Genome sequencing of disease and carriage isolates of nontypeable *Haemophilus influenzae* identifies discrete population structure

Matteo De Chiara[a], Derek Hood[b,c], Alessandro Muzzi[a], Derek J. Pickard[d], Tim Perkins[e], Mariagrazia Pizza[a], Gordon Dougan[d], Rino Rappuoli[a,1], E. Richard Moxon[f], Marco Soriani[a], and Claudio Donati[a,g,1]

[a]Novartis Vaccines, 53100 Siena, Italy; [b]Nuffield Department of Clinical Medicine, John Radcliffe Hospital, University of Oxford, Headington, Oxford OX3 9DU, United Kingdom; [c]Molecular Genetics Unit, Medical Research Council Harwell, Oxfordshire OX11 0RD, United Kingdom; [d]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom; [e]School of Pathology and Laboratory Medicine (M504), University of Western Australia, Crawley, WA 6009, Australia; [f]Department of Paediatrics, Medical Sciences Division, John Radcliffe Hospital, University of Oxford, Headington, Oxford OX3 9DU, United Kingdom; and [g]Department of Computational Biology, Research and Innovation Centre, Fondazione Edmund Mach, 38010 San Michele all'Adige, Trento, Italy

Contributed by Rino Rappuoli, February 25, 2014 (sent for review November 14, 2013)

One of the main hurdles for the development of an effective and broadly protective vaccine against nonencapsulated isolates of *Haemophilus influenzae* (NTHi) lies in the genetic diversity of the species, which renders extremely difficult the identification of cross-protective candidate antigens. To assess whether a population structure of NTHi could be defined, we performed genome sequencing of a collection of diverse clinical isolates representative of both carriage and disease and of the diversity of the natural population. Analysis of the distribution of polymorphic sites in the core genome and of the composition of the accessory genome defined distinct evolutionary clades and supported a predominantly clonal evolution of NTHi, with the majority of genetic information transmitted vertically within lineages. A correlation between the population structure and the presence of selected surface-associated proteins and lipooligosaccharide structure, known to contribute to virulence, was found. This high-resolution, genome-based population structure of NTHi provides the foundation to obtain a better understanding, of NTHi adaptation to the host as well as its commensal and virulence behavior, that could facilitate intervention strategies against disease caused by this important human pathogen.

NTHi | genomics | population genetics

The Gram-negative bacterium *Haemophilus influenzae* colonizes the human nasopharynx and can cause a spectrum of diseases (1). Members of this species can be separated into those that are encapsulated and those that do not express a capsule, so-called nontypeable *H. influenzae* (NTHi) (2). Encapsulated strains belong to one of six distinct capsular serotypes (a, b, c, d, e, and f) of which type b strains are notoriously associated with invasive disease (3). NTHi are associated with common pediatric diseases, including otitis media (OM) (4, 5), and with exacerbations of chronic obstructive pulmonary disease (COPD) in adults (6).

Although capsule-based vaccines against serotype b strains exist, NTHi vaccine candidates containing outer-membrane proteins have been unsuccessful due to their inability to induce functional antibodies to epitopes representative of the phenotypic variation within the population, resulting in poor coverage against heterologous strains (7–9). To devise containment strategies based on vaccination, it is therefore essential to characterize the population structure of the NTHi strains and their genomic variability. Classification schema based on ribotyping (10), multilocus enzyme electrophoresis (11, 12), and multilocus sequence typing (MLST) (13–15) have shown that isolates of encapsulated *H. influenzae* could be classified into a small number of monophyletic lineages, with reduced diversity (12, 16) and genetically distinct from NTHi strains, that constitute the vast majority of the circulating population (11). Despite these efforts, there is still a substantial lack of knowledge regarding the structure of the NTHi population, mainly attributable to the impact that homologous recombination has on the evolution of the genomes of this pathogen, which is higher in NTHi compared with capsulated strains (15). So far, data on isolates from carriers and those with disease have shown little correlation between MLST typing and the clinical source or the geographical origin of the strains studied (17).

Whole-genome sequencing can be used to characterize the population structure of large collections of isolates of bacterial pathogens (18–20) and to study the microevolution of virulent lineages (21, 22). Here, we use whole-genome sequencing of NTHi isolates of diverse clinical and geographical origin to assess population structure. Analysis of single nucleotide polymorphisms (SNPs) revealed six statistically supported clusters of isolates that correlated with the composition of the accessory genome. Our data lay the foundation for a comprehensive definition of the population structure of NTHi that can underpin the development of strategies to fight NTHi-associated disease.

Significance

Several human pathogens exploit genomic variability to adapt to the host environment. Genome sequencing of collections of isolates and classification of strains according to their genomic content are pivotal to the formulation of vaccines able to elicit broad protection. We sequenced a collection of carriage and disease isolates of nontypeable *Haemophilus influenzae*, a component of the microbial flora of the upper respiratory tract that can cause a spectrum of diseases, including otitis media and meningitis. We identified distinct evolutionary clades that correlate with the presence of selected surface-associated proteins and virulence determinants. The high-resolution definition of the population structure of nontypeable *Haemophilus influenzae* allowed by whole-genome sequencing will be key for the development of efficacious containment strategies for this important human pathogen.

