MICROBIAL GENOMICS

RESEARCH ARTICLE

Krisna *et al.*, *Microbial Genomics* 2024;10:001281

DOI 10.1099/mgen.0.001281

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Development and implementation of a core genome multilocus sequence typing scheme for *Haemophilus influenzae*

Made Ananda Krisna[1,2,3], Keith A. Jolley[2], William Monteith[2,4], Alexandra Boubour[5,*], Raph L. Hamers[1,3], Angela B. Brueggemann[5], Odile B. Harrison[2,5,*] and Martin C. J. Maiden[2]

Abstract

Haemophilus influenzae is part of the human nasopharyngeal microbiota and a pathogen causing invasive disease. The extensive genetic diversity observed in *H. influenzae* necessitates discriminatory analytical approaches to evaluate its population structure. This study developed a core genome multilocus sequence typing (cgMLST) scheme for *H. influenzae* using pangenome analysis tools and validated the cgMLST scheme using datasets consisting of complete reference genomes (*N* = 14) and high-quality draft *H. influenzae* genomes (*N* = 2297). The draft genome dataset was divided into a development dataset (*N* = 921) and a validation dataset (*N* = 1376). The development dataset was used to identify potential core genes, and the validation dataset was used to refine the final core gene list to ensure the reliability of the proposed cgMLST scheme. Functional classifications were made for all the resulting core genes. Phylogenetic analyses were performed using both allelic profiles and nucleotide sequence alignments of the core genome to test congruence, as assessed by Spearman's correlation and ordinary least square linear regression tests. Preliminary analyses using the development dataset identified 1067 core genes, which were refined to 1037 with the validation dataset. More than 70% of core genes were predicted to encode proteins essential for metabolism or genetic information processing. Phylogenetic and statistical analyses indicated that the core genome allelic profile accurately represented phylogenetic relatedness among the isolates (R^2 = 0.945). We used this cgMLST scheme to define a high-resolution population structure for *H. influenzae*, which enhances the genomic analysis of this clinically relevant human pathogen.

Impact Statement

Discriminating *Haemophilus influenzae* variants and evaluating population structure has been challenging and largely unstandardized. To address this, we have developed a core genome multilocus sequence typing (cgMLST) scheme for *H. influenzae*. Since an accurate typing approach relies on a precise reflection of the underlying population structure, we explored various methods to define the scheme. The core genes included in this scheme were predicted to encode functions in essential biological pathways, such as metabolism and genetic information processing, and could be reliably assembled from short-read sequence data. Single-linkage clustering, based on core genome allelic profiles, showed high congruence to genealogy reconstructed by maximum-likelihood methods from the core genome nucleotide alignment. The cgMLST scheme v1 enables rapid and accurate

Received 22 April 2024; Accepted 18 July 2024; Published 09 August 2024

Author affiliations: [1]Nuffield Department of Medicine, Centre for Tropical Medicine and Global Health, University of Oxford, Oxford, UK; [2]Department of Biology, University of Oxford, Oxford, UK; [3]Oxford University Clinical Research Unit Indonesia, Faculty of Medicine Universitas Indonesia, Jakarta, Indonesia; [4]Department of Biology and Biochemistry, University of Bath, Bath, UK; [5]Nuffield Department of Population Health, University of Oxford, Oxford, UK.

***Correspondence:** Odile B. Harrison, odile.harrison@ndph.ox.ac.uk; Alexandra Boubour, alexandraboubour@gmail.com

Keywords: cgMLST; core genome; *Haemophilus influenzae*; population genetics; typing scheme.

Abbreviations: CC, clonal complex; CGC, Core genome cluster; cgMLST, core genome MLSTmultilocus sequence typing; cgST, core genome sequence type; COG, Cluster of Orthologous Genes; Hib, *H. influenzae* serotype b; Hif, *H. influenzae* type f; KEGG, Kyoto Encyclopaedia of Genes and Genomes; ML, maximum-likelihood; MST, minimum-spanning tree; NCBI, National Center for Biotechnology Information; NTHi, nontypeable *H. influenzae*; ODC, ornithine decarboxylase; OG, orthologous; PHI, pairwise homoplasy index; seq-def, sequence definition; ST, sequence type; WGS, whole-genome sequencing.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Six supplementary figures, six supplementary tables and three supplementary files are available with the online version of this article.

001281 © 2024 The Authors

This is an open-access article distributed under the terms of the Creative Commons Attribution License. This article was made open access via a Publish and Read agreement between the Microbiology Society and the corresponding author's institution.

depiction of high-resolution *H. influenzae* population structure, and making this scheme accessible via the PubMLST database ensures that microbiology reference laboratories and public health authorities worldwide can use it for genomic surveillance.

DATA SUMMARY

Assembled whole genome sequence data are available from PubMLST, https://pubmlst.org/organisms/haemophilus-influenzae, and are publicly accessible and downloadable with the list of isolate IDs provided in Supplementary File 1. The *H. influenzae* core genome multilocus sequence typing (cgMLST) scheme is also accessible through PubMLST.

The pipeline for cgMLST scheme development and validation is published at dx.doi.org/10.17504/protocols.io.4r3l22o64l1y/v1. All in-house R and Python scripts for data processing and analysis are available from https://github.com/artmisk13/cgmlst_hinf.

A minimum-spanning tree (MST) based on the cgMLST scheme is accessible via Microreact:https://microreact.org/project/aBH3zddMifvweC2KJNEnN4-phylogenycgccgmlstproject

INTRODUCTION

Haemophilus influenzae is a fastidious Gram-negative commensal coccobacillus, which is also an accidental pathogen in humans. It is classified based on the expression of capsular polysaccharides, categorized as six serotypes, a–f, plus the unencapsulated or nontypeable [nontypeable *H. influenzae* (NTHi)], which do not express a capsule. *H. influenzae* exclusively inhabits the human host, primarily the upper respiratory tract, where it resides as a commensal member of the microbiota; however, infections caused by *H. influenzae* can manifest as non-invasive or invasive diseases, including otitis media, sinusitis, meningitis, epiglottitis, orbital cellulitis and septicaemia [1, 2]. Prior to the implementation of the *H. influenzae* serotype b (Hib) polysaccharide-conjugate vaccine, Hib was the predominant causative agent of *H. influenzae* invasive diseases worldwide. Currently, more than 70% of cases can be attributed to NTHi, which is also the most prevalent *H. influenzae* group [3, 4]. The incidence of these diseases has consistently increased across all age groups in recent years [3, 5, 6]. The circulating invasive NTHi has been reported to be increasingly resistant to multiple antibiotic groups such as beta-lactams, cephalosporins, fluoroquinolones and macrolides [7–9]. Additionally, vaccines for non-Hib *H. influenzae* are currently unavailable, with no vaccine developments in the past 5 years [10–13]. Although surveillance programmes exist [14], with most centres in high-income countries incorporating whole-genome sequencing (WGS), there is a lack of standardized, high-resolution methods and nomenclature to describe important lineages.

The molecular classification of human pathogens is an integral component of microbiological diagnostics and surveillance. Among the most widely employed methodologies is multilocus sequence typing (MLST), which is based on the genetic variability within a set of six to eight housekeeping gene fragments to categorize bacterial isolates into distinct sequence types (STs). These STs, in turn, can be organized into clonal complexes (CCs) based on similarities in their allelic profiles [15, 16]. Core genome MLST (cgMLST) extends the MLST framework to hundreds of complete core genes (i.e. genes shared among all or nearly all isolates of a species) to assess genomic variation with greater precision [17, 18]. Like conventional MLST, cgMLST treats all allelic changes as a single event, as a mitigation of the effects of recombination on tree branch lengths [19]. While MLST (STs and CCs) is still widely used to this date for the typing of multiple bacterial species when WGS is not possible, cgMLST improves accuracy when delineating lineages and dilutes the recombination effect [20]. Because of this, previous reports have concluded that it is the preferred method for assessing highly recombining organisms, including *H. influenzae* [21, 22]. *H. influenzae*, particularly NTHi, exhibits substantial genomic diversity primarily attributed to horizontal gene transfer through transformation and recombination [23].

