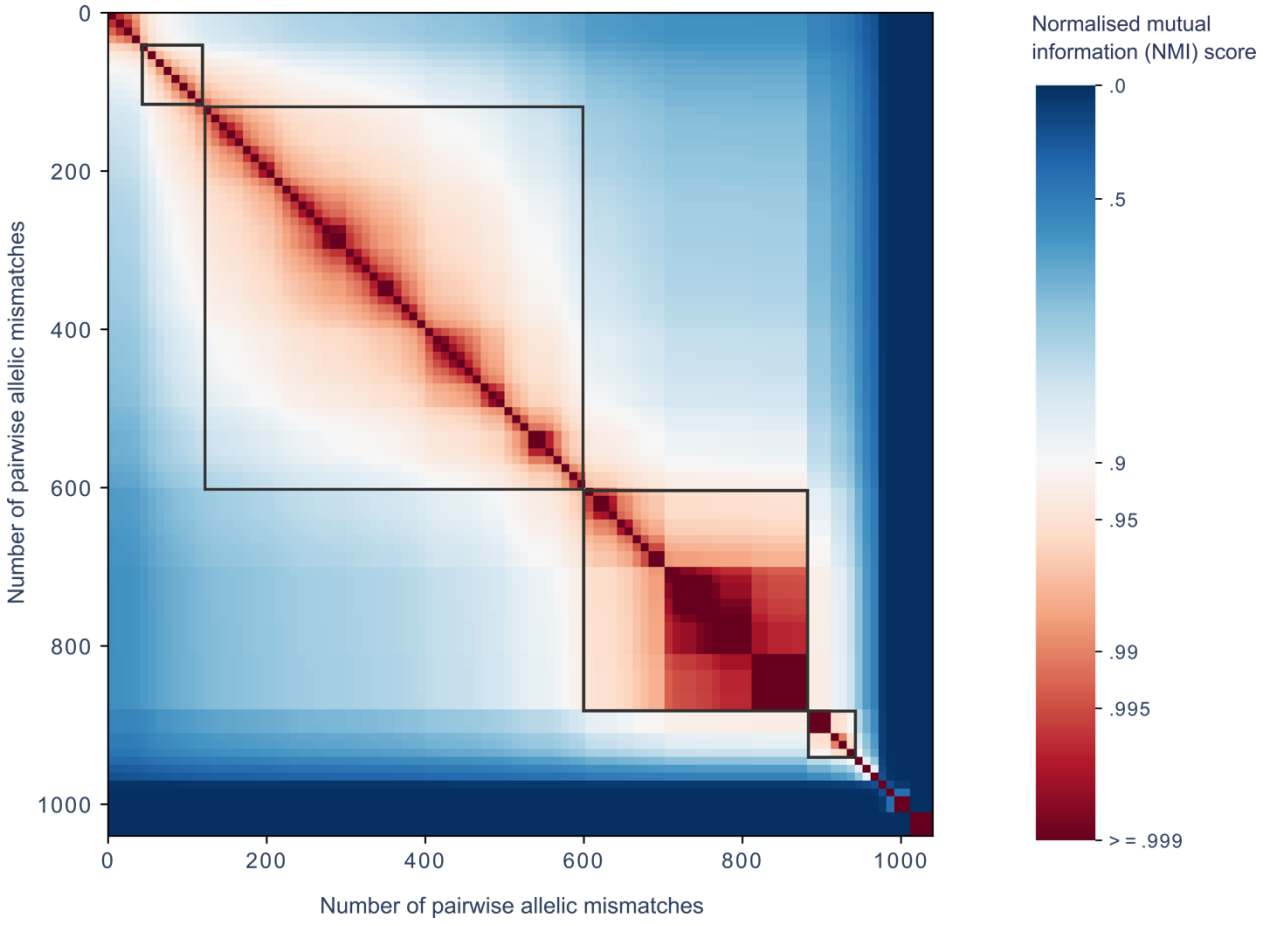
Appendix 3.3. Variation of average nucleotide identity (ANI) values within clusters formed at different LINcode thresholds.

With the main LINcode partitioning levels defined, *H. influenzae* genomes were clustered at each similarity threshold. Pairwise ANI comparison between genomes belonging to the same cluster was conducted and the intracluster variation of ANI at each LINcode threshold was evaluated. Although all pairwise comparisons resulted in ANI of ≥ 95%, clusters at higher mismatch threshold had a greater variability in ANI value. Circles in the plot indicate median value, with maximum and minimum value as the upper and lower limit, respectively. Each threshold was proposed a name, from superlineage at 910 allelic mismatches, to clonal subgroups at 40 allelic mismatches. This observation confirmed that the *H. influenzae* clusters defined using the LINcode effectively reflect the genetic relatedness among isolates.



