The major pathogen *Haemophilus influenzae* experiences pervasive recombination and purifying selection at local and global scales

Neil MacAlasdair[o1,2], **Anna K. Pöntinen**[o1], **Clare Ling**[3,4], **Sudaraka Mallawaarachchi**[1,], **Janjira Thaipadungpanit**[7], **Francois H. Nosten**[4,8], **Claudia Turner**[3,4], **Stephen D. Bentley**[2], **Nicholas J. Croucher**[9], **Paul Turner**[*3,4], **and Jukka Corander**[*1,2,10]

[1]Department of Biostatistics, Institute of Basic Medical Sciences, Faculty of Medicine, University of Oslo, Oslo, Norway
[2]Parasites and Microbes, Wellcome Sanger Institute, Hixton, United Kingdom
[3]Cambodia Oxford Medical Research Unit, Angkor Hospital for Children, Siem Reap, Cambodia
[4]Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK
[5]Peter MacCallum Cancer Centre, Melbourne, Victoria 3052, Australia;
[6]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Victoria 3052, Australia
[7]Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand
[8]Shoklo Malaria Research Unit, Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Mae Sot, Thailand
[9]MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, UK
[10]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
[o]These authors contributed equally.
[*]These authors contributed equally.

Haemophilus influenzae **is a major opportunistic human pathogen which causes both non-invasive and invasive disease. The *H. influenzae* type b (Hib) vaccine has led to a significant reduction of invasive Hib disease, but offers no protection against colonisation or disease by non-typeables (NT) or non-b serotypes, and *H. influenzae* remains a public health burden worldwide, with increasing reports of multi-drug resistance (MDR). Despite this, there is no comprehensive understanding of the species' global population structure. Here, we whole-genome sequenced >4,000 isolates of *H. influenzae* from an unvaccinated paediatric carriage and pneumonia cohort from northwestern Thailand. Despite no Hib immunisation, serotype b was uncommonly found, while 91.7% of isolates were NT. We identified a large number of nearly pan-resistant lineages and discovered that no lineages were enriched among disease samples, suggesting the ability to cause invasive disease is not restricted to any subpopulation. Population genetic analyses of our data combined with nearly 6,000 published genomes revealed a highly admixed population structure, low core genome nucleotide diversity, and evidence of pervasive negative selection. Global isolates confirm that MDR lineages are not confined to our newly-sequenced dataset, and their establishment globally is an urgent concern.**

Haemophilus influenzae **| Bacterial Population Structure | Bacterial Evolution | NTHi**

Correspondence: ***Jukka Corander (jukka.corander@medisin.uio.no), Paul Turner (pault@tropmedres.ac)***

Background

The nasopharynx is the natural habitat of the bacterium *Haemophilus influenzae*, where it exists in asymptomatic carriage, while frequently translocating to other body sites, such as inner ears, lungs and sinuses, causing a range of disease manifestations (1). Most common of these is acute otitis media (AOM), which is one of the leading causes of antibiotic prescriptions in children. Global estimates suggest over 700 million AOM cases per annum caused in total by any bacterial pathogen and a significant fraction of these lead to further complications and sequelae, particularly in low- and middle-income countries (LMICs) (2). After licensing of the polysaccharide-protein conjugate *H. influenzae* type b (Hib) vaccine in the late 1980s, its adoption into national vaccination programs worldwide has led to a significant reduction of Hib colonisation and its associated invasive disease manifestations, such as meningitis and pneumonia. However, the vaccine does not protect against colonisation by other serotypes or unencapsulated non-typeable *H. influenzae* (NTHi) . Therefore, *H. influenzae* remains a significant cause of AOM, sinusitis, conjunctivitis and pneumonia, and consequently is an important public health burden globally. A particular concern has arisen from the widespread antibiotic resistance observed in some strains of NTHi (3).

The evidence for NTHi as an important cause of paediatric community-acquired pneumonia (CAP) has been summarised in comprehensive reviews (4, 5). Determination of aetiology in paediatric CAP remains a challenge, with a minority of cases being bacteraemic. However, specimens obtained via bronchoscopy revealed NTHi to be the dominant bacterial pathogen in 250 Belgian children with recurrent or non-resolving CAP (6). Nasopharyngeal colonisation by non-Hib / NTHi, especially at higher densities, has also been shown to be associated with paediatric CAP in LMICs (7, 8).

Whole-genome sequencing studies of *H. influenzae* have been mainly conducted from smaller-scale collections of disease cases (9, 10), but rarely from large-scale collections of both carriage and disease isolates of the same population. Furthermore, few studies have been conducted in LMICs, where nasopharyngeal pathogen colonisation rates and the burden of CAP are generally much higher than in high income countries. As a consequence, the genetic population structure and evolutionary dynamics of the species remain poorly understood, particularly at a global scale (11).

This motivated us to conduct a longitudinal paediatric cohort study of both healthy colonisation and pneumonia among

Serotype	Number of Isolates	% of Maela collection
a	16	0.40%
b	227	5.7%
c	6	0.15%
d	8	0.20%
e	36	0.91%
f	36	0.91%
NT	3639	91.7%

Table 1. Proportion and counts of the number of isolates of each serotype and non-typable/capsule null isolates

a large birth cohort in a population located in Northwestern Thailand, the Maela camp for displaced persons. The densely populated camp is located on the Thailand-Myanmar border and provided a unique opportunity to systematically sample both carriage and disease cases in a pre-Hib vaccine population. Here we detail results from the whole-genome sequencing (WGS) of isolates from our longitudinal cohort study, and also of further analyses performed on the Maela data combined with all publicly available *H. influenzae* WGS data. This combined collection of nearly 10,000 genomes allowed us to conduct genomic analyses of the species at an unprecedented scale and use both local and global scales of interpretation.

Results

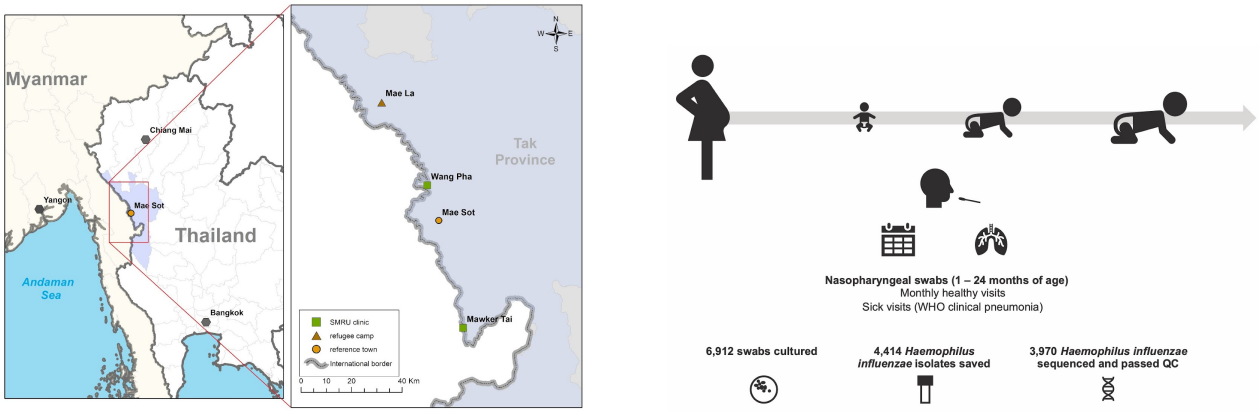
Serotype distribution across the Maela paediatric population. Infants in the Maela cohort (Fig. 1) were predominantly colonised by NT *H. influenzae*, despite lacking immunisation against Hib. Out of 3,970 isolates that passed final QC filters, the counts and estimated frequencies of the six different serotypes (12) and non-typable (capsule null) isolates are listed in Table 1. Notably, non typable isolates made up 91.7% of all isolates, and serotype b isolates are the second most prevalent, making up 5.7% of the population. The remaining 5 serotypes account for less than 1% of the population each. Serotypable isolates generally form monophyletic clades on the tree (Fig. 2). Two isolates gave only partial *in silico* capsule typing results. In silico capsule typing was generally congruent with the agglutination-based phenotypic serotyping, and was corrected by the latter in those 28 cases where the serological typing indicated a serotype for a non-typeable *in silico* type. These included serotype a ($n = 1$), b ($n = 19$), d ($n = 3$), and e ($n = 5$).

Genetic population structure of *H. influenzae* in Maela. Popular epidemiological clustering schemes, such as multilocus sequence typing (MLST), failed to provide meaningful insight into the population structure of H. influenzae. While MLST sequence types (STs) and clonal complexes (CCs) can be used to partition the Maela *H. influenzae* population into groups of closely related isolates, they were unable to provide meaningful insight into the population structure. A majority of the Maela isolates (2,344/3,970) contained combinations of alleles not present in the current MLST database, likely reflecting the extent to which research on *H. influenzae* has been biased towards high-income countries. Of the 1,626 isolates which were successfully typed, many STs and CCs remained paraphyletic, with isolates from the ST or CC being found interspersed with isolates from other STs/CCs on the estimated core genome phylogeny (Supplementary Figure 1). Using PopPUNK to cluster all the isolates was largely able to identify monophyletic clusters (Fig. 2), however, the optimal clustering divided the population into many clusters. The largest PopPUNK cluster contained 349 isolates, with only 13 PopPUNK clusters in total each representing at least 100 isolates (out of a total of 122 clusters). A large fraction of clusters (50%) contained 10 or fewer isolates. NT *H. influenzae* were observed as the dominating serotype in both healthy carriage and pneumonia time point samples (Fig. 2), and no particular genetic lineage was overrepresented in pneumonia (Fisher's exact test, p-value $= 0.084$) (Methods).