Previous investigations into the core genome of *H. influenzae* have yielded insights, with one study culminating in the development of a cgMLST scheme; however, these studies typically included fewer than 500 draft genomes, which has been identified as a threshold for stable paralogous locus detection and a key step in core genome analyses [24–26]. Paralogs are duplicated genes derived from a single gene, and their presence can mislead genetic relationships among genomes [27, 28]. Furthermore, neither of the existing cgMLST schemes has been validated or implemented on publicly accessible platforms. Previous research has identified various *in silico* parameters affecting cgMLST precision, the most important of which was the completeness of a cgMLST profile [29]. The BIGSdb software that hosts the PubMLST website periodically scans for new alleles within defined schemes, such as cgMLST. These alleles are stored in the sequence definition (seq-def) database and assigned allele numbers, with the resulting cgMLST profile defined in the same database, ensuring the fulfilment of completeness criteria [30]. At the time of writing, the PubMLST website had implemented cgMLST schemes for various bacterial species, including *Neisseria meningitidis* [31], *Neisseria gonorrhoeae* [32], *Campylobacter* sp [33], *Streptococcus agalactiae*, *S. pneumoniae* [34], *S. uberis* [35], *Vibrio cholera* [36], *V. parahaemolyticus* [37], *Bacillus anthracis* [38], *Bacillus cereus* [39], *Burkholderia mallei* [40], *Acinetobacter baumannii* [41] and *Clostridium perfringens* [42]. This illustrated the importance of a high-resolution scheme for pathogens of public health importance.

The present study developed and validated a cgMLST scheme for *H. influenzae*, which was then implemented in the PubMLST database. The scheme is publicly available and can be utilized by public health authorities worldwide to characterize *H. influenzae* genomes in a more standardized manner and with greater detail. It also serves as a tool to improve our understanding of the population biology of the bacterium, particularly the NTHi, as this approach is minimally biased by recombination events frequently occurring in this population. In turn, this can be used to aid in vaccine development and contextualize antimicrobial resistance spread.

METHODS

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 analysis (May 2023), including Roary [43], PIRATE [44], PanX [45], PGAP [46], PPanGGOLiN [47], MetaPGN [48], PEPPAN [49], chewBBACA [26] and Panaroo [50]. Tools without any updates for 5 years or more were not used for this analysis. Roary and PPanGGOLiN were excluded because, based on a recent report, the accessory genome size was likely to be inflated using these programmes, largely due to paralogs over-splitting. In addition, this could also lead to an underestimation of the number of core genes [50]. Therefore, four software packages, PIRATE, PEPPAN, chewBBACA and Panaroo, were employed. A 'two-step' approach was implemented to evaluate the optimal pipeline for *H. influenzae* core gene identification:

(1) Assessment of the methods and unique features provided by each software (Table S1, available in the online version of this article) and
(2) Pangenome analysis of a set of reference genomes using each of the software packages and comparison of the outputs.

The complete workflow and decision-making that followed for each step are published in dx.doi.org/10.17504/protocols.io.4r3l22o64l1y/v1.

Development and validation datasets of high-quality draft genomes

The dataset for this study was compiled from draft genomes stored in the PubMLST database, accessed on 24 September 2022. All publicly available *H. influenzae* genomes in the database underwent quality checks based on several parameters (Table S2) [51]. Genomes not fulfilling these criteria were individually reviewed for inclusion in the dataset. After the review, 2397 *H. influenzae* isolates were chosen, and their provenance data were obtained.

Genomes were systematically allocated into two independent datasets: the development dataset (N = 986) and the validation dataset (N = 1411). The development dataset consisted of well-characterized isolate records with good provenance information. This provided reliability for implementation in PubMLST, particularly given the automated annotation process. The validation dataset consisted of isolate records with less detailed provenance information and allowed proposed workflows to be tested. Provenance information of importance included the capsule genotype, isolate source, disease and region of isolation (Fig. S1).

Core gene identification and curation

Draft genome assemblies in both datasets were annotated for protein-coding genes using Prokka [command: prokka --outdir /output/directory/ --cpus (number of CPU) --compliant --prefix (Genome ID) /path/to/FASTA/files] [52]. For the purpose of scheme generation, core genes were defined as genes present in at least 95% of isolates to account for the draft status of the genomes and to cater for genes that may occasionally be missing in the originating isolates [17, 20, 31]. After finalizing the computational pipeline, PIRATE [command: PIRATE -I /path/to/Prokka/GFF/files/ -o /output/directory/ -a -r -t (number of CPU)] was mainly used for core gene identification. The command above ran BLASTP over a range of amino acid similarity thresholds (50, 60, 70, 80, 90, 95 and 98%). PIRATE clusters genes into 'gene families' and gene families present in 95% of isolates at only a single copy, hence excluding all putative paralogs, were included in the provisional list of core genes. Subsequently, we verified whether (1) these provisional core genes were also identified as 'core' by Panaroo [command: panaroo -I /path/to/Prokka/GFF/files/ -o /output/directory/ -t (number of CPU) --clean-mode strict -a core --search_radius 1000 --refind_prop_match 0.7], which incorporates additional stages for annotation error correction, and (2) any of these core genes were identified as paralogs by either PEPPAN [command 1: PEPPAN --match_identity 0.7 --orthology ml /path/to/Prokka/GFF/files/ -t (number of CPU); command 2: PEPPAN_parser -g /path/to/PEPPAN/output/PEPPAN.PEPPAN.gff -s /output/directory/ -t -c -a 95] or Panaroo to increase the sensitivity of paralog detection and their exclusion from the final cgMLST scheme (Fig. 1). Complete exclusion of any putative paralogs allows the automated annotation of the core genomes for *H. influenzae* genomes deposited in PubMLST, which is one of the goals of this scheme. Results from the three programmes were consolidated by utilizing an in-house Python script, available at https://github.com/artmisk13/cgmlst_hinf.

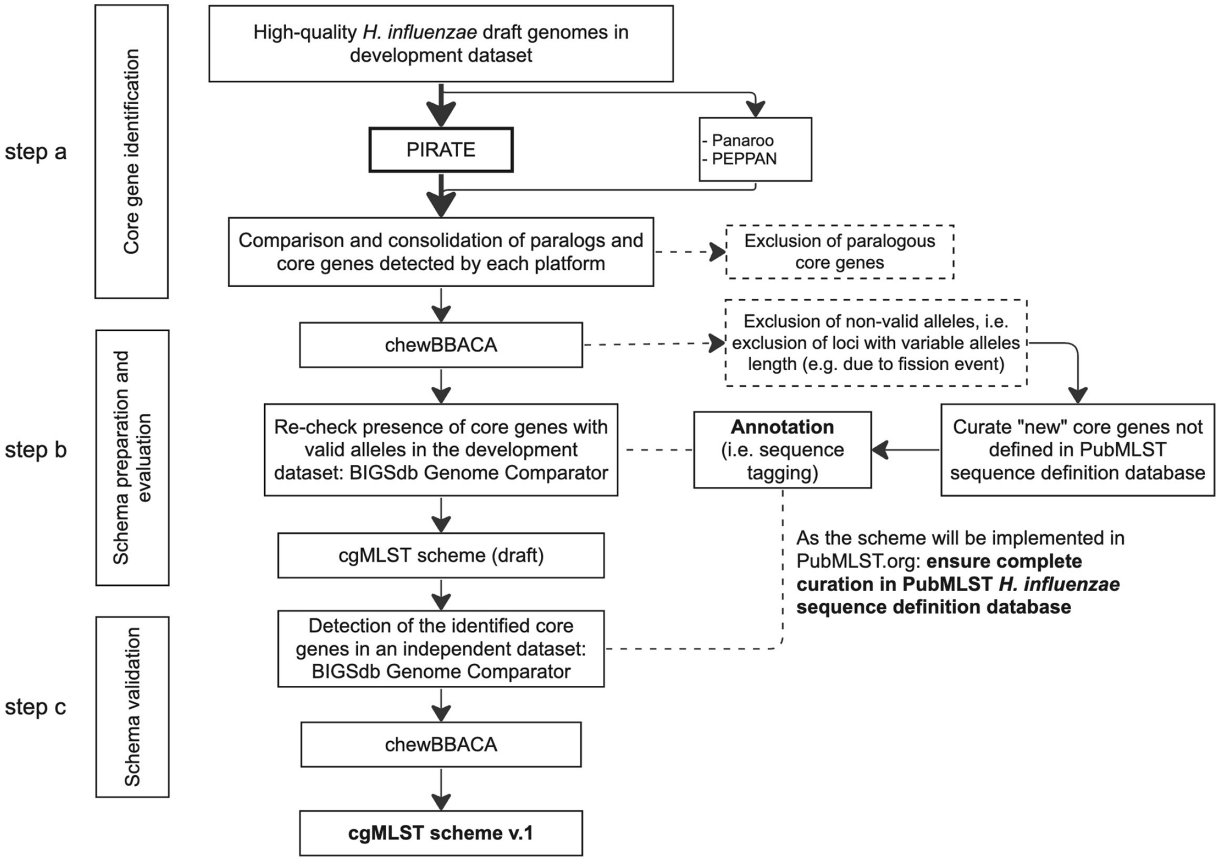

Fig. 1. The workflow of cgMLST scheme development and validation.