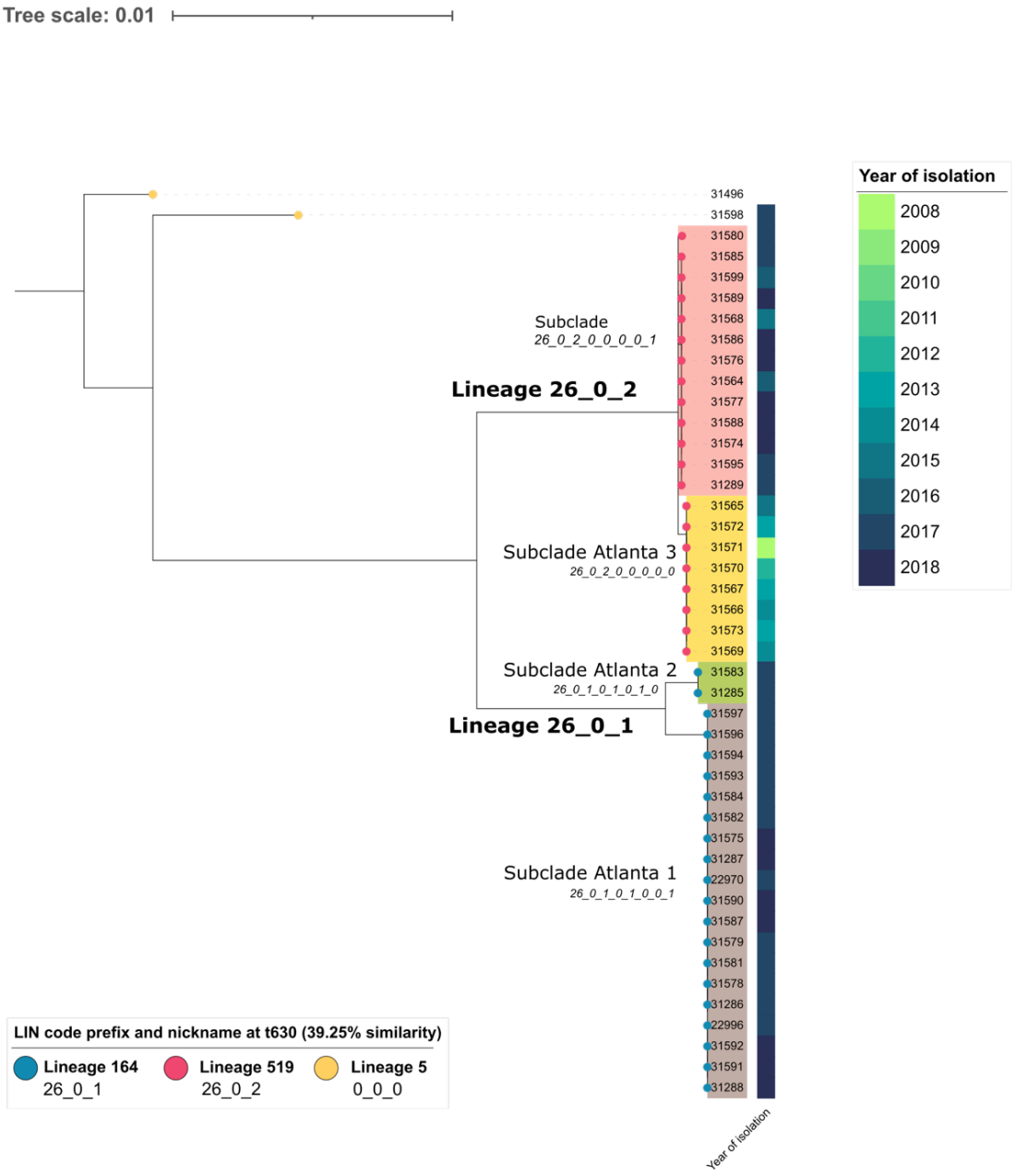
Appendix 3.4. Heat map of Normalised Mutual Information (NMI) calculated for pairwise clustering result comparison for each possible threshold.

Areas marked by black rectangles indicated a stable clustering threshold range.