Distribution of AMR determinants. AMR determinants were frequently identified across the phylogeny and strongly associated with MDR lineages (Fig. 2). Only one of these MDR lineages was clearly associated with serotype b (Fig. 2). 41 PopPUNK clusters contained at least one MDR isolate, indicating repeated acquisition of AMR determinants across the population.

Quantification of homologous recombination. Since the acquisition of AMR determinants is likely aided by horizontal gene transfer in this naturally transforming species, we quantified the extent of homologous recombination. Mapping Illumina reads from isolates in the same cluster against long-read reference assemblies, to produce whole genome pseudoalignments that were used as inputs to SNP-density based recombination analysis (Gubbins) was not feasible to quantify recombination in the entire Maela cohort due to the size of the dataset and the large number of clusters present.

Consequently, we used a complementary approach by leveraging the aligned pangenome genes for the 3,970 isolates (Methods). Of the pangenome of 7015 genes, at least one recombination event between PopPUNK lineages was identified in 2672 genes (38%). On average, 193.36 recombination events were identified per gene (including recombination free genes) and the frequency of recombination events was significantly correlated with the estimated nucleotide diversity per gene (Spearman's $r = 0.49$, $p < 7.15$Ö10^{-293}). A substantial proportion of genes with no detected recombination (64.0%) also had zero nucleotide diversity (Fig. 4A). Finally, we also quantified the rate of decay of linkage disequilibrium (LD) in the core genome and compared this with several other common bacterial pathogens analysed in (13). This shows that the decoupling of SNPs as a function of base pair distance happens fastest in *H. influenzae* (Fig. 4B), and the rate is considerably elevated compared with other species known to be highly recombinant, such as *Campylobacter jejuni* and *Enterococcus faecalis*. Taken together, these results suggest that the *H. influenzae* population within Maela is extremely recombinant, to the extent that it reduces the overall level of diversity within the population.



(a) **(b)**

Fig. 1. A: Geographical location of the study site. B: Cohort design and sample processing.

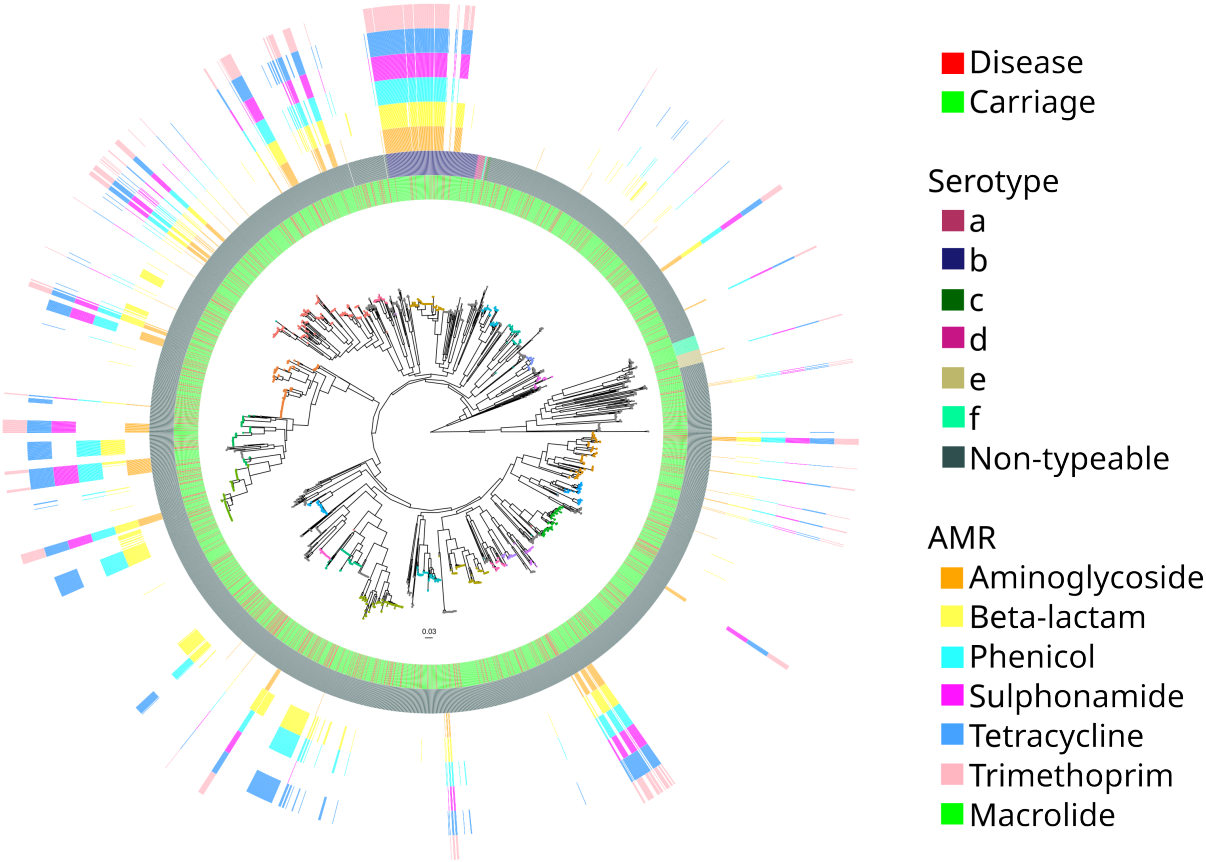
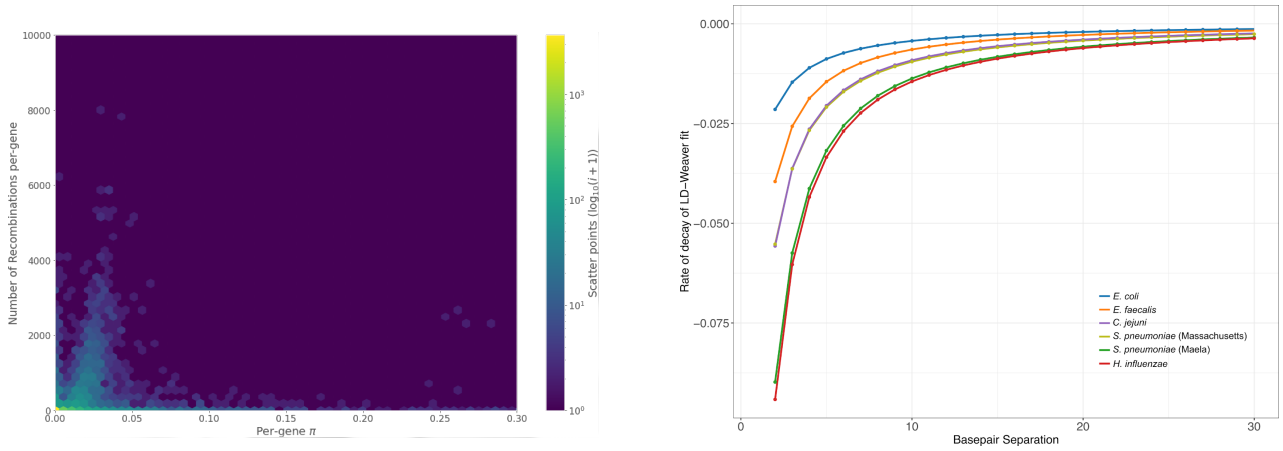


Fig. 2. Neighbour-joining phylogeny of Maela *H. influenzae* genomes for 3970 isolates from carriage and pneumonia samples (inner ring), estimated using FastTree v.2.1.10 on the core-genome alignment mapped against the *H. influenzae* reference 86-028NP (NC_007146.2). The 20 largest PopPUNK clusters (>50 isolates) are indicated by coloured dots at the tips of the phylogeny, while smaller clusters are grey. In silico serotypes (second ring), inferred by using Hicap v.1.0.3, and AMR profiles (six outer rings), screened with AMRFinderPlus v.3.10.18, are shown by colour as indicated in the legend. AMR = antimicrobial resistance. An interactive online phylogeny, with additional metadata including MLST and Clonal Complex data is available at the following link: https://microreact.org/project/oMm8PFCoG2429JwiDBpdru-maela-h-influenzae



(a) **(b)**

Fig. 3. A: Log-scaled hexagon-density scatter plot of the estimated per-gene nucleotide diversity, π, versus the number of recombination events inferred per-gene. Hexagon colours indicate the number of scatter plot points (log-scaled) present within each hexagon. B: The estimated rate of decay in linkage disequilibrium (LD) for the *H. influenzae* Maela cohort, compared with estimated rates for the four species with LD decay functions fitted in (Mallawaarachchi et al. 2024). The shown curves correspond to gradients of the LD decay function, with smaller values indicating faster decoupling of SNPs as a function of distance in base pairs.