Each core gene identified in step a (Fig. 1) was evaluated for the presence of invalid alleles using chewBBACA with default parameters (Fig. 1, step b). Invalid alleles were defined either as: (i) sequences with ambiguous characters; (ii) a length not divisible by 3; (iii) the presence of in-frame stop codon(s) and (iv) sequences that could not be translated due to missing start/stop codons [26]. Valid core genes were subsequently cross-referenced with the PubMLST database to identify genes that had not been previously defined in the database and allow these to be curated in the database. To automatically detect new alleles, the BIGSdb 'sequence tagging' function was utilised (90% minimum identity; 70% minimum alignment; and a BLASTN word size of 20) [53]. All genes in the PubMLST *H. influenzae* database were assigned a unique locus name starting with 'HAEM' for '*Haemophilus*', followed by an arbitrary number. Each HAEM locus can be associated with the common gene name, e.g. *bexA* gene, which encodes a protein involved in capsule transport, assigned as 'HAEM1156' in the PubMLST database.

Validation analyses

The initial cgMLST scheme was applied to the validation dataset to confirm that the core genes were still defined as core using a different dataset. This was implemented using the BIGSdb Genome Comparator tool available in PubMLST (default parameters). All genes that remained core after this step would constitute the final *H. influenzae* cgMLST scheme. Additional analyses were performed on these core genes, including allelic variability, functional classification and intragenic recombination analysis.

To assess the allelic variability of core genes, the total number of alleles and their length were calculated using an in-house Python script (https://github.com/artmisk13/cgmlst_hinf). Functional classification of the core genes was completed using eggnog-mapper (version 2.1.11), Diamond in blastx mode and HMMER methods with default parameters [54]. Each gene was assigned a Cluster of Orthologous Genes (COG) category [55]. The COG category was further grouped based on the Kyoto Encyclopaedia of Genes and Genomes (KEGG) BRITE functional hierarchy system [56]. Lastly, intragenic recombination analysis was conducted using the pairwise homoplasy index (PHI) method with PhiPack software. PHI is a measure of recombination by assessing the incompatibility between each site and its downstream sites within a global alignment. To evaluate its statistical

significance, it calculates a *P*-value by comparing the observed PHI statistic to a normal distribution derived from the expected mean and variance [57, 58].

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 [59, 60]. 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**. Every genome in the dataset received a core genome sequence type (cgST) based on the allelic profile of the 1037 core genes contained within the scheme. For a cgST to be assigned, a maximum of 25 genes can be missing. This threshold was set because the scheme was applied to draft genomes, which may lack complete annotation.
   All genomes with a cgST assigned were further grouped into CGCs using a single-linkage clustering method with the pre-determined similarity thresholds at 500, 200 and 50 allelic mismatches. Each member of the CGC group has a maximum of the specified allelic differences with at least one other member of the group.
(2) **Capsule type based on genome sequences** (i.e. capsule genotype) as assigned by the Hicap suite software [61].
(3) **CCs** were defined using the globally optimized eBURST algorithm implemented in PHYLOViZ 2.0.
(4) **Pathotype clade classification system for NTHi**. Previous studies on NTHi population structure revealed six distinct clades (clade I–VI), based on the presence or absence of 17 accessory genes (Table S3), which was reported to reflect the phylogeny based on the core genome single nucleotide polymorphisms (SNPs) [25, 62, 63]. The presence of these genes was concluded by conducting a BLAST search of the genomes in the dataset with the following parameters: 90% identity, 70% alignment and BLASTN word 20.
(5) **Biotype** based on the presence or absence of genes (Table S4) encoding three metabolic enzymes: ornithine decarboxylase (ODC), urease, and tryptophanase [64, 65].

Additionally, a minimum-spanning tree (MST) based on the core genome allelic profile was constructed, utilizing the GrapeTree plugin on the PubMLST isolate database [53, 66].

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

Genetic relatedness among *H. influenzae* isolates was assessed based on their core genome by two approaches: either using the core genome allelic profile or the nucleotide sequence identity. To evaluate the congruence between these approaches, scatter plots were generated, incorporating a correlation test and simple linear regression analysis with ordinary least square (OLS) method. Distance matrices based on the cgMLST scheme pairwise allelic mismatch and core genome alignment ML tree branch length were produced using the Genome Comparator plug-in on the PubMLST website and an in-house Python script (https://github.com/artmisk13/cgmlst_hinf). Matrices were converted into a frequency table, and a scatter plot was generated. If at least one group showed a non-Gaussian distribution, the Pearson correlation test was employed for calculating a *P*-value. Lastly, regression analysis was conducted to obtain a mathematical function explaining the relationship between the two methods, as well as the coefficient of determination (R^2) which measures the goodness-of-fit of the function. These steps were conducted with R statistical software and Python statistics packages. All scripts are available at https://github.com/artmisk13/cgmlst_hinf.

RESULTS

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

The reference genome dataset comprised 14 complete reference *H. influenzae* genomes, 11 of which were listed in the KEGG Organisms: Complete Genomes database (https://www.genome.jp/kegg/catalog/org_list.html accessed 26 September 2022). Three additional clinically important complete genomes were included after a publication search through the National Center for Biotechnology Information (NCBI) PubMed [67, 68]. All associated provenance data and complete genome sequences are available from PubMLST (Table S5).

There were 986 draft genomes in the development dataset that fulfilled the quality check criteria; however, this number was reduced to 921 genomes after initial pangenome analysis with PIRATE. The 65 excluded genomes exhibited high duplication events and corresponded to the outliers in the box plot for genome length (Fig. S2) of the original development dataset, and likely represent sequencing errors or lower assembly quality. Therefore, the pangenome analysis was rerun using 921 draft genomes in the development dataset, which identified 1063 core genes for the first draft of the cgMLST scheme. Subsequently, genomes in the validation dataset ($N = 1411$) were annotated for these core genes, and genomes with <95% of the core genes annotated were excluded ($N = 35$), resulting in the final validation dataset of 1376 genomes. There was no notable difference in the distribution of other variables between the initial and final datasets (Supplementary File 1).

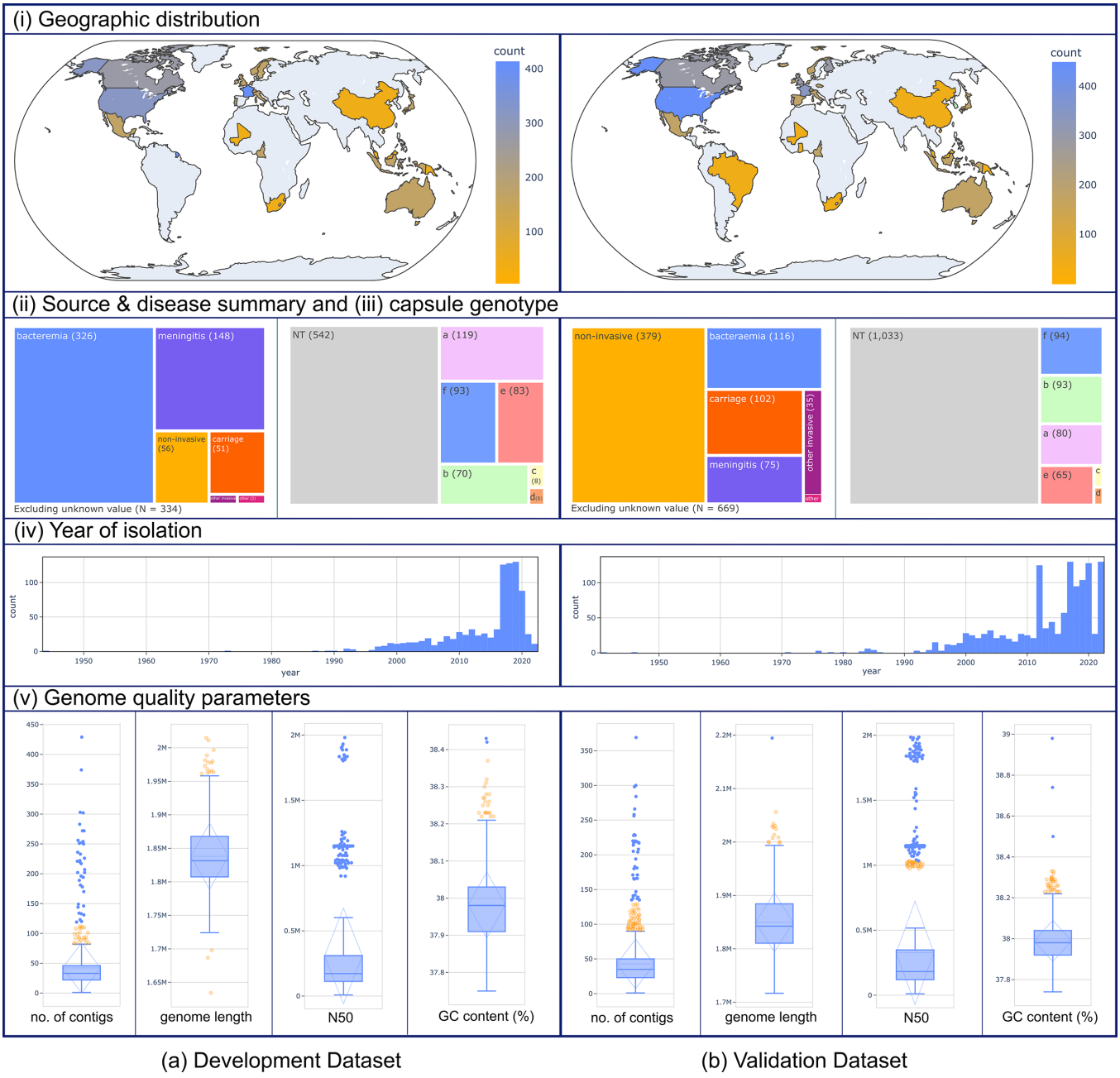

Fig. 2. Characteristics of the datasets and genomes employed for developing the *H. influenzae* cgMLST scheme: (a) development (*N* = 921) and (b) validation (*N* = 1376).