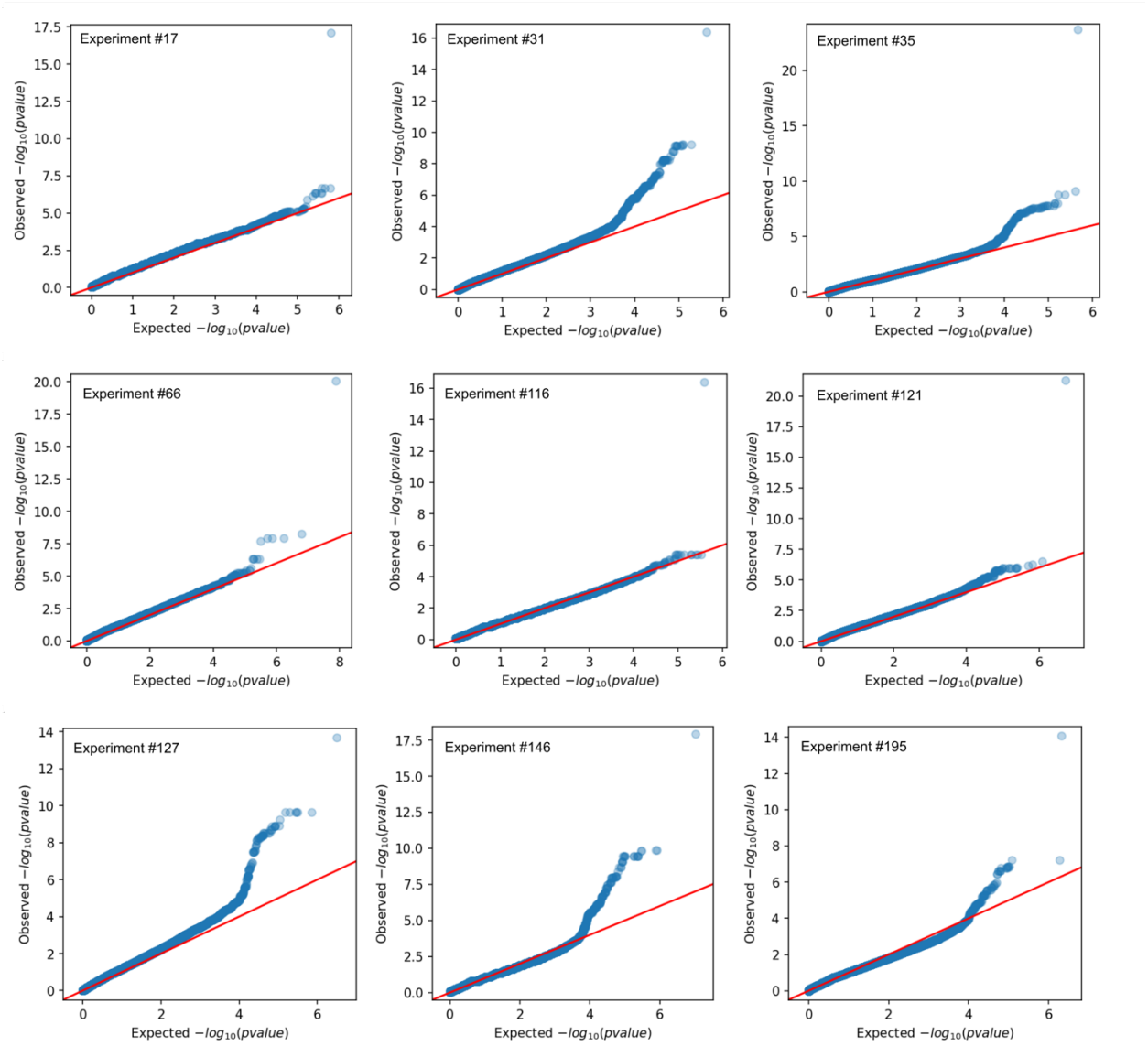
Appendix 3.5. A neighbor-joining tree of H. influenzae isolates of clinical and public health importance, as described by Collins et al (2019), was annotated with LIN code prefixes and corresponding nicknames.

These isolates were derived from invasive infections in adults with HIV, identified in Atlanta, U.S.A, between 2008 and 2018.