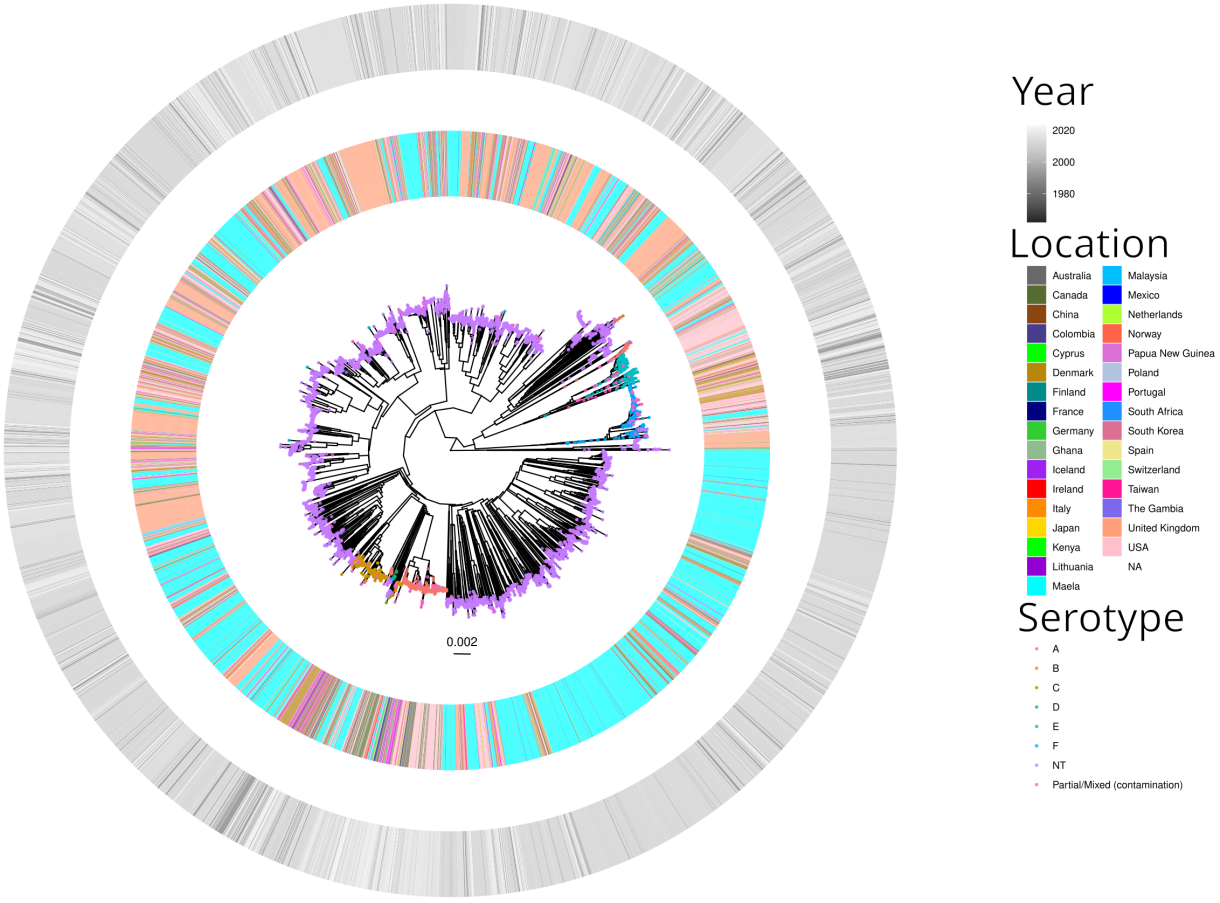


Fig. 4. Core genome distance neighbour-joining phylogeny of 9,849 *H. influenzae* isolates, combining the Maela cohort and a systematically identified collection of published isolates from around the globe. The phylogeny was estimated using core-genome distances inferred with PopPUNK v.2.4.0. In silico serotypes are indicated by the circles on the tips of the phylogeny, isolation location is indicated on the inner ring, while the outer ring indicates the year of isolates collection, as shown by colour as indicated in the legend. An interactive online phylogeny, with additional metadata including MLST, Clonal Complex, and partial disease state data is available at the following link: https://microreact.org/project/ioyt4oJRSJgeFGK9KmFyVk-global-h-influenzae-core-tree

Population genetic analyses of the global dataset. To understand how the genetic variation observed in the Maela cohort samples relates to internationally circulating *H. influenzae*, we combined the study data with a systematic collection of all publicly available *H. influenzae* genome data with basic metadata available (country and year of collection), for a total dataset comprising 9,849 isolates (Methods). Clustering the global population with MLST resulted in many fewer non-database sequence types, only 2,669 of the total (including 2,344 from Maela). Similarly to the Maela data, many STs and CCs were paraphyletic and split across the phylogeny (Supplementary Figure 1). PopPUNK clustering of the combined dataset successfully identified 752 monophyletic clusters, with the largest cluster composed of 483 isolates, and 19 clusters containing at least 100 isolates. 595 clusters contained fewer than 10 isolates, leading to many larger clusters being paraphyletic according to the core genome tree, with the monophyletic clade containing small or singleton clusters, likely reflecting a change in the accessory genome of the smaller clade which brought the pairwise distances above PopPUNK's clustering threshold.

The core genome phylogeny of the combined collections (Figure 3, Maela isolates in light blue) clearly demonstrates that the Maela isolates are extensively interspersed within the global population of the species, suggesting rapid cross-border and intercontinental transmission of *H. influenzae*. Furthermore, isolates spanning the sampling window, from 1962-2023 are distributed across the phylogeny and do not form monophyletic clusters made up of temporally restricted isolates. Together, these patterns strongly suggest that the history of migration within the global *H. influenzae* population is sufficiently frequent and extensive to overwhelm the phylogeographical signal of local clonal expansion of lineages.

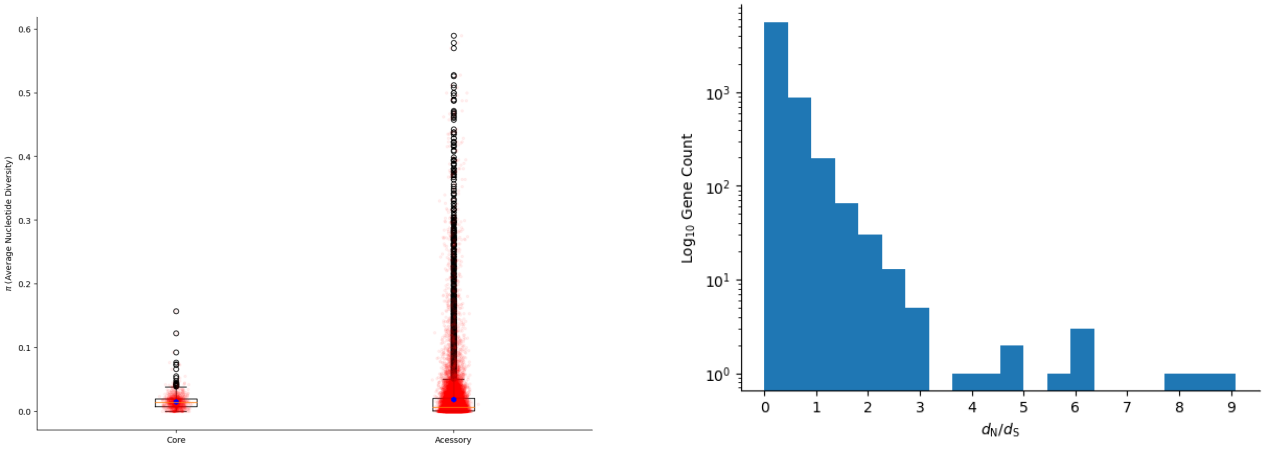
Due to the extremely low level of nucleotide diversity evident during the recombination analysis of the Maela cohort, we further investigated the overall level of nucleotide diversity across the aligned pangenome of the entire global collection, which consisted of 18,265 genes. Nucleotide diversity in both core and accessory genes was generally low (Figure 5A), with core genes significantly less diverse than accessory genes (Mann-Whitney U test, $p = 1.205 \times 10^{-5}$, Figure 5A).

To understand how selective forces may be influencing the diversity observed within the pangenome, we further estimated d_N/d_S, the ratio of nonsynonymous to synonymous nucleotide mutations within every gene of the pangenome (Methods). Consistent with the low level of diversity observed, the average d_N/d_S value was 0.28, and 96% of the 6,853 genes for which it was successfully estimated (Methods) had $d_N/d_S < 1$ (Figure 5B). This implies that negative selection is widespread across the coding regions of the *H. influenzae* genome. Of the remaining 256 genes (4%) with a d_N/d_S estimate, 45 had $d_N/d_S > 2$, indicating potential positive directional selection. Further analysis of these genes was undertaken using three statistical tests implemented in the HYPHY v.2.5.60 (Methods), and a few accessory genes possessed extremely strong evidence of selection, where at least two of the three statistical tests rejected the null hypothesis of neutral evolution (Methods). The genes involved included an unnamed gluconate transporter, the BrnT toxin protein, and a third small protein of unknown function. The results of these analyses are illustrated in Supplementary Figures 2-3, and briefly summarised as follows. The unnamed gluconate transporter showed statistically significant results in all three HYPHY tests used, and these tests indicated a branch of the gene phylogeny containing eight isolates, and a specific codon (185) in the protein alignment which have been positively selected for. This branch consists of seven Maela isolates and one isolate from elsewhere, which all possess a structural variant of the unnamed gluconate transporter with a large deletion of a transmembrane domain. The brnT toxin gene, the toxin from the BrnT/BrnA type II toxin-antitoxin system (14) also showed statistically significant results in all three HYPHY tests, which indicated that a branch of gene phylogeny consisting of two Maela isolates with a large deletion of an alpha helix had recently been subject to positive selection. Finally, the protein of unknown function showed a statistically significant result in two of the three HYPHY tests, identifying a glutamine/valine variable site, with the valine variant primarily associated with Maela isolates, and the glutamine variant primarily associated with isolates from elsewhere. All three of these proteins are low-frequency accessory genes which are globally distributed. Notably, the variants identified as under selection in the *brnT* toxin and the unnamed gluconate transporter are unique or much more prevalent in Maela isolates, with a similar split association between the two variants of the unnamed protein, suggesting that the either the unique sampling strategy of the dataset or environment of the camp may be resulting in more statistical power to detect selection, or genuinely stronger positive selection and rapid local adaptation.

Finally, to explore the geographic distribution of MDR lineages in greater detail, we focused on the lineages with at least 50 isolates and 30% resistance prevalence for at least 3 of 7 antibiotic classes. This revealed that all large MDR lineages (n = 3) are widely disseminated internationally, i.e. observed in at least 11 different locations (Fig. 6). Only one of these lineages was dominated by Hib strains and showed evidence of independent capsule switches to serotype a (Fig. 6). The others were composed of non-typable isolates.