H. influenzae isolates in both datasets were globally distributed; however, most isolates originated from North America (356/921 and 548/1376) and Europe (431/921 and 547/1376). The two datasets also showed a similar temporal distribution, with the most isolated after 2016. There were more clinical isolates from invasive disease cases (i.e. bacteraemia, meningitis and other invasive) in the development dataset, which was reflected in a similar proportion of non-typeable (542/921) and typeable (379/986) in the dataset. This was performed to ensure that the developed schema would be applicable to all typeable *H. influenzae* as well as the NTHi (Fig. 2 (i)–(iv)). Additionally, 242 (61%) unique STs and 7 (11%) unique CCs found in the validation dataset were not present in the development dataset (Supplementary File 1), indicating the larger diversity of *H. influenzae* that has been captured by the inclusion of the validation dataset. Based on several metrics, genomes in both datasets were of good quality, with the majority (893/921 and 1346/1376) having less than 150 contigs and an N50 parameter of more than 23 000.

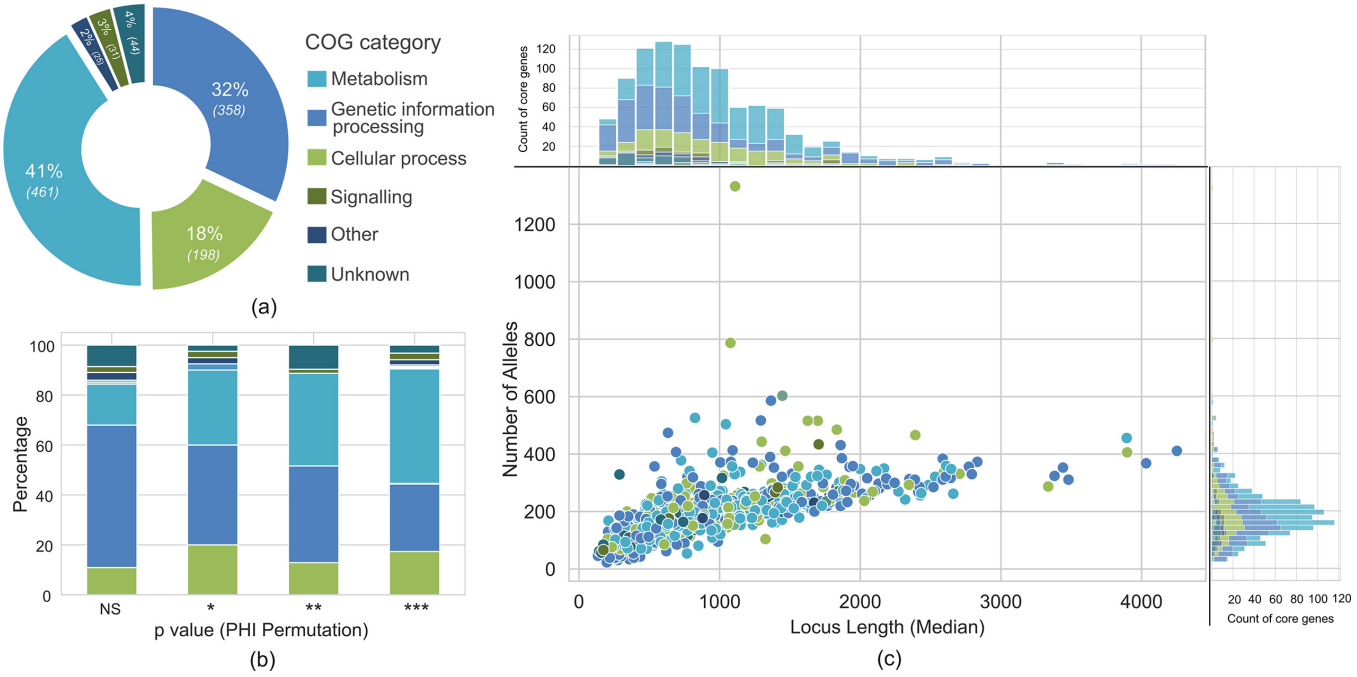

Fig. 3. Functional classification, recombination and allele variability analysis of *H. influenzae* core genes in the cgMLST scheme. (a) Functional classification was achieved with eggnog-mapper, assigning the COG category for each core gene. (b) Intragenic recombination analysis of each core gene based 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.

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

A total of 1392 core genes were identified in the combined pangenome analysis with the three tools, PIRATE, Panaroo and PEPPAN (Fig. 1, step a), with 144 paralogous core genes excluded at this stage. Only 18 out of the 1248 non-paralogous core genes were not previously defined in the PubMLST *H. influenzae* seq-def database and were therefore added. After filtering for invalid alleles (Fig. 1, step b), a total of 185 genes were no longer defined as core, resulting in 1063 genes included in the cgMLST scheme draft (Supplementary File 2). Lastly, 1037 genes remained core after validation with the validation dataset (Fig. 1, step c), excluding only 26 genes, 25 of which were present in 90–95% of the genomes (Supplementary File 2). The draft genomes in this dataset were successfully annotated for those core genes using the automated annotation process of the PubMLST database.

A total of 1024 core genes in the scheme were assigned an NCBI COG category, 37 of which had more than one COG category defined [55]. Among the core genes with one COG, 149 (14%) had 'function unknown', or denoted as 'S', as the COG category. After re-assessment, 95 had at least one specific, non-S COG based on the orthologous (OG) hit from any higher taxonomy levels. The remaining genes were curated manually using the KEGG orthology classification system or protein product-based search on the Protein Families Database (v 95.0) or both (Supplementary File 3). For the 13 genes without an initial COG category, re-searching using the HMMER method resulted in identical results. These were categorized as having an unknown function and assigned the 'S' COG category (Fig. 3a).

Nearly 90% (909/1037) of the genes within the cgMLST scheme exhibited evidence of intragenic recombination events, as indicated by a *P*-value of <0.05 in PHI statistics. When assessing the percentage of recombination events in each COG category, the majority (73/128) of core genes with non-significant PHI statistics belonged to the genetic information processing group (Fig. 3b). Conversely, nearly all core genes in the metabolism group (405/461) displayed a *P*-value of <0.05. In addition, while core genes with significant PHI statistics exhibited a range of allele counts and median lengths, those with non-significant PHI statistics were shorter and had lower allele counts (Fig. S3).

The number of alleles and median length of each core gene were correlated (Fig. 3c), although with a shallow slope. Furthermore, 65% (675/1037) of the core genes had a median locus length of <1000 nucleotides, and 95% (986/1037) were <2000 nucleotides

long. The former had a median allele count of 151, and the latter was 175. There were short core genes with as much diversity as the long ones, the most extreme example being HAEM0191 (hypothetical protein) and HAEM1295 (outer membrane protein P5), which exhibited higher allele counts compared to other genes of similar median locus length (1332 and 787 alleles, respectively).

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

Within the validation dataset encompassing 1376 genomes, 1320 (95.9%) had a cgST assigned and thus could be classified into CGC groups at different similarity thresholds (Supplementary File 1). The rest of the genomes without cgST still had an assigned allelic profile for a substantial number of their core genes, ranging from 989 to 1011 (95.4–97.5%), out of the 1037 core genes in the scheme.

The population structure of *H. influenzae* can be systematically explored using hierarchical core genome clustering at different thresholds. This hierarchical classification, derived from the allelic profiles of the core genome, effectively mirrors the phylogenetic relationships among the isolates when annotated to the ML tree (Fig. 4). Additionally, the clustering pattern observed on the phylogeny was replicated on the MST, which was generated solely based on the allelic difference of the core genome (Fig. 5).