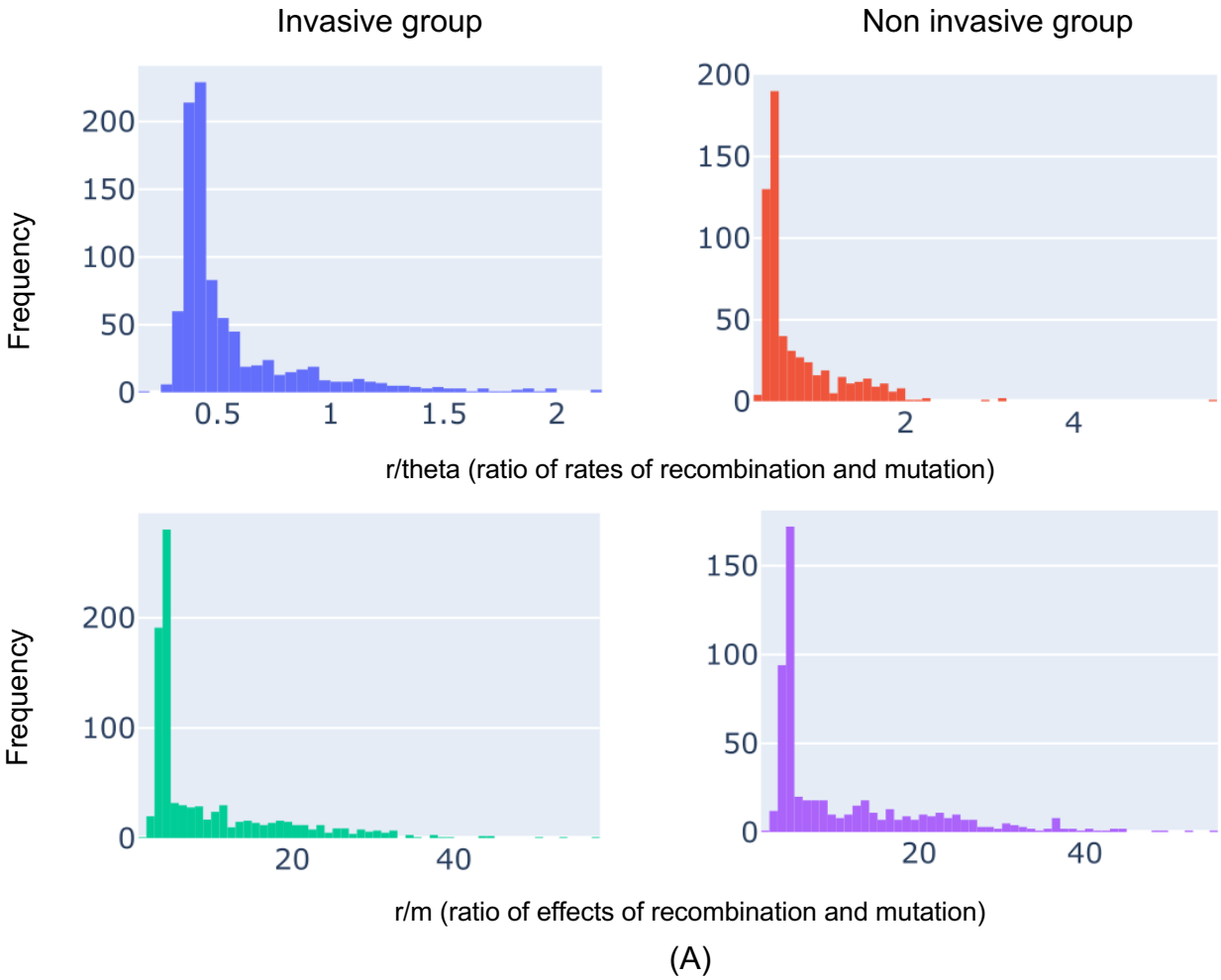


Appendix 4

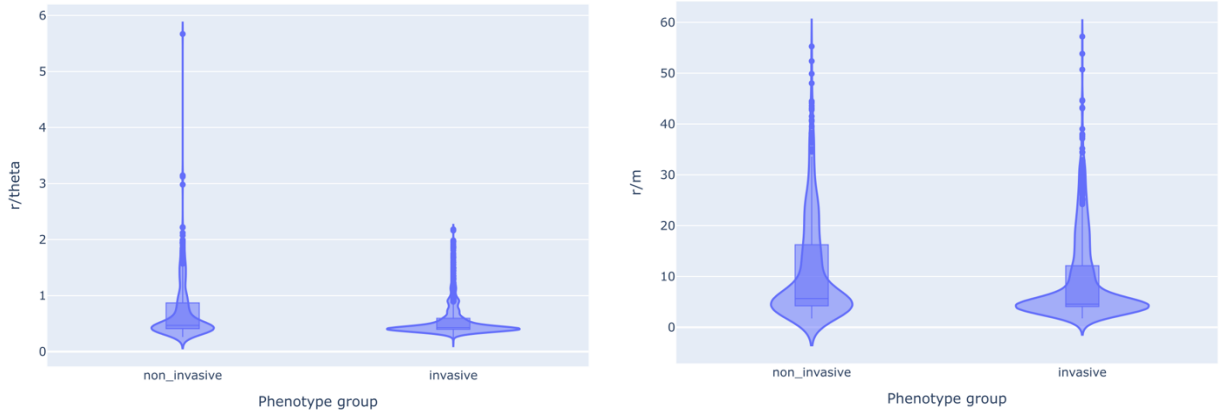
Appendix 4.1. Q-Q plot of randomly chosen GWAS experiments.