Discussion

The genomic epidemiology of non-b *H. influenzae* has remained largely elusive to date, due to the lack of carriage studies in high-burden settings, particularly in populations from before the rollout of the Hib vaccine. Our study provides the first comprehensive evidence that NT *H. influenzae* are equally capable of causing invasive disease irrespective of their genetic background, even in a pre-Hib vaccine host population. While only colonising isolates were available from the Maela cohort, sampling during episodes of clinical pneumonia has provided insights into disease-associated strains. Results from the multi-country PERCH pneumonia aetiology study (8) confirmed a positive association between non-b



(a) **(b)**

Fig. 5. A: Boxplots of the estimated average pairwise nucleotide diversity, π, in each gene of the aligned pan-genome of the combined dataset, split into genes present in 80% or more isolates (core), and genes in less than 80% of isolates (accessory). Blue hexagons indicate gene-frequency weighted average nucleotide diversity across all genes. B: Log-scaled histogram of the estimated d_N/d_S values, the ratio of nonsynonymous to synonymous mutations, across 6,853 aligned genes from pan-genome of the combined dataset.

H. influenzae upper respiratory colonisation and chest x-ray confirmed pneumonia. The same study demonstrated an etiologic fraction of 4.5% for non-b *H. influenzae* amongst HIV negative, chest x-ray confirmed cases, which is comparable to the 6.7% fraction estimated for *S. pneumoniae*. There is evidence from several countries of an increasing burden of invasive non-b *H. influenzae* disease, notably in neonates and older adults, with the vast majority being NTHi infections (1). The high burden of pneumonia attributable to non-b *H. influenzae*, combined with the frequent emergence of MDR lineages as identified in the current study, serves thus as a reminder of the significant health benefits of developing an immunisation program targeting eradication of these pathogens. This would contribute not only towards removing the public health burden of non-B *H. influenzae* invasive disease, but also to significantly reduce both the incidence of AOM and the need to prescribe antibiotics to children.

Apart from the genomic epidemiology of *H. influenzae*, the overall understanding of the species' population structure has also remained largely elusive, particularly at a global scale, despite various efforts to elucidate it over the past decade. This study, through analysing a large cohort of isolates from an understudied region combined with a systematic collection of publicly available data, suggests that the global *H. influenzae* population is not structured into independently evolving lineages which are predominant in certain regions. This is unlike other well-studied bacterial species that colonise the same niche, such as *Streptococcus pneumoniae* or *Neisseria meningitidis*, where distinct independently evolving lineages have been readily identified for decades (15–17). *H. influenzae* instead appears to have a population structure reminiscent of panmixia, where routine gene flow between members of the species prevents the formation of stable lineages. This type of population structure would account for the limited success of various methods used to cluster the population in this study, and the difficulties previous efforts have encountered when using smaller datasets (11). Technically, clustering methods have likely been limited by the low levels of nucleotide diversity observed within the *H. influenzae* genome, particularly the core genome, even at a global scale. This, however, is not readily apparent in the results of clustering methods, and only becomes evident when population genomics analyses are conducted at scale.

Despite the low levels of nucleotide diversity, phylogenetic analysis of the combined Maela and globally sequenced isolates remains possible and clearly demonstrates a persistent lack of phylogeographical signal, where closely related isolates are highly co-localised, in *H. influenzae*, even with this collection of isolates spanning over 50 years. This strongly suggests that inter-regional and intercontinental transmission of these bacteria happens frequently. This is consistent with the high levels of recombination observed in the Maela dataset, as that would facilitate the efficient admixture of migrating isolates with the destination population. Furthermore, the frequent migration and recombination, when combined with widespread evidence of negative selection across coding regions of the genome implied by the low d_N/d_S values, corresponds to a pool of biological forces that likely explains the low nucleotide diversity within coding regions of the *H. influenzae* genome, particularly the core. It is difficult to disentangle the individual contributions of migration, recombination and negative selection in producing low levels of diversity, and indeed they are likely acting in concert, as has been demonstrated previously in other ecological settings (18, 19).

These results underscore the importance of a global perspective on disease surveillance when developing public health strategies for managing invasive *H. influenzae* disease, as it is clear that pathogenic adaptations which arise in one part of the world have ample opportunity for global spread. Although the near pan-resistant (five antibiotic classes) MDR lineages we have identified in this study are primarily located in the Maela data, this remains of particular concern as the bias toward high-income settings in our systematic collection of previously sequenced and published *H. influenzae* does not exclude the possibility that these lineages may

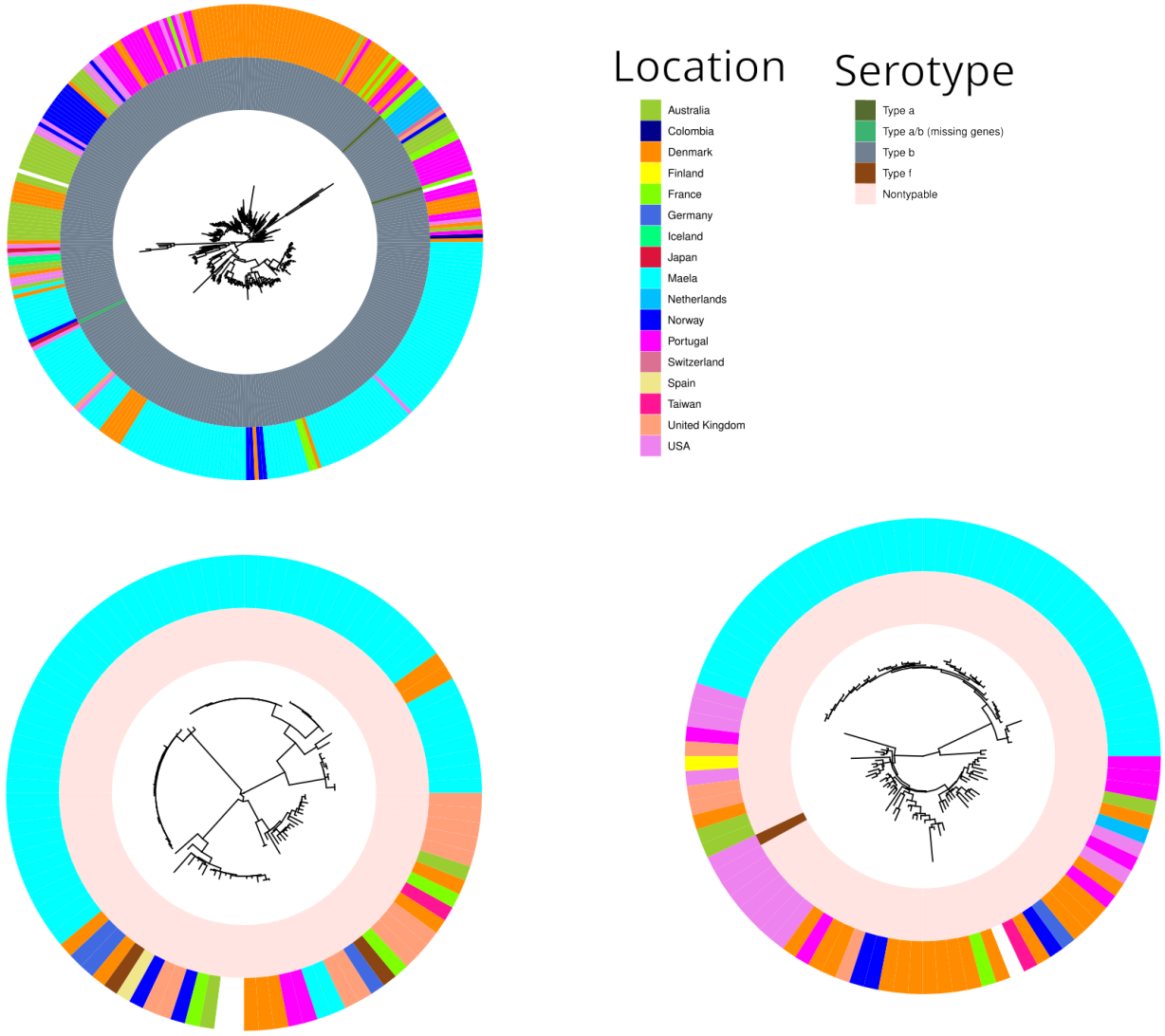


Fig. 6. Recombination-free maximum-likelihood phylogenies for each PopPUNK MDR cluster (Clusters 3, 19, and 20), with more than 30% resistance prevalence for a minimum of 3/6 antibiotic classes, comprising at least 50 genomes. The phylogenies were inferred by using Gubbins on whole-genome pseudoalignments of each cluster, separately mapped against the *H. influenzae* reference 86-028NP (NC_007146.2). *In silico* serotypes (inner ring) and isolation location (second ring) are shown by colour as indicated in the legend. MDR stands for multidrug resistance.

be further transmitted and established in LMIC populations with high antibiotic use. Intensified efforts should be made to include *H. influenzae* into AMR surveillance programs as widely as possible. Such surveillance should preferably not be limited to including only bloodstream isolates, because it will otherwise underestimate the prevalence of circulating AMR determinants among pneumonia and AOM clinical cases. Given the significant evidence of adaptation in accessory genes in the Maela population, and all but one of the pan-resistant MDR lineages was predominantly identified in the newly sequenced Maela isolates, an unlikely cause for optimism may still be that the camp host population may be uniquely suited to evolutionary adaptation. This could be due to either the human population density resulting in high colonisation and transmission success, or the level of antibiotic use in the camp, and it is possible that the fitness cost of maintaining high levels of resistance beyond these settings is prohibitive. However, it is probably more likely that the difference in sampling strategy in the Maela data leads to higher statistical power to identify adaptation when it has occurred, suggesting that similar adaptation is likely elsewhere. Widespread genomic genomic surveillance in such settings is most likely to lead to the early detection of the spread of extensive levels of AMR, and allow for targeted intervention. Importantly, this type of surveillance data would also be crucial in developing an understanding of the parameters of how selection drives the evolution and maintenance of AMR in *H. influenzae*.