Typeable *H. influenzae* isolates sharing the same capsule genotype formed clusters in the phylogeny, consistent with their close genetic relatedness. Several CGC groups, spanning various allelic mismatch thresholds, aligned with specific capsule genotypes (Fig. 4). For instance, among *H. influenzae* type a (Hia) isolates, four distinct lineages were delineated, each with its own CGC group at 500 allelic mismatches (CGC500) – namely, CGC500 48, 46, 79 and 81 – which correlated with the CC assignments (Fig. 5 and Table S6). The largest Hia lineage, CGC500 48, which corresponds to the ST-23 complex, harboured a significant subgroup, mainly comprised of isolates originating from Canada. This subgroup was clustered together at 50 allelic mismatch thresholds (Fig. 4 yellow highlight and subpanel i). A similar pattern emerged within the Hib in which there were three distinct lineages in the ML tree and MST (Figs 4 and 5), two of which correlated with the CGC group at 500 allelic mismatches and their respective CCs (Table S6). The least represented lineage consisted of a single isolate (isolate ID 5105) of ST 464, and this isolate did not have CC or CGC groups assigned (Figs 4 and 5, black arrow). At the time of writing (October 2023), there is only one other genome in the PubMLST database with this ST (isolate ID 15982). This isolate did not cluster with the rest of Hib and was more closely related to Hia when the allelic profile was evaluated.

The NTHi isolates were much more diverse, with 110 CGC 500 groups, only 8 of which contained at least 30 isolates in the group (Fig. S4). Nevertheless, a correlation between CGC grouping and the *H. influenzae* phylogeny was also evident among NTHi. To exemplify this, a subset of the ML tree incorporating 97 NTHi within the CGC500 23 group was highlighted (Fig. 4 grey highlight and subpanel (ii) and Fig. 5). Within this group, there were 7 clusters at the 200 allelic mismatches threshold (CGC200), and this clustering was highly congruent with the tree topology. However, while typeable *H. influenzae* isolates sharing the same CC exhibited clustering congruence in the phylogeny, this was not observed in the case of NTHi. For example, the ST-3 complex was distributed throughout the tree (Figs 4 and 5 grey nodes with red lines). This pattern was observed in 19 out of 57 CCs within the NTHi population in the validation dataset (Table S6). Conversely, the NTHi CC assignments did not consistently align with the CGC500 grouping. For instance, within the CGC500 23 group mentioned earlier, three different CCs were identified (Table S6). Therefore, the CGC groups offered a more precise representation of the NTHi phylogenetic structure (Fig. 5).

Although, in general, there were specific CCs for each *H. influenzae* serotype, this was not always the case. Within the ST-124, ST-210 and ST-422 complexes, there were NTHi and *H. influenzae* type f (Hif) isolates. Conversely, the CGC500 group assignment for these isolates was consistent with both the serotype and phylogeny, also showing the expected advantage of cgMLST compared to 7-loci MLST. Taking this into account, a few NTHi isolates ($N = 5$) were observed within the encapsulated clusters, with the CGC500 group assignment identical to the encapsulated isolates (Fig. 4 arrow heads and Table S6). This marked the possible rare event of capsule loss in previously encapsulated *H. influenzae*.

The ability of the core genome allelic profile within the scheme to depict the phylogenetic relationships in *H. influenzae* was quantitatively assessed (Fig. 6). This assessment was performed by comparing the number of allelic mismatches with the branch length value from the ML tree for each pair of genomes in the validation dataset. The coefficient of determination (R^2) of 0.945 indicates that 0.945 of the variance in branch lengths, in the context of an ML tree unaffected by recombination events, can be elucidated through the allelic pairwise mismatches observed across the 1037 core genes. A Spearman correlation test also yielded a statistically significant P value of <0.001. On the left side of the plot, the larger deviation of data points from the regression line indicates that when two genomes are highly similar (i.e. differing by only a few nucleotides), the dissimilarity between them might not be accurately captured by the number of core genome allelic mismatches.

Lastly, we assessed two previously utilised classification systems aimed at characterizing the population structure of *H. influenzae* to determine the extent of their concordance with both the cgMLST scheme and the overall phylogeny of this bacterial species. First, the pathotype clade classification system, specifically applicable to NTHi, was concluded and annotated in the ML tree of genomes in the validation dataset (Fig. S4). Out of 1008 NTHi isolates, 160 could not be assigned to any of the six specified clades.

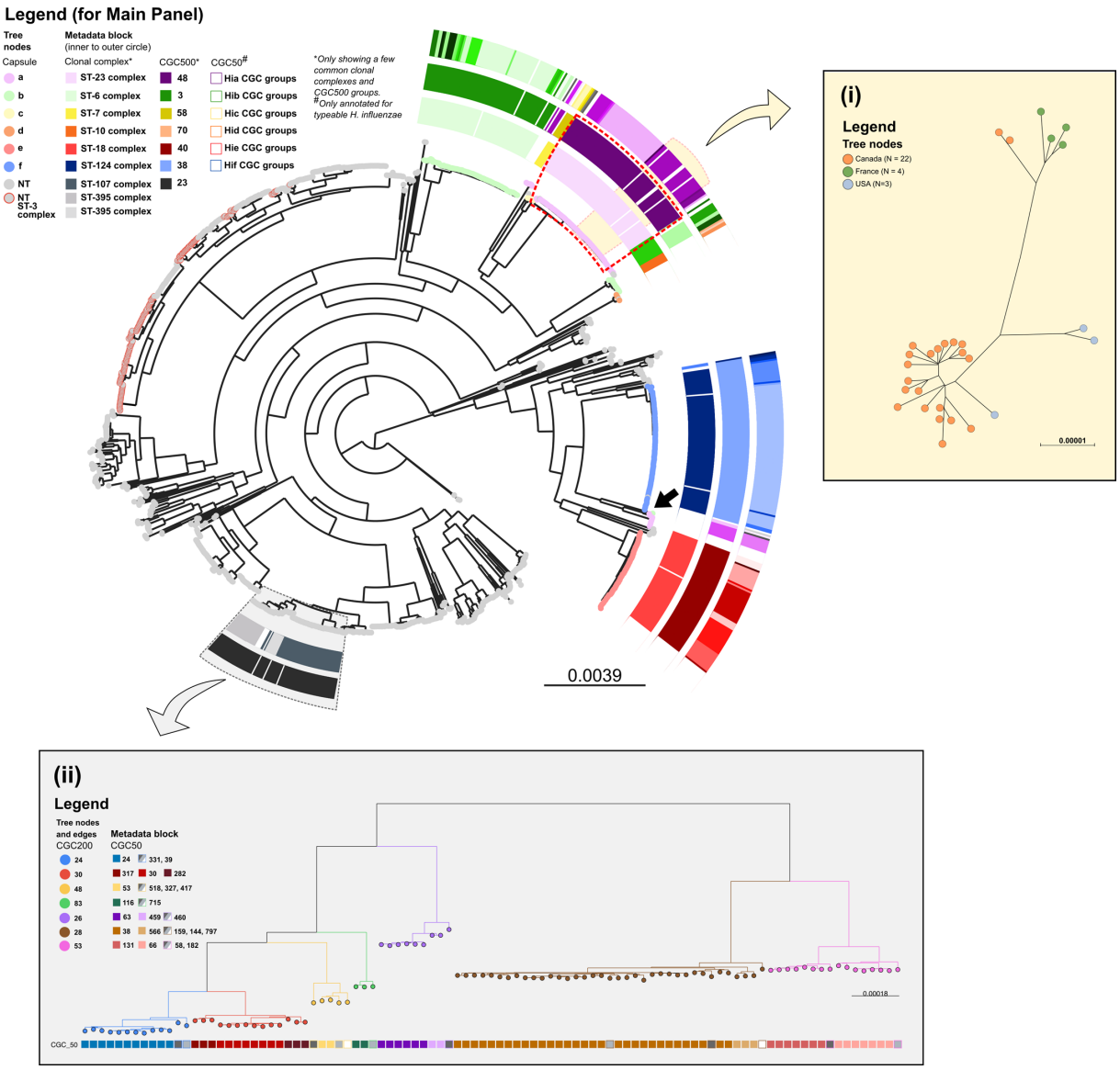

Fig. 4. Population structure of 1,376 *H. influenzae* genomes from the validation dataset. Main panel: a ML phylogenetic tree generated from a concatenated core gene nucleotide sequence alignment. Tree nodes were coloured by capsule type. Each capsule type clustered together, except for *H. influenzae* type a (Hia) and 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*. [74]. 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.

Additionally, although the clade classification system followed the phylogeny topology to some extent, within each identified clade, the genetic relatedness among genomes varied greatly. For example, the median pairwise allelic mismatch in clade VI was 991 (minimum–maximum = 0–1028) and in clade II the median was 640 (minimum–maximum = 0–766) (Fig. S5). Second, we examined the biotype classification system, which relied on the distinct production of three specific enzymes, each directly corresponding to the presence of a particular gene. None of the eight biotypes demonstrated a clear correspondence with specific CGC groups at any threshold, the underlying phylogenetic structure, or any discernible clinical or demographic characteristics (Fig. S6).