Appendix 4.2. Recombination parameters distribution between invasive and non invasive group among NTHi population as histogram (A) and violin plot (B)



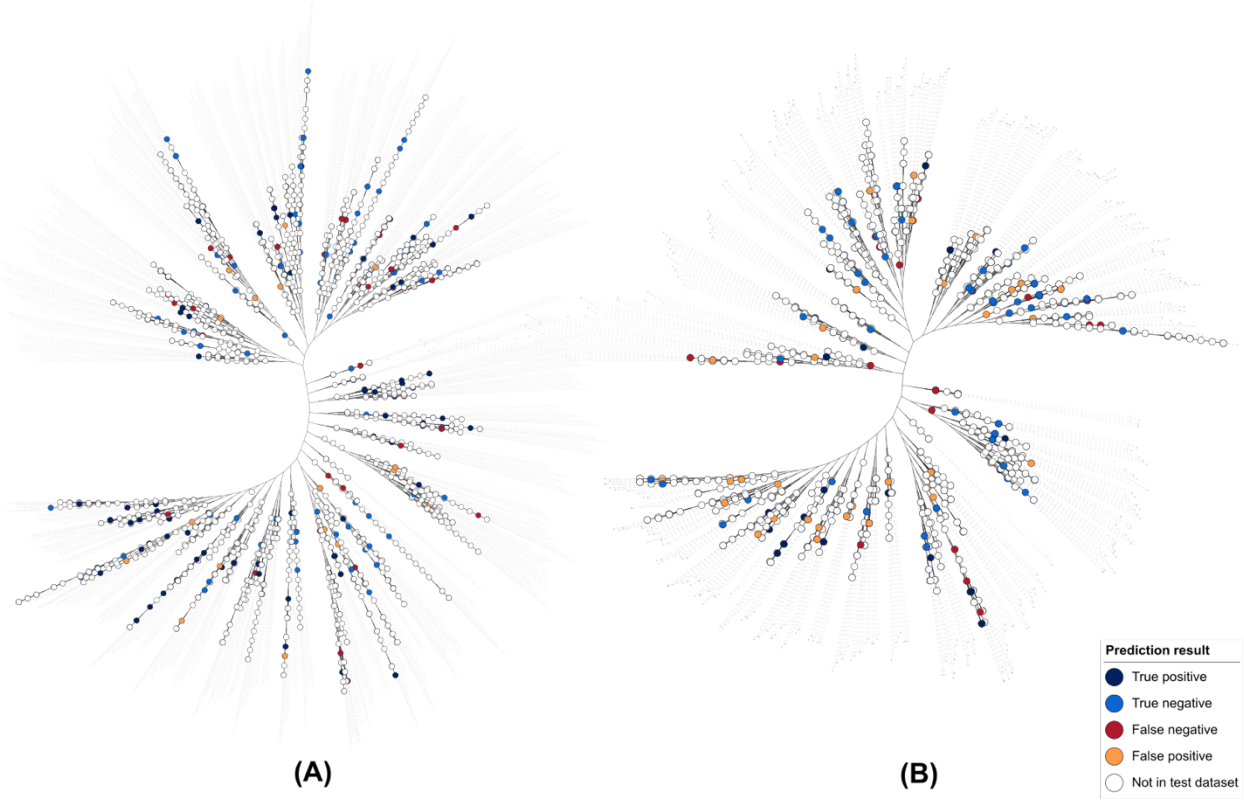
(A)