Apart from the concerning implication regarding the possibility of the global spread of AMR in *H. influenzae*, the results

of this study also suggest that vaccination may be a particularly effective strategy to control invasive *H. influenzae* disease irrespectively of the serotype, due to the lower level of diversity present within its core genome relative to the accessory, and its highly admixed population structure. Given the low level of observed allelic diversity, the pervasive negative selection we detected throughout the *H. influenzae* genome at a global scale may be strong enough to overcome selection driving compensatory adaptations which would generally reduce vaccine efficacy in response to rollout. Although this is cause for optimism, it must be tempered by the fact that the high levels of recombination observed in *H. influenzae* may also increase the efficacy of positive selection on any mutations which do arise, as has been observed in other species (20). In any case, the stark contrast between the *H. influenzae* population structure identified in this work and the highly stratified population structure of S. pneumoniae, both globally (16) and in the Maela host population (21), strongly suggests that vaccine evasion through inter-lineage competition and replacement, as has repeatedly been observed in *S. pneumoniae* (22) would be much less likely to happen in *H. influenzae*, due to the absence of a deeply structured population and local variants. Although there are many complications involved in the design of protein based bacterial vaccines which would need to be overcome (23) this work supports the conjecture that a single universal vaccine could possibly be developed to combat invasive *H. influenzae* disease, and suggests that the eradication of invasive disease caused by *H. influenzae* may be a feasible end goal of widespread vaccination campaigns.

Methods

Study design and collections. A total of 4,474 *H. influenzae* isolates were retrieved from a mother-infant cohort of 999 pregnant women from the Maela camp for displaced persons, Thailand, from October 2007 to November 2008 (24, 25). Within 24 months' postpartum period, infants were sampled by nasopharyngeal swabs monthly and when the infant presented symptoms of pneumonia. Of the whole-genome sequences, 3970 passed the quality control and were included in the genomic analyses. For comparative analyses, a systematic search was conducted for publicly available short-read genome sequences for which country and year of isolation metadata was available. 6129 isolate data were retrieved from the ENA ((9, 10, 26–54), of which 5,879 passed quality control, resulting in a final dataset size of 9849 isolates.

Sampling and Sequencing Procedures. Between October 2007 and November 2008, 999 pregnant women from the Maela camp for displaced persons (located on the Thailand-Myanmar border in Tak province, NW Thailand, Fig. 6a) were recruited into a mother-infant colonisation study. Infants were followed from birth for 24 months and a nasopharyngeal swab specimen collected (dacron tipped swabs; Medical Wire & Equipment, Corsham, UK) at monthly intervals and if the infant presented to the Shoklo Malaria Research Unit clinic with symptoms and signs compatible with WHO clinical pneumonia (Fig 6b).

Following sampling, the NPS tip was excised immediately into a sterile cryovial containing 1mL STGG (skim milk, tryptone, glucose, glycerol medium; prepared in-house) using 70% ethanol-cleaned scissors. NPSS-STGG specimens were transferred to the SMRU microbiology laboratory in a cool box, within eight hours of collection, and were frozen at -80°C until culture.

Ten microlitres of thawed NPS-STGG specimen was cultured onto plain chocolate agar (Clinical Diagnostics, Bangkok, Thailand), a 10 unit bacitracin disc (Oxoid, Basingstoke, UK) applied to the first streak, and the plate incubated overnight at 36°C in 5% CO2. Bacitracin resistant colonies were confirmed as *H. influenzae* by Gram stain and X+V factor dependent growth. Serotype was determined by slide agglutination (Becton Dickinson, Franklin Lakes NJ, USA). Pure isolates of *H. influenzae* were harvested from an overnight culture plate into 1mL STGG and stored at -80°C prior to DNA extraction and sequencing.

Short read whole-genome sequencing of the 4,474 *H. influenzae* isolates was performed at the Sanger Wellcome Institute on Illumina-HTP NovaSeq 6000 platform with 150bp paired-end sequencing.

For long-read sequencing, one reference isolate was selected per each of the 48 largest PopPUNK clusters, covering 3558 (89.6%) of 3970 isolates of the study cohort. Reference isolates were selected based on the gene presence absence matrix from the estimated pangenome, using a published selection pipeline (https://gitlab.com/sirarredondo/long_read_selection) (55).

The selected *H. influenzae* strains were subcultured on chocolate agar and incubated overnight at 35-37°C in 5% CO_2. Genomic DNA was extracted using the Qiagen MagAttract HMW DNA Kit. Whole-genome sequencing libraries were constructed using the Oxford Nanopore Technologies SQK-NBD112.96 Native Barcoding Kit and all 48 strains were pooled together and sequenced on one ONT R9.4.1 flowcell using a MinION Mk1c. Hybrid assembly of the reference isolates was performed using a publicly available pipeline (https://github.com/arredondo23/hybrid_assembly_slurm) with minimum ONT coverage of 40x and a phred score of 20 to trim the Illumina reads, resulting in 40 (83.3%) complete hybrid assemblies of the 48 reference isolates.

Genomic Analysis. A total of 4474 *Haemophilus influenzae* isolates were sequenced at the Sanger Wellcome Institute (Hinxton, UK) on Illumina-HTP NovaSeq 6000 150 bp paired-end platform. Species contamination was identified by using Kraken v.0.10.6 (56). The sequence data failed quality control if the depth of coverage was <20x, or if there was evidence of contamination or mixed strains, poor assembly or extreme violation of any of the quality control parameters. Short-read genome sequences, both newly sequenced and publicly available, were assembled and annotated using a published pipeline with default parameters (57). Isolates were clustered using PopPUNK v.2.4.0 (58) separately on the Maela data, with core threshold of 0.11, and on the combined global data with default thresholds. Antimicrobial

resistance genes and point mutations were screened from assemblies using AMRFinderPlus v.3.10.18 (59) and database version 2022-04-04.1 with minimum identity of 75% and minimum coverage of 80%. Hicap v.1.0.3 (12) was used to infer capsule type from assemblies. Sequence type (ST) was identified from assembled genomes using mlst (60) and the *H. influenzae* MLST database, and the Clonal Complex typing scheme from the latter was used to group isolates into clonal complexes. For phylogenetic analyses on the Maela data, the sequence reads of the 3,970 quality control-passed genomes were mapped to the complete genome of *H. influenzae* 86-028NP (NC_007146.2) (61) using Snippy v.4.6.0 (https://github.com/tseemann/snippy) and a SNP-only sequence alignment was created using snp-sites v.2.5.1 (62). A neighbour-joining phylogeny was inferred using FastTree v.2.1.10 with a generalised time-reversible model (63), and a neighbour-joining tree was inferred with PopPUNK v.2.4.0 using core-genome distances to analyse the combined global collection.

The pangenome was inferred for the Maela genome collection using Panaroo v.1.2.9 (64) using sensitive mode and merging paralogs. The pangenome was further inferred for the entire combined global dataset running Panaroo in strict mode, merging paralogs. FastGEAR v.2016-12-16 (65) was used to infer recombinations, and pixy v.1.2.7.beta1 (66) was used to infer per-gene nucleotide diversity, π, both from aligned pan-genome gene sequences. Genomegamap v.1.0.1 (67) (Wilson and CRyPTIC Consortium 2020) was used to infer maximum-likelihood estimates of each aligned gene average d_N/d_S for each gene in the pangenome of the entire collection. In keeping with the method as described, only genes with an estimated nucleotide diversity (theta) value greater than 0.005 were considered robust enough estimates for further interpretation, leading to successful estimates for only 6853 genes. Genes with robust estimates of d_N/d_S greater than 2 were further analysed using the HYPHY package v.2.5.60 (68), specifically the FUBAR, FEL, and ABSREL statistical tests for pervasive gene-wide directional selection, specific sites subject to directional selection, and subsets of branches subject to directional selection respectively. Genes which had significant results to at least one of these HYPHY tests were manually investigated by searching the consensus and variant nucleotide sequences against the non-redundant protein database with tblastx, searching ESMFold-predicted protein structures against the AlphaFold, Uniprot, and Swiss-prot database using Foldseek. 3D structural alignments of consensus, variant, and reference protein structures were created for specific genes using TM-align.

Multidrug resistant (MDR) clusters in the combined collection were defined as PopPUNK clusters with a minimum of 50 isolates per cluster, of which at least 50% harboured resistance determinants to at least 3 of 7 antibiotic classes (aminoglycoside, β-lactam, phenicol, sulphonamide, tetracycline, trimethoprim, macrolide). For phylogenetic reconstruction of the global MDR clusters, assemblies of each cluster were mapped to the reference *H. influenzae* 86-028NP (61) using Snippy and phylogenies inferred using Gubbins (69).

For the NT isolates spanning 142 PopPUNK clusters in the Maela cohort, a permutation test of association between genetic lineages and pneumonia was performed by randomly shuffling "Status" labels (carriage and pneumonia). A two-sided Fisher's exact test with 10000 Monte Carlo replications was used.

DATA AVAILABILITY

All sequence data generated as part of this research is available in the ENA/SRA/DDBJ under project accession PRJEB41043. Metadata for both newly sequenced data and the systematic collection is available on microreact at the links as indicated in figure captions.

FUNDING

The funders had no role in study design, data collection, and analysis nor interpretation of the results, writing of the report, and the decision to submit the paper for publication. This work was supported by European Research Council (grant 742157; to JC), Wellcome Trust Grants (206194, 083735; to PT, MORU core award 220211; to PT; 098051 to Wellcome Sanger Institute), Trond Mohn Foundation (BATTALION grant; to JC, AKP, SM, and NM), Marie Skłodowska-Curie Actions (Grant 801133; to AKP, SM, and NM).

COMPETING FINANCIAL INTERESTS

The authors declare that they have no conflict of interest.