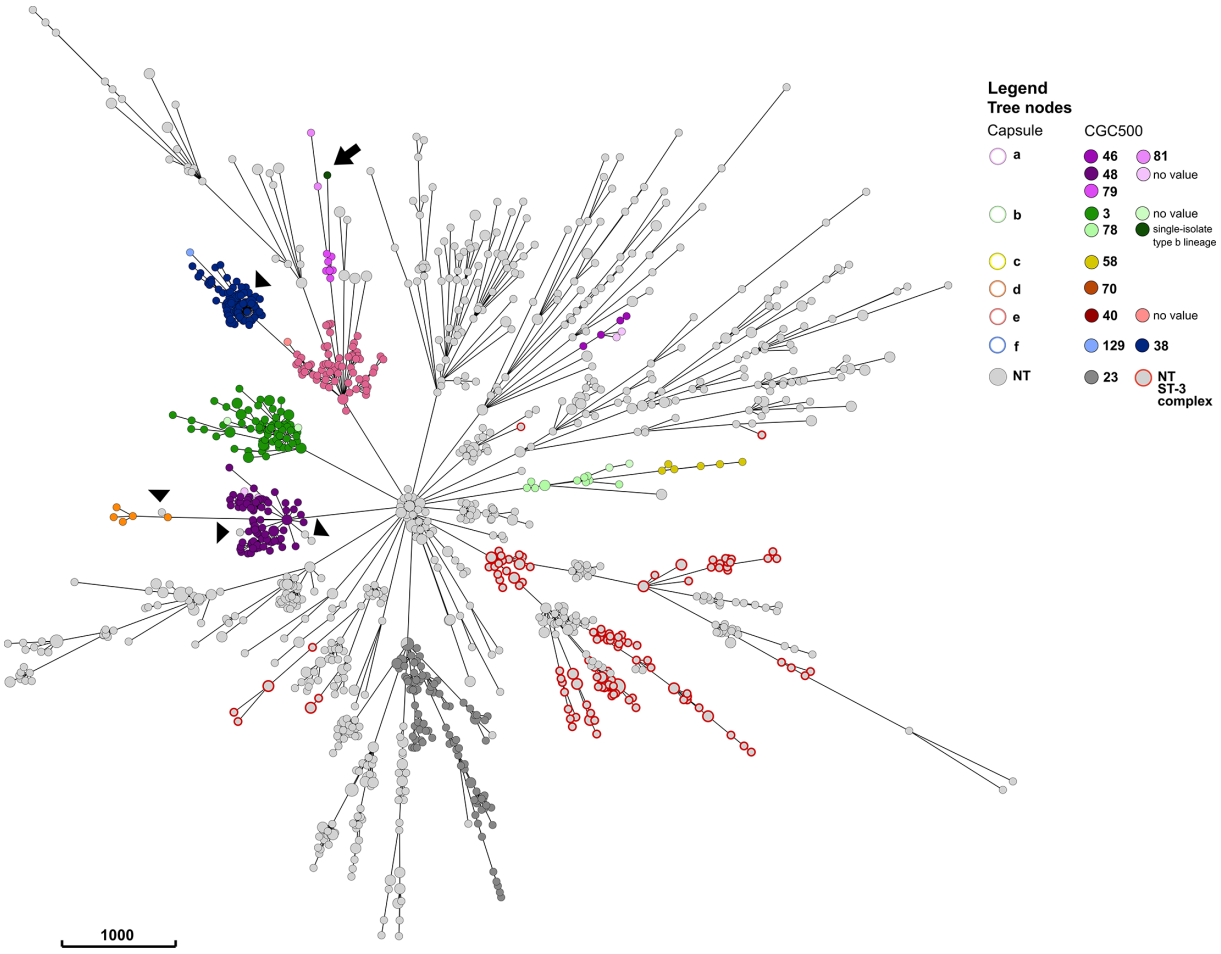

Fig. 5. MST of the validation dataset (*N* = 1376 genomes) constructed from the core genome allelic profile. Tree nodes were coloured by capsule type and CGC group at the 500 allelic mismatches threshold (CGC500), using the same colour scheme as Fig. 4. The pattern observed in the phylogeny was replicated 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 was also replicated (grey with an outline). Additionally, 5 NTHi isolates found within the encapsulated clusters were shown in greater detail here (black arrow heads).

DISCUSSION

At the time of writing, there were over 560000 prokaryote genome assemblies in the NCBI Genome Library (https://www.ncbi.nlm.nih.gov/genome/microbes/). The accessibility and relative affordability of high-throughput sequencing technologies have resulted in an unprecedented amount of available bacterial genome data [69]. cgMLST is a gene-by-gene typing approach able to assess genomic variation within a bacterial species or genus by utilizing WGS data. The gene-by-gene approach has several advantages compared to other methods for population genetic evaluation. Firstly, it does not require a reference genome, and secondly, it focuses on functional protein-coding genes. More importantly, this approach treats any allelic changes as single events, a feature useful for highly recombining organisms such as *H. influenzae*. With 1037 core genes included in the cgMLST scheme, the scheme can be used to generate a high-resolution population structure largely independent of recombination events.

There are variations in the reported number of *H. influenzae* core genes in the published literature. Hoggs *et al.* and Eutsey *et al.* found a range of 1450–1485 genes shared by 100% of genomes, commonly referred to as 'hard core' genes [70, 71]. In 2019, Pinto *et al.* examined the pan-genome of over 200 NTHi genomes, identifying 1400 genes shared by at least 95% of the genomes [25]. Rajendra KC *et al.* conducted a similar pangenome analysis in 568 NTHi genomes but found only 853 core genes from the total of 12 249 pan-genes [72]. Gonzalez-Diaz and colleagues revealed a spectrum of 1470–1627 core genes for each *H. influenzae* capsule type, with a total of 10–234 genomes in each capsule group. However, when combined (*N* = 800 genomes), the core genes shared by all capsulated *H. influenzae* was 1037 [73]. Several factors are known to cause variations in pangenome analyses. Firstly, core genome size estimation decreases as more genomes are integrated into the analysis until it eventually reaches an asymptote [70]. Secondly, different approaches and thresholds, used for clustering sequences into the same OG group, can either over- or

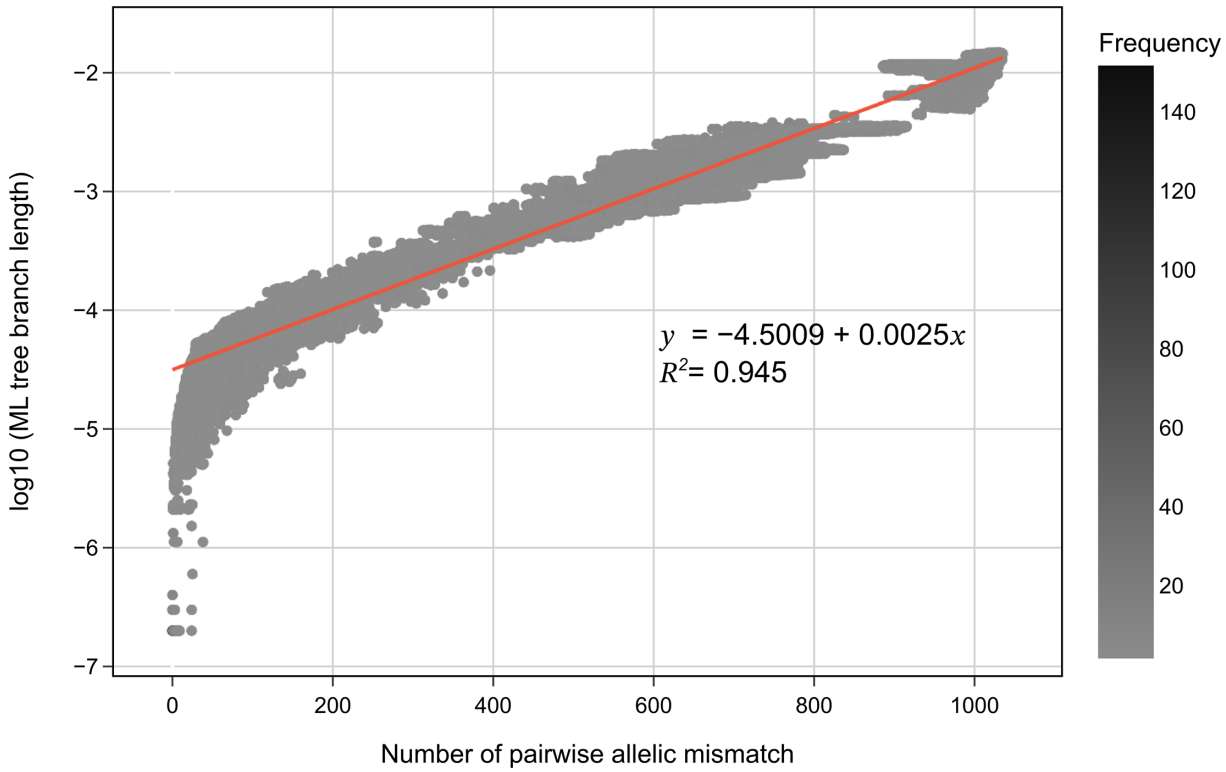

Fig. 6. Comparison of pairwise allelic mismatch of cgMLST core genes with the log10 branch length values from the ML tree, implemented in the validation dataset (*N* = 1,376). The more closely related genomes in a pair, the lower the allelic mismatch and log10 branch length value found. The log10 branch 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 branch length.