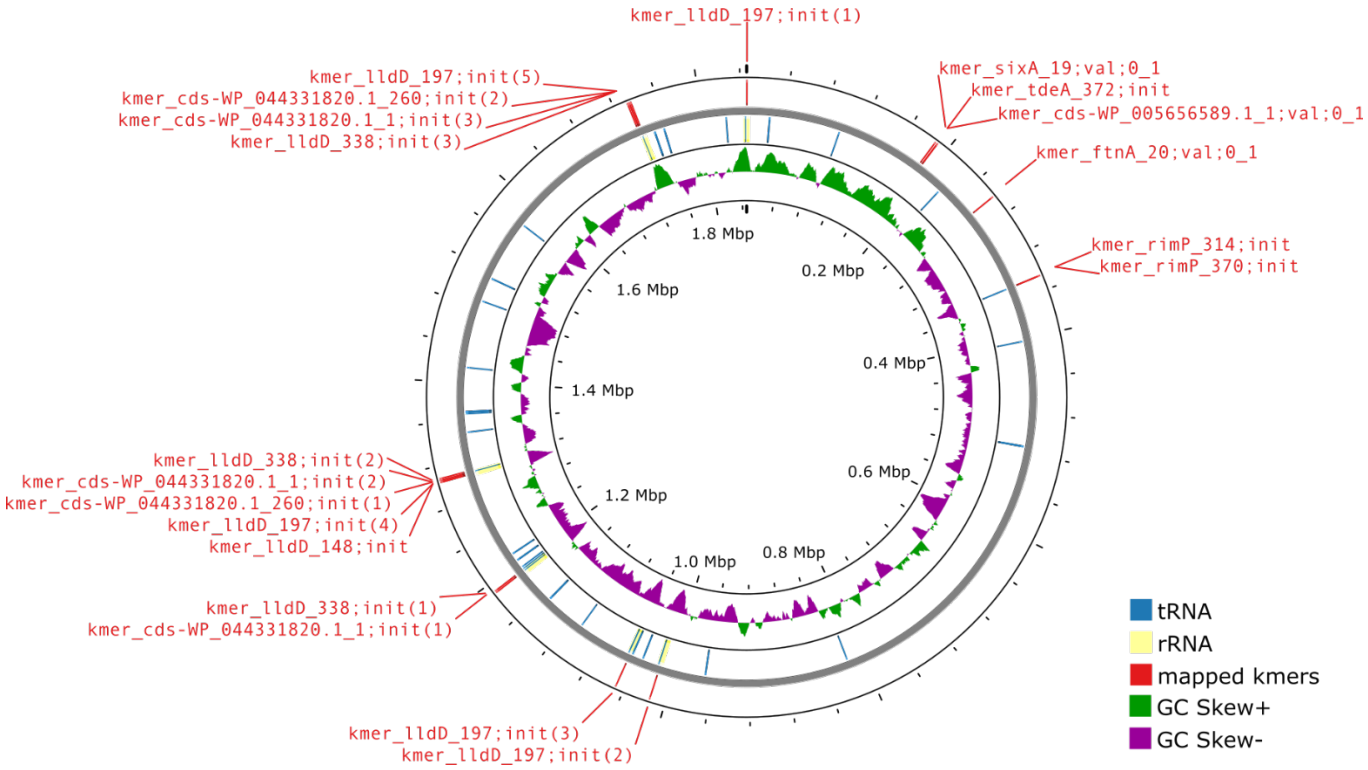
(B)

Appendix 4.3. The phylogenetic tree of NTHi genomes in (A) dataset 1 and (B) dataset 2, annotated by the logistic regression model result.

Only genomes which were randomly assigned to the test dataset (20%) in the model construction had the prediction results.

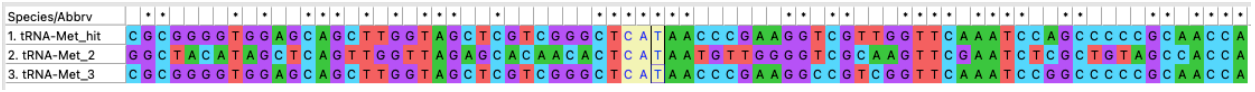


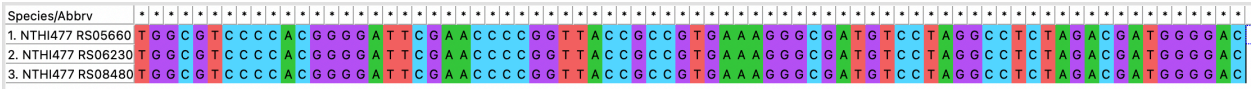
Appendix 4.4. Variants associated with invasive NTHi infection in non-protein coding and intergenic regions as aligned to *H. influenzae* reference genome (GCF_000931575.1)



Appendix 4.5. Multiple sequence alignments of (a) tRNA-Methionine (tRNA-Met) and (b) tRNA-Glutamate (tRNA-Glu) genes present in the NTHi reference genome (GCF_000931575.1).