Bibliography

1. M. P. E Slack, A. W Cripps, K Grimwood, G. A Mackenzie, and M Ulanova. Invasive *Haemophilus influenzae* infections after 3 decades of hib protein conjugate vaccine use. *Clin. Microbiol. Rev.*, 34(3):e0002821, June 2021.
2. L Monasta, L Ronfani, F Marchetti, M Montico, L Vecchi Brumatti, A Bavcar, D Grasso, C Barbiero, and G Tamburlini. Burden of disease caused by otitis media: systematic review and global estimates. *PLoS One*, 7(4):e36226, April 2012.
3. J Van Eldere, M. P. E Slack, S Ladhani, and A. W Cripps. Non-typeable *Haemophilus influenzae*, an under-recognised pathogen. *Lancet Infect. Dis.*, 14(12):1281–1292, December 2014.
4. A. W Cripps. Nontypeable *Haemophilus influenzae* and childhood pneumonia. *P. N. G. Med. J.*, 53(3-4):147–150, September 2010.
5. M. P. E Slack. The evidence for non-typeable *Haemophilus influenzae* as a causative agent of childhood pneumonia. *Pneumonia (Nathan)*, 9(1):9, June 2017.
6. I De Schutter, E De Wachter, F Crokaert, J Verhaegen, O Soetens, D Piérard, and A Malfroot. Microbiology of bronchoalveolar lavage fluid in children with acute nonresponding or recurrent community-acquired pneumonia: identification of nontypeable *Haemophilus influenzae* as a major pathogen. *Clin. Infect. Dis.*, 52(12):1437–1444, June 2011.
7. S Chochua, V D'Acremont, C Hanke, D Alfa, J Shak, M Kilowoko, E Kyungu, L Kaiser, B Genton, K. P Klugman, and J. E Vidal. Increased nasopharyngeal density and concurrent carriage of *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Moraxella catarrhalis* are associated with pneumonia in febrile children. *PLoS One*, 11(12):e0167725, December 2016.
8. Pneumonia Etiology Research for Child Health (PERCH) Study Group. Causes of severe pneumonia requiring hospital admission in children without HIV infection from africa and asia: the PERCH multi-country case-control study. *Lancet*, 394(10200):757–779, August 2019.
9. A Carrera-Salinas, A González-Díaz, L Calatayud, J Mercado-Maza, C Puig, D Berbel, J Càmara, F Tubau, I Grau, M. Á Domínguez, C Ardanuy, and S Martí. Epidemiology and population structure of *Haemophilus influenzae* causing invasive disease. *Microb. Genom.*, 7(12), December 2021.
10. M De Chiara, D Hood, A Muzzi, D. J Pickard, T Perkins, M Pizza, G Dougan, R Rappuoli, E. R Moxon, M Soriani, and C Donati. Genome sequencing of disease and carriage isolates of nontypeable *Haemophilus influenzae* identifies discrete population structure. *Proc. Natl. Acad. Sci. U. S. A.*, 111(14):5439–5444, April 2014.
11. M. A Krisna, K. A Jolley, W Monteith, A Boubour, R. L Hamers, A. B Brueggemann, O. B Harrison, and M. C. J Maiden. Development and implementation of a core genome multilocus sequence typing scheme for *Haemophilus influenzae*. *Microb. Genom.*, 10(8), August 2024.
12. S. C Watts and K. E Holt. hicap: In silico serotyping of the *Haemophilus influenzae* capsule locus. *J. Clin. Microbiol.*, 57(6), June 2019.
13. S Mallawaarachchi, G Tonkin-Hill, A. K Pöntinen, J. K Calland, R. A Gladstone, S Arredondo-Alonso, N MacAlasdair, H. A Thorpe, J Top, S. K Sheppard, D Balding, N. J Croucher, and J Corander. Detecting co-selection through excess linkage disequilibrium in bacterial genomes. *NAR Genom. Bioinform.*, 6(2):lqae061, June 2024.
14. B. E Heaton, J Herrou, A. E Blackwell, V. H Wysocki, and S Crosson. Molecular structure and function of the novel BrnT/BrnA toxin-antitoxin system of *Brucella abortus*. *J. Biol. Chem.*, 287(15):12098–12110, April 2012.
15. C. O Buckee, K. A Jolley, M Recker, B Penman, P Kriz, S Gupta, and M. C Maiden. Role of selection in the emergence of lineages and the evolution of virulence in Neisseria meningitidis. *Proc Natl Acad Sci U S A*, 105(39):15082–15087, Sep 2008.
16. R. A Gladstone, S. W Lo, J. A Lees, N. J Croucher, A. J van Tonder, J Corander, A. J Page, P Marttinen, L. J Bentley, T. J Ochoa, P. L Ho, M du Plessis, J. E Cornick, B Kwambana-Adams, R Benisty, S. A Nzenze, S. A Madhi, P. A Hawkins, D. B Everett, M Antonio, R Dagan, K. P Klugman, A von Gottberg, L McGee, R. F Breiman, and S. D Bentley. International

genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine*, 43:338–346, May 2019.
17. P. H. C Kremer, J. A Lees, B Ferwerda, M. W Bijlsma, N MacAlasdair, A van der Ende, M. C Brouwer, S. D Bentley, and D van de Beek. Diversification in immunogenicity genes caused by selective pressures in invasive meningococci. *Microb. Genom.*, 6(9), September 2020.
18. N Takeuchi, K Kaneko, and E. V Koonin. Horizontal gene transfer can rescue prokaryotes from muller's ratchet: benefit of DNA from dead cells and population subdivision. *G3 (Bethesda)*, 4(2):325–339, February 2014.
19. M Hartfield, S. P Otto, and P. D Keightley. The maintenance of obligate sex in finite, structured populations subject to recurrent beneficial and deleterious mutation. *Evolution*, 66 (12):3658–3669, December 2012.
20. N MacAlasdair, M Pesonen, O Brynildsrud, V Eldholm, P. A Kristiansen, J Corander, D. A Caugant, and S. D Bentley. The effect of recombination on the evolution of a population of *Neisseria meningitidis*. *Genome Res*, Jun 2021.
21. C Chewapreecha, S. R Harris, N. J Croucher, C Turner, P Marttinen, L Cheng, A Pessia, D. M Aanensen, A. E Mather, A. J Page, S. J Salter, D Harris, F Nosten, D Goldblatt, J Corander, J Parkhill, P Turner, and S. D Bentley. Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet*, 46(3):305–309, Mar 2014.
22. S. W Lo, K Mellor, R Cohen, A. R Alonso, S Belman, N Kumar, P. A Hawkins, R. A Gladstone, A von Gottberg, B Veeraraghavan, K. L Ravikumar, R Kandasamy, S. A. J Pollard, S. K Saha, G Bigogo, M Antonio, B Kwambana-Adams, S Mirza, S Shakoor, I Nisar, J. E Cornick, D Lehmann, R. L Ford, B Sigauque, P Turner, J Moïsi, S. K Obaro, R Dagan, I Diawara, A Skoczyńska, H Wang, P. E Carter, K. P Klugman, G Rodgers, R. F Breiman, L McGee, S. D Bentley, C Muñoz-Almagro, E Varon, and Global Pneumococcal Sequencing Consortium. Emergence of a multidrug-resistant and virulent streptococcus pneumoniae lineage mediates serotype replacement after PCV13: an international whole-genome sequencing study. *Lancet Microbe*, 3(10):e735–e743, October 2022.
23. N. J Croucher. Immune interface interference vaccines: An evolution-informed approach to anti-bacterial vaccine design. *Microb. Biotechnol.*, 17(3):e14446, March 2024.
24. C Turner, P Turner, V Carrara, K Burgoine, S Tha Ler Htoo, W Watthanaworawit, N. P Day, N. J White, D Goldblatt, and F Nosten. High rates of pneumonia in children under two years of age in a south east asian refugee population. *PLoS One*, 8(1):e54026, January 2013.
25. P Turner, C Turner, A Jankhot, N Helen, S. J Lee, N. P Day, N. J White, F Nosten, and D Goldblatt. A longitudinal study of streptococcus pneumoniae carriage in a cohort of infants and their mothers on the Thailand-Myanmar border. *PLoS One*, 7(5):e38271, May 2012.
26. A Aziz, D. S Sarovich, E Nosworthy, J Beissbarth, A. B Chang, H Smith-Vaughan, E. P Price, and T. M Harris. Molecular signatures of non-typeable *Haemophilus influenzae* lung adaptation in pediatric chronic lung disease. *Front. Microbiol.*, 10:1622, July 2019.
27. A Carrera-Salinas, A González-Díaz, R. L Ehrlich, D Berbel, F Tubau, X Pomares, J Garmendia, M. Á Domínguez, C Ardanuy, D Huertas, A Marín, C Montón, J. C Mell, S Santos, and S Marti. Genetic adaptation and acquisition of macrolide resistance in *Haemophilus* spp. during persistent respiratory tract colonization in chronic obstructive pulmonary disease (COPD) patients receiving long-term azithromycin treatment. *Microbiol. Spectr.*, 11 (1):e0386022, February 2023.
28. L. F Collins, F. P Havers, A Tunali, S Thomas, J. A Clennon, Z Wiley, M Tobin-D'Angelo, T Parrott, T. D Read, S. W Satola, R. A Petit, 3rd, and M. M Farley. Invasive nontypeable *Haemophilus influenzae* infection among adults with HIV in metropolitan atlanta, georgia, 2008-2018. *JAMA*, 322(24):2399–2410, December 2019.
29. D. W Cleary, V. T Devine, D. E Morris, K. L Osman, R. A Gladstone, S. D Bentley, S. N Faust, and S. C Clarke. Pneumococcal vaccine impacts on the population genomics of non-typeable *Haemophilus influenzae*. *Microb. Genom.*, 4(9), September 2018.
30. A.-E Deghmane, E Hong, S Chehboub, A Terrade, M Falguières, M Sort, O Harrison, K. A Jolley, and M.-K Taha. 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.*, 79(1):7–14, July 2019.
31. M Diricks, T. A Kohl, N Käding, V Leshchinskiy, S Hauswaldt, O Jiménez Vázquez, C Utpatel, S Niemann, J Rupp, and M Merker. Whole genome sequencing-based classification of human-related *Haemophilus* species and detection of antimicrobial resistance genes. *Genome Med.*, 14(1):13, February 2022.
32. A Fernández-Calvet, B Euba, C Gil-Campillo, A Catalan-Moreno, J Moleres, S Martí, A Merlos, J. D Langereis, F García-Del Portillo, L. O Bakaletz, G. D Ehrlich, E. A Porsch, M Menéndez, J. C Mell, A Toledo-Arana, and J Garmendia. Phase variation in HMW1A controls a phenotypic switch in *Haemophilus influenzae* associated with pathoadaptation during persistent infection. *MBio*, 12(3):e0078921, June 2021.
33. A. C Fluit, J. R Bayjanov, B. J Benaissa-Trouw, M. R. C Rogers, M Díez-Aguilar, R Cantón, M. M Tunney, J. S Elborn, and M. B Ekkelenkamp. Whole-genome analysis of *Haemophilus influenzae* strains isolated from persons with cystic fibrosis. *J. Med. Microbiol.*, 71(8), August 2022.
34. W Fong, E Martinez, V Timms, A Ginn, T Nguyen, H Rahman, and V Sintchenko. Increase in invasive *Haemophilus influenzae* serotype a infections during the COVID-19 pandemic in new south wales, australia. *Pathology*, 56(5):696–701, August 2024.
35. A Gonzalez-Diaz, A Carrera-Salinas, M Pinto, M Cubero, A van der Ende, J. D Langereis, M. Á Domínguez, C Ardanuy, P Bajanca-Lavado, and S Marti. Comparative pangenome analysis of capsulated *Haemophilus influenzae* serotype f highlights their high genomic stability. *Sci. Rep.*, 12(1):3189, February 2022.
36. C.-H Ho, C.-W Chen, and P.-Y Su. Membranome-based identification of amino acid substitution in *Haemophilus influenzae* multidrug efflux pump HmrM for reduced chloramphenicol susceptibility. *Arch. Microbiol.*, 206(7):298, June 2024.
37. C Izydorczyk, B. J Waddell, R. B Weyant, M. G Surette, R Somayaji, H. R Rabin, J. M Conly, D. L Church, and M. D Parkins. The natural history and genetic diversity of *Haemophilus influenzae* infecting the airways of adults with cystic fibrosis. *Sci. Rep.*, 12(1):15765, September 2022.
38. R Kc, K. W. C Leong, B McEwan, J Lachowicz, N. M Harkness, S Petrovski, G Karupiah, and R. F O'Toole. Draft genome sequence of an isolate of nontypeable *Haemophilus influenzae* from an acute exacerbation of chronic obstructive pulmonary disease in tasmania. *Microbiol. Resour. Announc.*, 9(19), May 2020.
39. J. A Lees, P. H. C Kremer, A. S Manso, N. J Croucher, B Ferwerda, M. V Serón, M. R Oggioni, J Parkhill, M. C Brouwer, A van der Ende, D van de Beek, and S. D Bentley. Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. *Microb Genom*, 3(1):e000103, 01 2017.
40. B Lobb, M. C Lee, C. L McElheny, Y Doi, K Yahner, A Hoberman, J. M Martin, J. A Hirota, A. C Doxey, and N Shaikh. Genomic classification and antimicrobial resistance profiling of streptococcus pneumoniae and *Haemophilus influenza* isolates associated with paediatric otitis media and upper respiratory infection. *BMC Infect. Dis.*, 23(1):596, September 2023.
41. J. C Mell, C Viadas, J Moleres, S Sinha, A Fernández-Calvet, E. A Porsch, J. W St Geme, 3rd, C Nislow, R. J Redfield, and J Garmendia. Transformed recombinant enrichment profiling rapidly identifies HMW1 as an intracellular invasion locus in *Haemophilus influenza*. *PLoS Pathog.*, 12(4):e1005576, April 2016.
42. S Nag, G Larsen, J Szarvas, L. E. K Birkedahl, G. M Gulyás, W. J Ciok, T. M Lagermann, S Tafaj, S Bradbury, P Collignon, D Daley, V Dougnon, K Fabiyi, B Coulibaly, R Dembélé, G Nikiema, N Magloire, I. J Ouindgueta, Z. Z Hossain, A Begum, D Donchev, M Diggle, L Turnbull, S Lévesque, L Berlinger, K. K Sogaard, P. D Guevara, C. D Valderrama, P Maikanti, J Amlerova, P Drevinek, J Tkadlec, M Dilas, A Kaasch, H. T Westh, M. A Bachtarzi, W Amhis, C. E. S Salazar, J Villacis, M. A. D Lúzon, D. B Palau, C Duployez, M Paluche, S Asante-Sefa, M Moller, M Ip, I Mareković, A Pál-Sonnevend, C. E Cocuzza, A Dambrauskiene, A Macanze, A Cossa, I Mandomando, P Nwajiobi-Princewill, I. N Okeke, A. O Kehinde, I Adebiyi, I Akintayo, O Popoola, A Onipede, A Blomfeldt, N. E Nyquist, K Bocker, J Ussher, A Ali, N Ullah, H Khan, N. W Gustafson, I Jarrar, A Al-Hamad, V Luvira, W Paveenkittiporn, I Baran, J. C. L Mwansa, L Sikakwa, K Yamba, R. S Hendriksen, and F. M Aarestrup. Whole genomes from bacteria collected at diagnostic units around the world 2020. *Sci. Data*, 10(1):628, September 2023.
43. K. L Osman, J. M. C Jefferies, C. H Woelk, N Devos, T. G Pascal, M.-C Mortier, J.-M Devaster, T. M. A Wilkinson, D. W Cleary, S. C Clarke, and AERIS Study Group. Patients with chronic obstructive pulmonary disease harbour a variation of *Haemophilus* species. *Sci. Rep.*, 8(1):14734, October 2018.
44. M Pinto, A González-Díaz, M. P Machado, S Duarte, L Vieira, J. A Carriço, S Marti, M. P Bajanca-Lavado, and J. P Gomes. Insights into the population structure and pan-genome of *Haemophilus influenzae*. *Infect. Genet. Evol.*, 67:126–135, January 2019.
45. C. C Potts, N Topaz, L. D Rodriguez-Rivera, F Hu, H.-Y Chang, M. J Whaley, S Schmink, A. C Retchless, A Chen, E Ramos, G. H Doho, and X Wang. Genomic characterization of *Haemophilus influenzae*: a focus on the capsule locus. *BMC Genomics*, 20(1):733, October 2019.
46. C. C Potts, L. D Rodriguez-Rivera, A. C Retchless, S. A Buono, A. T Chen, H Marjuki, A. E Blain, and X Wang. Antimicrobial susceptibility survey of invasive *Haemophilus influenzae* in the united states in 2016. *Microbiol. Spectr.*, 10(3):e0257921, June 2022.
47. E. P Price, D. S Sarovich, E Nosworthy, J Beissbarth, R. L Marsh, J Pickering, L.-A. S Kirkham, A. D Keil, A. B Chang, and H. C Smith-Vaughan. *Haemophilus influenzae*: using comparative genomics to accurately identify a highly recombinogenic human pathogen. *BMC Genomics*, 16(1):641, August 2015.
48. A. R Rebelo, T Ibfelt, V Bortolaia, P Leekitcharoenphon, D. S Hansen, H. L Nielsen, S Ellermann-Eriksen, M Kemp, B. L Røder, N Frimodt-Møller, T. S Søndergaard, J. E Coia, C Østergaard, M Pedersen, H Westh, and F. M Aarestrup. One day in denmark: Nationwide point-prevalence survey of human bacterial isolates and comparison of classical and whole-genome sequence-based species identification methods. *PLoS One*, 17(2): e0261999, February 2022.
49. D. J Roach, J. N Burton, C Lee, B Stackhouse, S. M Butler-Wu, B. T Cookson, J Shendure, and S. J Salipante. A year of infection in the intensive care unit: Prospective whole genome sequencing of bacterial clinical isolates reveals cryptic transmissions and novel microbiota. *PLoS Genet.*, 11(7):e1005413, July 2015.
50. H.-C Slotved, T. B Johannesen, M Stegger, and K Fuursted. Evaluation of molecular typing for national surveillance of invasive clinical *Haemophilus influenzae* isolates from denmark. *Front. Microbiol.*, 13:1030242, November 2022.
51. Y Terrat, L Farnaes, J Bradley, N Tromas, and B. J Shapiro. 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.*, 6(4), April 2020.
52. N Topaz, R Tsang, A.-E Deghmane, H Claus, T.-T Lâm, D Litt, M. P Bajanca-Lavado, M Pérez-Vázquez, D Vestrheim, M Giufrè, A Van Der Ende, O Gaillot, A Kuch, M McElligott, M.-K Taha, and X Wang. Phylogenetic structure and comparative genomics of multi-national invasive *Haemophilus influenzae* serotype a isolates. *Front. Microbiol.*, 13:856884, March 2022.
53. R. S. W Tsang, M Shuel, T Ahmad, K Hayden, N Knox, G Van Domselaar, L Hoang, G. J Tyrrell, J Minion, P Van Caeseele, J. V Kus, M Ulanova, B Lefebvre, D Haldane, R Garceau, G German, G Zahariadis, B Hanley, K Kandola, and M Patterson. Whole genome sequencing to study the phylogenetic structure of serotype a *Haemophilus influenzae* recovered from patients in canada. *Can. J. Microbiol.*, 66(2):99–110, February 2020.
54. S. C Watts, L. M Judd, R Carzino, S Ranganathan, and K. E Holt. Genomic diversity and antimicrobial resistance of *Haemophilus* colonizing the airways of young children with cystic fibrosis. *mSystems*, 6(4):e0017821, August 2021.
55. S Arredondo-Alonso, A. K Pöntinen, F Cléon, R. A Gladstone, A. C Schürch, P. J Johnsen, Ø Samuelsen, and J Corander. A high-throughput multiplexing and selection strategy to complete bacterial genomes. *Gigascience*, 10(12), December 2021.
56. D. E Wood and S. L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46, 2014.
57. A. J Page, N De Silva, M Hunt, M. A Quail, J Parkhill, S. R Harris, T. D Otto, and J. A Keane. Robust high-throughput prokaryote de novo assembly and improvement pipeline for illumina data. *Microbial genomics*, 2(8), 2016.
58. J. A Lees, S. R Harris, G Tonkin-Hill, R. A Gladstone, S. W Lo, J. N Weiser, J Corander, S. D Bentley, and N. J Croucher. Fast and flexible bacterial genomic epidemiology with poppunk. *Genome research*, 29(2):304–316, 2019.
59. M Feldgarden, V Brover, N Gonzalez-Escalona, J. G Frye, J Haendiges, D. H Haft, M Hoff-