under-estimate the core genome size. In turn, these are influenced by the quality of genome sequencing and assembly [49, 50]. Lastly, putative paralogous genes are not automatically excluded from the list of core genes estimated using pangenome analysis tools [26–28]. In developing the cgMLST scheme, we incorporated steps to control for these factors by excluding low-quality genomes based on a set of predetermined criteria and utilizing different pangenome analysis tools, particularly to increase the sensitivity of detecting potential paralogous genes. These are all in line with our main aim for the scheme to be used as a robust typing method that reflects the genealogical relationship of *H. influenzae*.

The *H. influenzae* genome clustering based on cgST showed congruence with the genealogy reconstructed using the gold standard ML methods from the core genome nucleotide alignment. Based on the clustering of 500 allelic mismatches, we identified four Hia lineages as described by Topaz *et al.* [74]. They also reported a sublineage mostly isolated from North America, a finding replicated in our study when using the cgST clustering at 50 allelic mismatches [74]. Another report from the U.S. surveillance programme discovered two main Hib lineages, also delineated in two different CGC500 groups [75]. However, we identified one Hib of ST 464 (PubMLST id: 5105) located away from the two lineages, indicating a third Hib cluster. In addition, two major Hif lineages clustered at 500 allelic mismatches. A study by Gonzalez-Diaz *et al.* reported only one of these lineages and found three arbitrary clades, which were not discovered in our study [73]. This difference likely stemmed from variations in approaches used to describe Hif population structure (all *H. influenzae* vs. Hif-specific core genome) and in the methodology for generating the phylogeny, i.e. core genome nucleotide alignment vs. core SNP using a reference genome [73]. Generally, the cgST clustering at 500 allelic mismatches corresponded to CC assignment, with some exceptions. Some discrepancy was expected as there are 150 times more genes in the cgMLST scheme compared to the seven-locus MLST, resulting in higher resolution for depicting the population structure. However, the overall correlation between the CC and CGC500 groups showed that 7-gene MLST remains useful for describing a broader level of genetic relatedness of *H. influenzae* isolates when WGS is not possible. Accordingly, thresholds for CCs should still be defined based on the 7-locus MLST.

The five NTHi isolates located within the encapsulated clusters in the ML tree, with their CGC500 group assignment identical to the encapsulated isolates, may signify rare capsule loss events. In 2019, the first case report of a Hib isolate from an invasive case exhibiting loss of capsule expression was reported [76]. Although Potts *et al.* observed a similar finding in the same year through

a population genetic study of the US *H. influenzae* collected from surveillance programmes [75], the implementation of cgMLST and the subsequent cgST clustering will circumvent the need to construct phylogeny in order to observe such phenomena.

Finally, our cgMLST scheme was compared with two existing classification systems for *H. influenzae*. The first and oldest, 'biotyping', was utilized prior to the development of molecular techniques in microbiology and relied on the differential production of tryptophanase (indole test), urease and ODC [77]. Slotved *et al.* demonstrated that the production of these enzymes can be inferred from the molecular detection of the genes encoding them, although their presence/absence did not reflect the phylogeny, which we also demonstrated here [65]. The second was proposed more recently and focused on the NTHi, the NTHi clade typing. This typing system was developed based on core genome SNPs, which were defined as 'the portion of the reference genome (isolate 86–026 NP, PubMLST id: 5068, NCBI RefSeq GCF_000012185.1) that could be aligned against all of the other sequences' [63, 72]. In two previous studies, each NTHi isolate was assigned to one clade, with isolates in clades I, IV and V forming a monophyletic group in the phylogeny [62, 63]. These findings were not replicated in the current study. Our results suggested that biotyping and NTHi clade typing had limited discriminatory power and precision for describing *H. influenzae* phylogenetic relatedness [64]. The cgMLST method utilizes more genes and a gene-by-gene approach and can depict a high-resolution population structure free from reference bias.

In conclusion, a strength of the cgMLST scheme proposed here is its provision of a high-resolution representation of *H. influenzae* population structure, offering a reliable and accessible method to use for characterizing phylogenetic relatedness between *H. influenzae* isolates that can be utilized by microbiology reference laboratories and public health authorities worldwide. The cgST assignment gives a stable nomenclature to individual *H. influenzae* genomes, and this represents an advance in typing methodologies, particularly for the NTHi population as the predominant cause of *H. influenzae* invasive disease. However, grouping genomes based solely on the core genome allelic profile does not result in a stable clustering system. Combining the cgMLST scheme with a stable clustering approach such as life identification number coding will be beneficial to provide a high-resolution typing method that is stable for both naming individual isolates and clustering at higher levels [78]. Nevertheless, the thorough characterization of the bacterial core genome using the proposed cgMLST scheme also serves as a valuable resource for enhancing molecular diagnostics and advancing vaccine development.

Funding information
This study was funded by a Wellcome Trust Biomedical Resource Grant to M.J.C.M., A.B.B. and K.A.J. (grant number 218205/Z/19/Z) and a National Institute for Health and Care Research (NIHR) Grant called MEVacP to M.J.C.M., A.B.B. and O.B.H. Studentship for M.A.K. was funded by the Ministry of Education Indonesia in collaboration with the Medical Science Division, University of Oxford.

Acknowledgements
The authors wish to thank Prof Susanna Dunachie and Prof Samuel K. Sheppard for their valuable advice on the work.

Author contributions
M.C.J.M. and O.B.H. designed the study and wrote the manuscript with M.A.K.; M.C.J.M., O.B.H., A.B.B. and R.H. supervised the activities; M.A.K. and A.B.B. curated the draft genomes; A.B.B. and A.B. identified central genotypes for *H. influenzae* clonal complexes description; M.A.K. performed the pangenome analyses, created all figures and curated loci in cgMLST schemes; M.A.K. and W.M. wrote in-house R and Python scripts; K.A.J. implemented the cgMLST scheme in the PubMLST database.

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

Ethical statement
Not applicable

References

1. Carrol KC, Funke G, Landry ML, Richter SS, Warnock DW. *Manual of Clinical Microbiology*. Washington, DC: ASM Press; 2019.
2. Brooks GF, JawetzE, Melnick JL, AdelbergEA. *Jawetz, Melnick, &Amp; Adelberg's Medical Microbiology*. New York: McGraw Hill Medical; 2019.
3. Soeters HM, Blain A, Pondo T, Doman B, Farley MM, *et al*. Current epidemiology and trends in invasive haemophilus influenzae disease-United States, 2009-2015. *Clin Infect Dis* 2018;67:881–889.
4. Suga S, Ishiwada N, Sasaki Y, Akeda H, Nishi J, *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.
5. Bertran M, D'Aeth JC, Hani E, Amin-Chowdhury Z, Fry NK, *et al*. 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.
6. McTaggart LR, Cronin K, Seo CY, Wilson S, Patel SN, *et al*. Increased incidence of invasive Haemophilus influenzae disease driven by non-type B isolates in Ontario, Canada, 2014 to 2018. *Microbiol Spectr* 2021;9:e0080321.
7. 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.
8. Yamada S, Seyama S, Wajima T, Yuzawa Y, Saito M, *et al*. *β*-Lactamase-non-producing ampicillin-resistant Haemophilus influenzae is acquiring multidrug resistance. *J Infect Public Health* 2020;13:497–501.
9. Zhou Y, Wang Y, Cheng J, Zhao X, Liang Y, *et al*. Molecular epidemiology and antimicrobial resistance of *Haemophilus influenzae* in Guiyang, Guizhou, China. *Front Public Health* 2022;10:947051.
10. Cox AD, Williams D, Cairns C, St Michael F, Fleming P, *et al*. 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.