(a) one k-mer variant associated with the NTHi invasive phenotype only mapped to the first tRNA-Met ("tRNA-Met_hit"). Nucleotides with yellow highlight were the anticodon position. (b) Identical tRNA-Glu genes were present in triplicate in the NTHi reference genome.





Appendix 5

Appendix 5.1. Probe and primers for *hpd* gene detection and serotyping of *Haemophilus influenzae.*

Target gene	Primer or Probe Name	Real-time Primers and Probes Nucleotide Sequence (5' to 3')	Working Stock* (µM)	Final* (nM)	Suggested Probe Modifications
hpd [298]	hpdF822	GGTTAAATATGCCGATGGTGTTG	1.25	100	
	hpdR952	TGCATCTTTACGCACGGTGTA	3.75	300	
	Pb896i	TTGTGTACACTCCGT"T"GGTAAAAGAACTTGCAC	1.25	100	5' FAM, BHQ1 on "T", 3' SpC6
Serotyping [10, 299]					
Hia	F261	GGT CTG CGG TGT CCT GTG T	1.25	100	
acsB	R427	CCG GTC ATC TTT TAT GCT CCA A	3.75	300	
	Pb375i	TAA TTT TCT TGC "T"CA ATA CCG CCT TCC CA	1.25	100	5' FAM, BHQ1 on "T", 3' SpC6
Hib	F192	TGA TGC ATT GAA AGA AGG TGT AAT TT	3.75	300	
bcsB	R359	CCT GCG GTA ATA ACA TGA TCA TAA A	7.5	600	
	Pb244i	TGT CGT GCA G"T"A GCA AAC CGT AAC CTT ACT C	1.25	100	5' FAM, BHQ1 on "T", 3' SpC6
Hic	F7667	CAT TGG TGA TGG TTC AGT TAT TGG	7.5	600	
ccsD	R7784	TAC AGC ATT CAG CAA TAA TGG G	3.75	300	
	Pb7726i	ATT GCA "T"CG CCG CAG GAG TTC CCG	2.5	200	5' FAM, BHQ1 on "T", 3' SpC6
Hid	F2211	CCT AAA ATA CGG ACC TAG TGC AC	7.5	600	
dcsE	R2255	CCG ATG AGA CCA AGT ATG GTT A	1.25	100	
	Pb2221i	AAC GAG C"T"A GAG CTG GTG CTG AA	3.75	300	5' FAM, BHQ1 on "T", 3' SpC6
Hie	F1523	ACT AAA ATA TGG CCC AAA CCC AC	7.5	600	
ecsH	R1589	CCG ATG AGC CCA AGT ATG ATG A	7.5	600	
	Pb1555i	AAC GAG CAA AAG CCG G"T"G CGG AT	2.5	200	5' FAM, BHQ1 on "T", 3' SpC6
Hif	F7164	CCC TGA AAA GCG TTG ACT TTG	7.5	600	
bexD	R7313	CCA ACT TCA GGA CCA AGT CAT TC	3.75	300	
	Pb7242i	TGC TGC TAA C"T"C AGA TGC ATC AGC TCC TT	2.5	200	5' FAM, BHQ1 on "T", 3' SpC6

Note: Thermal cycler settings
1x: 50°C for 2min
1x: 95°C for 10min
40x:
95°C for 15sec

60°C for 1min

Appendix 5.2. Genome length and ribosomal MLST (rMLST) designation for four *H. influenzae* isolates from Indonesia excluded from downstream analyses.

PubMLST ID	Source	Genome length	rMLST result			Note
			rST	Species	Support[b]	
19483[a]	Carriage	3,615,653	288316	*H. influenzae* *Thermus scotoductus*	57% 21%	Concluded as contamination, as supported by the unusually long genome and rMLST result.
19495[a]	Carriage	3,428,049	190875	*H. influenzae* *T. scotoductus*	65% 10%	
19516[a]	Carriage	1,966,079	277088	*H. influenzae* *T. scotoductus*	95% 2%	Concluded as contamination, as supported but the rMLST result and that the 2% support was for the same contaminating species as 2 other contaminated isolates.
19551	Carriage	1,778,263	190920	*H. haemolyticus*	69%	Concluded as misidentification. *In silico* PCR confirms the false positive *hpd* gene detection[c] (ruled out cross-contamination during qPCR process)

[b] Support percentage is interpreted as the percentage of the designated ribosomal gene alleles for that isolate which were associated with a specific species.
[a] DNA extraction and library preparation for whole-genome sequencing were performed concurrently in a single batch.
[c] Both *H. influenzae* and *H. haemolyticus* possessed *hpd* gene and misidentification between the two species had been reported [344, 345].