mann, J. B Pettengill, A. B Prasad, G. E Tillman, G. H Tyson, and W Klimke. AMRFinderPlus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci. Rep.*, 11(1):12728, June 2021.

60. M. C Maiden, J. A Bygraves, E Feil, G Morelli, J. E Russell, R Urwin, Q Zhang, J Zhou, K Zurth, D. A Caugant, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95(6):3140–3145, 1998.
61. A Harrison, D. W Dyer, A Gillaspy, W. C Ray, R Mungur, M. B Carson, H Zhong, J Gipson, M Gipson, L. S Johnson, L Lewis, L. O Bakaletz, and R. S Munson, Jr. Genomic sequence of an otitis media isolate of nontypeable *Haemophilus influenzae*: comparative study with h. influenzae serotype d, strain KW20. *J. Bacteriol.*, 187(13):4627–4636, July 2005.
62. A. J Page, B Taylor, A. J Delaney, J Soares, T Seemann, J. A Keane, and S. R Harris. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genom.*, 2 (4):e000056, April 2016.
63. M. N Price, P. S Dehal, and A. P Arkin. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, 5(3):e9490, March 2010.
64. G Tonkin-Hill, N MacAlasdair, C Ruis, A Weimann, G Horesh, J. A Lees, R. A Gladstone, S. W Lo, C Beaudoin, R. A Floto, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol*, 21(1):180, 07 2020.
65. R Mostowy, N. J Croucher, C. P Andam, J Corander, W. P Hanage, and P Marttinen. Efficient inference of recent and ancestral recombination within bacterial populations. *Molecular biology and evolution*, 34(5):1167–1182, 2017.
66. K. L Korunes and K Samuk. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol Ecol Resour*, 21(4):1359–1368, May 2021.
67. D. J Wilson and CRyPTIC Consortium. GenomegaMap: Within-species genome-wide dN/dS estimation from over 10,000 genomes. *Mol. Biol. Evol.*, 37(8):2450–2460, August 2020.
68. S. L. K Pond and S. V Muse. HyPhy: hypothesis testing using phylogenies. In *Statistical methods in molecular evolution*, pages 125–181. Springer, 2005.
69. N. J Croucher, A. J Page, T. R Connor, A. J Delaney, J. A Keane, S. D Bentley, J Parkhill, and S. R Harris. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using gubbins. *Nucleic acids research*, 43(3):e15–e15, 2014.

Supplementary Figures:

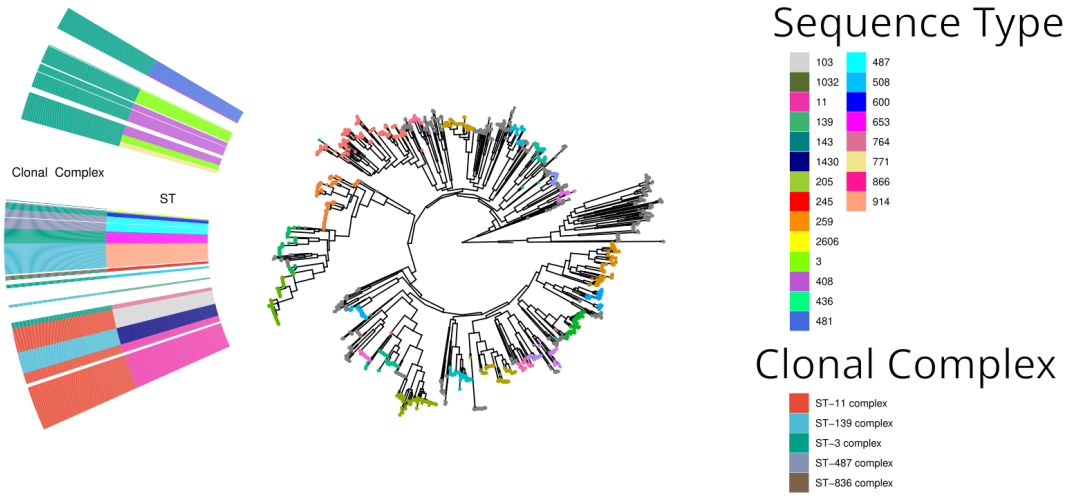
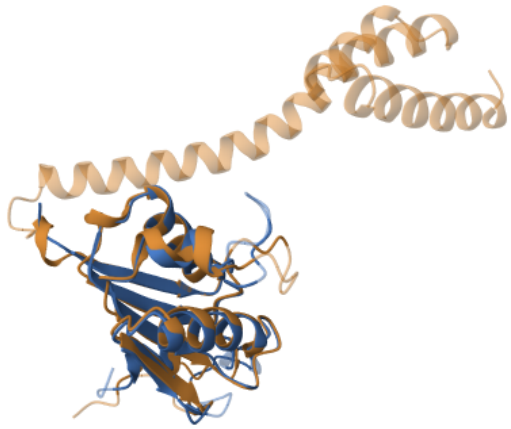
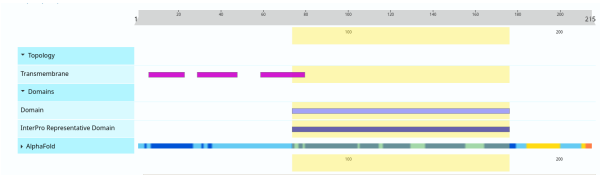


Fig. S1. Clonal Complexes(CC) CC11, CC139, CC3, CC487, and CC836 plotted with constituent sequence types on the core distance phylogeny of all global isolates. The 20 largest PopPUNK clusters of greater than 50 isolates are coloured on the tips of the phylogeny, with smaller clusters in grey. This figure clearly demonstrates that the ST-3, ST-11, and ST-139 complexes are paraphyletic, consisting of highly diverged isolates. An interactive online phylogeny, with all sequence types and Clonal Complexes, as well as additional metadata is available at the following link: https://microreact.org/project/oMm8PFCoG2429JwiDBpdru-maela-h-influenzae

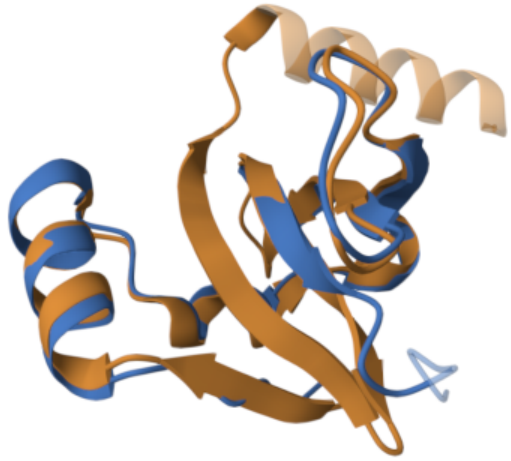


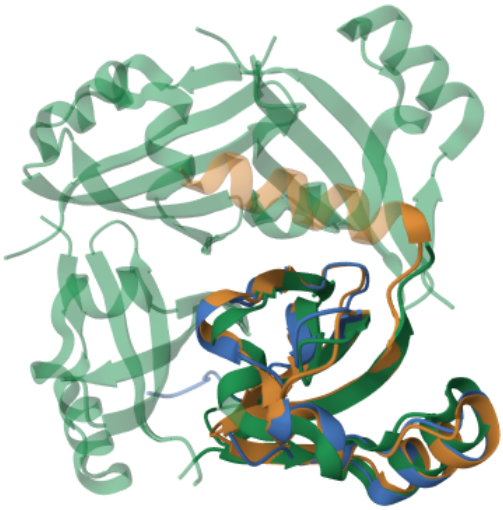


(a) Structural alignment of variant under selection in Maela data (Blue) and wild type variant (Gold), with the large domain composed of alpha helices at the C-terminus deleted in the Maela variant.

(b) Interpro scan domain annotation of the wild-type protein indicating that the deleted C-terminus domain in the Maela data is transmembrane

Fig. S2. Unnamed Gluconate Transporter (pangenome ID: group_3361) Identified as being under selection in some Maela isolates





(a) Structural alignment of the structural variant found in Maela isolates (blue) compared with the wild type *H. influenzae* structure (gold). The variant under selection contains some amino acid changes, as well as a small deletion, leading to the absence of an alpha helix when compared to the wild type *H. influenzae* structure.

(b) Structural alignment of both *H. influenzae* BrnT genes compared to the reference structure from *Brucella abortus*.

Fig. S3. BrnT toxin (pangenome ID: group_4059) variant includes an altered alpha helix domain