11. Shoukat A, Van Exan R, Moghadas SM. Cost-effectiveness of a potential vaccine candidate for Haemophilus influenzae serotype "a." *Vaccine* 2018;36:1681–1688.
12. 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.
13. Wilkinson TMA, Schembri S, Brightling C, Bakerly ND, Lewis K, *et al*. Non-typeable Haemophilus influenzae protein vaccine in adults with COPD: a phase 2 clinical trial. *Vaccine* 2019;37:6102–6111.
14. Haemophilus influenzae: Surveillance standard. Geneva: World Health Organization; 2018. p. 1-14.
15. Maiden MC. Multilocus sequence typing of bacteria. *Annu Rev Microbiol* 2006;60:561–588.
16. Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, *et al*. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol* 2013;11:728–736.
17. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. *Nat Microbiol* 2017;2:17040.
18. Schürch 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.
19. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, *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.
20. Preska Steinberg A, Lin M, Kussell E. Core genes can have higher recombination rates than accessory genes within global microbial populations. *elife* 2022;11:e78533.
21. de Been M, Pinholt M, Top J, Bletz S, Mellmann A, *et al*. Core genome multilocus sequence typing scheme for high- resolution typing of *Enterococcus faecium*. *J Clin Microbiol* 2015;53:3788–3797.
22. Blanc DS, Magalhães B, Koenig I, Senn L, Grandbastien B. Comparison of whole genome (wg-) and core genome (cg-) MLST (BioNumerics™) versus SNP variant calling for epidemiological investigation of *Pseudomonas aeruginosa Front Microbiol* 2020;11:1729.
23. 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.
24. Iskander M. *Development and Evaluation of Core Genome MLST Schema for Haemophilus Influenzae*. University of Manitoba, 2017.
25. Pinto M, González-Díaz A, Machado MP, Duarte S, Vieira L, *et al*. Insights into the population structure and pan-genome of *Haemophilus influenzae*. *Infect Genet Evol* 2019;67:126–135.
26. Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, *et al*. chewBBACA: a complete suite for gene-by-gene schema creation and strain identification. *Microb Genom* 2018;4:e000166.
27. Zallot R, Harrison KJ, Kolaczkowski B, de Crécy-Lagard V. Functional annotations of paralogs: a blessing and a curse. *Life* 2016;6:39.
28. Zhang J, Halkilahti J, Hänninen M-L, Rossi M. Refinement of whole-genome multilocus sequence typing analysis by addressing gene paralogy. *J Clin Microbiol* 2015;53:1765–1767.
29. Palma F, Mangone I, Janowicz A, Moura A, Chiaverini A, *et al*. *In vitro* and in silico parameters for precise cgMLST typing of *Listeria* monocytogenes. *BMC Genomics* 2022;23:235.
30. Jolley KA, Maiden MCJ. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 2010;11:595.
31. 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.
32. Harrison OB, Cehovin A, Skett J, Jolley KA, Massari P, *et al*. *Neisseria gonorrhoeae* population genomics: use of the gonococcal core genome to improve surveillance of antimicrobial resistance. *J Infect Dis* 2020;222:1816–1825.
33. Cody AJ, Bray JE, Jolley KA, McCarthyND, 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.
34. Rensburg M, Berger DJ, Fohrmann A, Bray JE, Jolley KA, *et al*. 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.
35. 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.
36. Liang KYH, Orata FD, Islam MT, Nasreen T, Alam M, *et al*. A vibrio cholerae core genome multilocus sequence typing scheme to facilitate the epidemiological study of cholera. *J Bacteriol* 2020;202:e00086-20.
37. 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.
38. Abdel-Glil MY, Chiaverini A, Garofolo G, Fasanella A, Parisi A, *et al*. A whole-genome-based gene-by-gene typing system for standardized high-resolution strain typing of *Bacillus anthracis*. *J Clin Microbiol* 2021;59:e0288920.
39. Tourasse NJ, Jolley KA, Kolstø A-B, Økstad OA. Core genome multilocus sequence typing scheme for *Bacillus cereus* group bacteria. *Res Microbiol* 2023;174:104050.
40. Appelt S, Rohleder AM, Jacob D, von Buttlar H, Georgi E, *et al*. Genetic diversity and spatial distribution of *Burkholderia mallei* by core genome-based multilocus sequence typing analysis. *PLoS One* 2022;17:e0270499.
41. Moreno-Manjón J, Jolley KA, Maiden MC. *Acinetobacter baumannii* core genome multilocus sequence typing. 2022.
42. Abdel-Glil MY, Thomas P, Linde J, Jolley KA, Harmsen D, *et al*. Establishment of a publicly available core genome multilocus sequence typing scheme for *Clostridium perfringens*. *Microbiol Spectr* 2021;9:e0053321.
43. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, *et al*. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.
44. 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:giz119.
45. Ding W, Baumdicker F, Neher RA. panX: pan-genome analysis and exploration. *Nucleic Acids Res* 2018;46:e5.
46. Zhao Y, Wu J, Yang J, Sun S, Xiao J, *et al*. PGAP: pan-genomes analysis pipeline. *Bioinformatics* 2012;28:416–418.
47. Gautreau G, Bazin A, Gachet M, Planel R, Burlot L, *et al*. PPanGGOLiN: depicting microbial diversity via a partitioned pangenome graph. *PLoS Comput Biol* 2020;16:e1007732.
48. Peng Y, Tang S, Wang D, Zhong H, Jia H, *et al*. MetaPGN: a pipeline for construction and graphical visualization of annotated pangenome networks. *Gigascience* 2018;7:giy121.
49. Zhou Z, Charlesworth J, Achtman M. Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. *Genome Res* 2020;30:1667–1679.
50. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, *et al*. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* 2020;21:180.

51. Boubour A. *Genomic Characterisation of Haemophilus Influenzae Capsular Locus*. University of Oxford, 2021.
52. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
53. 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.
54. Cantalapiedra CP, Hernández-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.
55. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, *et al*. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res* 2021;49:D274–D281.
56. 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.
57. Bruen TC, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 2006;172:2665–2681.
58. Lai YP, Ioerger TR. A statistical method to identify recombination in bacterial genomes based on SNP incompatibility. *BMC Bioinformatics* 2018;19:450.
59. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* 2015;11:e1004041.
60. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.
61. Watts SC, Holt KE. In silico serotyping of the *Haemophilus influenzae* capsule locus. *J Clin Microbiol* 2019;57.
62. Carrera-Salinas A, González-Díaz A, Calatayud L, Mercado-Maza J, Puig C, *et al*. Epidemiology and population structure of *Haemophilus influenzae* causing invasive disease. *Microb Genom* 2021;7:000723.
63. De Chiara M, Hood D, Muzzi A, Pickard DJ, Perkins T, *et al*. 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.
64. Nørskov-Lauritsen N. Classification, identification, and clinical significance of *Haemophilus* and *Aggregatibacter* species with host specificity for humans. *Clin Microbiol Rev* 2014;27:214–240.
65. 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.
66. Zhou Z, Alikhan N-F, Sergeant MJ, Luhmann N, Vaz C, *et al*. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res* 2018;28:1395–1404.
67. 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:e00821-18.
68. Harrison A, Dyer DW, Gillaspy A, Ray WC, Mungur R, *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.
69. Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. *Nat Rev Microbiol* 2015;13:787–794.
70. Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, *et al*. 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.
71. Eutsey RA, Hiller NL, Earl JP, Janto BA, Dahlgren ME, *et al*. Design and validation of a supragenome array for determination of the genomic content of *Haemophilus influenzae* isolates. *BMC Genomics* 2013;14:484.
72. Kc R, Leong KWC, Harkness NM, Lachowicz J, Gautam SS, *et al*. 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:mgen000405.
73. Gonzalez-Diaz A, Carrera-Salinas A, Pinto M, Cubero M, van der Ende A, *et al*. Comparative pangenome analysis of capsulated *Haemophilus influenzae* serotype f highlights their high genomic stability. *Sci Rep* 2022;12:3189.
74. Topaz N, Tsang R, Deghmane A-E, Claus H, Lâm T-T, *et al*. Phylogenetic structure and comparative genomics of multi-national invasive *Haemophilus influenzae* serotype a isolates. *Front Microbiol* 2022;13:856884.
75. Potts CC, Topaz N, Rodriguez-Rivera LD, Hu F, Chang H-Y, *et al*. Genomic characterization of *Haemophilus influenzae*: a focus on the capsule locus. *BMC Genomics* 2019;20:733.
76. Meyler K, Meehan M, Bennett D, Mulhall R, Harrison O, *et al*. Spontaneous capsule loss in *Haemophilus influenzae* serotype b associated with Hib conjugate vaccine failure and invasive disease. *Clin Microbiol Infect* 2019;25:390–391.
77. Kilian M. A taxonomic study of the genus Haemophilus, with the proposal of a new species. *J Gen Microbiol* 1976;93:9–62.
78. 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.

The Microbiology Society is a membership charity and not-for-profit publisher.

Your submissions to our titles support the community – ensuring that we continue to provide events, grants and professional development for microbiologists at all career stages.

Find out more and submit your article at microbiologyresearch.org