Appendix 5.3. *ftsI* alleles assigned to non-sensitive *H. influenzae* isolates from Indonesia and the point mutations associated with with each allele.

***ftsI* allele**	**D350**	**S357**	**A568**	**A437**	**M377**	**S385**	**L389**	**I449**	**G490**	**A502**	**R517**	**N526**	***ftsI* group**
1	D350N	WT (S)	WT (A)	WT (A)	M377I	WT (S)	WT (L)	WT (I)	WT (G)	A502V	WT (R)	N526K	IIB[a]
2	D350N	WT (S)	WT (A)	WT (A)	M377I	WT (S)	WT (L)	WT (I)	WT (G)	A502V	WT (R)	N526K	IIB[a]
5	WT (D)	WT (S)	WT (A)	WT (A)	WT (M)	WT (S)	WT (L)	I449V	WT (G)	WT (A)	WT (R)	N526K	IID
6	D350N	WT (S)	WT (A)	WT (A)	WT (M)	WT (S)	WT (L)	WT (I)	WT (G)	WT (A)	WT (R)	WT (N)	WT
15	WT (D)	WT (S)	WT (A)	WT (A)	WT (M)	WT (S)	WT (L)	WT (I)	WT (G)	WT (A)	WT (R)	WT (N)	WT
23	D350N	WT (S)	WT (A)	WT (A)	M377I	WT (S)	WT (L)	WT (I)	G490E	A502V	WT (R)	N526K	IIB[b]
26	D350N	S357N	WT (A)	WT (A)	M377I	S385T	L389F	WT (I)	WT (G)	WT (A)	R517H	WT (N)	III-like+
27	WT (D)	WT (S)	WT (A)	WT (A)	WT (M)	WT (S)	WT (L)	WT (I)	WT (G)	WT (A)	WT (R)	WT (N)	WT
37	WT (D)	WT (S)	WT (A)	WT (A)	WT (M)	WT (S)	WT (L)	WT (I)	WT (G)	WT (A)	WT (R)	WT (N)	WT
43	D350N	WT (S)	WT (A)	WT (A)	WT (M)	WT (S)	WT (L)	WT (I)	G490E	WT (A)	WT (R)	N526K	IIA[c]
151	D350N	WT (S)	WT (A)	WT (A)	WT (M)	WT (S)	WT (L)	WT (I)	WT (G)	WT (A)	WT (R)	N526K	IIA
193	WT (D)	WT (S)	WT (A)	WT (A)	WT (M)	WT (S)	WT (L)	WT (I)	WT (G)	WT (A)	WT (R)	WT (N)	WT

Highlighted columns indicated that mutations in these positions were not considered in the *ftsI* grouping scheme.
[a] Isolates possessing these *ftsI* alleles were grouped as IIB because they had all the definitive point mutations associated with this group. However, the alleles had one extra point mutation (M→I) at 377 amino acid location.
[b] Isolates possessing these *ftsI* alleles were grouped as IIB because they had all the definitive point mutations associated with this group. However, the alleles had one extra point mutation (G→E) at 490 amino acid location.

Appendix 5.4. List of point mutations in the *gyrA*, *parC*, and *parE* in 2 *H. influenzae* isolates which are phenotypically resistant to ciprofloxacin and moxifloxacin.

Point mutations were defined using the *H. influenzae* reference genome Rd KW20 which was known to be sensitive to fluoroquinolone antibiotics. However, this reference genome had been noted to have many frameshifted coding sequences therefore results were also double checked using another complete genome, the strain FDAARGOS_1560 which is currently the NCBI reference genome. The Rd KW20 and FDAARGOS_1560 possessed identical alleles for the 3 genes encoding products targeted by the fluoroquinolone antibiotics.

HAEM1394 (*gyrA*)	HAEM1650 (*parC*)	HAEM1649 (*parE*)
• **S84L** • **D88N** • A407T • T416A • E433D • I472V • E505D • E872D • S879N • E881K	• K20R • **S84I** • N138S • I248T • E276D • D356A • E395K • N496S • S534N • L536R • T582A • H705Y • A723T • V745I	• S53T • T78I • S145N • G594E