Author contributions: M.P., G.D., R.R., E.R.M., M.S., and C.D. designed research; D.J.P. and T.P. performed research; D.H. and E.R.M. contributed new reagents/analytic tools; M.D.C., D.H., A.M., G.D., and C.D. analyzed data; and M.D.C., D.H., A.M., G.D., R.R., E.R.M., M.S., and C.D. wrote the paper.

Conflict of interest statement: M.D.C., A.M., M.P., R.R., and M.S. are full-time employees of Novartis Vaccines. C.D. was a full-time Novartis Vaccines employee when research was performed.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. E-mail: rino.rappuoli@novartis.com or claudio.donati@fmach.it.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1403353111/-/DCSupplemental.

Results

A Global Collection of 97 Strains, Including Those with Completed Genome Sequences, Is Representative of the Genetic Diversity of NTHi. The collection was composed of 97 NTHi isolates (89 sequenced for this study and 8 publicly available) (Dataset S1). Of the sequenced isolates, 48 were from Finland (32 obtained by tympanocentesis from individuals with otitis media and 16 from nasopharyngeal cultures of healthy individuals), 19 from Spain (from patients with COPD), and 22 from a global collection [5 from the United Kingdom, 1 from the United States, 2 from South Korea, 1 from South Africa, 3 from Papua New Guinea (PNG), 1 from Malaysia, 4 from Iceland, 1 from Ghana, 1 from China, 1 from Australia, and 2 for which the geographical locus of isolation was not known]. Complete and draft genome sequences of 8 isolates of encapsulated and noncapsulated *H. influenzae* [86-028NP (23, 24), F3031 and F3047 (25), PittEE, PittGG, R3021, and R2866 (26), and Rd_KW20 (27)] were downloaded from the National Center for Biotechnology Information Web site in March 2013. Assembly of the sequenced genomes produced draft sequences with a number of contigs ranging from 18 to 55 (average 31). The average amount of sequence assembled for each isolate was 1.8 Mb (1.73–1.99 Mb).

For an initial assessment of the genetic diversity of the collection of NTHi isolates, we used MLST to determine their sequence type (ST) and compared the distribution of STs to that of the MLST database. Seventy-five isolates were assigned to a known ST whereas 12 had MLST profiles not present in the database and 10 could not be assigned a complete MLST profile due to the draft status of the sequences. In Fig. 1*A*, we show a minimum spanning tree (MST) obtained using the goeBurst algorithm (28) based on the allelic profiles of all isolates included in the MLST database, plus the 12 profiles identified here. Capsulated strains were concentrated into a small number of serotype-specific clusters that probably originated by the expansion of single, successful clones that acquired the ability to synthesize a specific type of polysaccharide capsule (15). Two groups of isolates of serotypes a and b could be identified whereas isolates of serotypes c, d, e, and f were each in a single, serotype-specific group within the tree. The population of NTHi was composed of a heterogeneous group of isolates that constituted most of the sparse structure of the MST.

On the same representation, we highlighted the STs of the isolates in the sequenced collection (Fig. 1*B*). The collection showed no bias in term of MLST profile, covering almost uniformly the entire MST. In some cases the sequenced NTHi isolates had STs that are usually associated with capsulated strains (Fig. 1 *A* and *B* and Dataset S1). Although in all other isolates the nontypeable phenotype was due to a complete deletion of the capsule biosynthesis locus, in those isolates, the lack of capsule expression was related to the deletion of the *bexA* gene, an energy-coupling component of the capsule polysaccharide export apparatus. The rest of the capsule locus was similar to that of the corresponding capsulated isolates, suggesting that the loss of the encapsulated phenotype might be the result of a recent event.

Sequencing of 97 Isolates Identifies a Population Structure Defined by Six Clades. Draft and complete genome sequences were aligned against the chromosome of isolate 86–026NP (23) used as reference, and the core genome: i.e., the portion of the reference genome that could be aligned against all of the other sequences was determined to be 1,134,504 bp. This portion represented on average 63% of the genome of any *H. influenzae* isolate. From these alignments, we generated a list of 149,214 polymorphic sites, 134,702 of which were biallelic. To this dataset, we applied the discriminant analysis of principal components (DAPC) method (29) to define the structure of the population of NTHi and to identify groups of related strains. The Bayesian information criterion (BIC) (Fig. S1) supported the partitioning of this set of strains into six groups that were clearly distinct in a scatter plot of the three principal components of the DAPC (Fig. 2).

To compare this partitioning to standard phylogenetic analyses, we built a neighbor-joining (NJ) phylogenetic tree using the SNPs (Fig. 3). All of the branches were supported by bootstrap values >90%. Except for the position of isolate RM603375, each group identified a monophyletic clade in the tree (clades I–VI in Fig. 3). The classification of the isolates into clades could not be predicted on the basis of the MLST data (Fig. S2). In particular, whereas isolates of clades II–V had a certain degree of correlation with the structure of the MST, isolates of clades I and VI were scattered apparently in a random manner.

Strains of Common Clinical or Geographical Source Do Not Group. To test whether there was evidence of geographical isolation between subpopulations in our collection, we overlaid information on the source of the isolate on the NJ phylogenetic tree (Fig. S3*A*). There was no absolute separation of the strains according to geography. In particular, Finnish and Spanish isolates were present in all groups and scattered throughout the whole tree. However, it was clear that clade VI, the largest subpopulation in the collection, was more heterogeneous in terms of the origin of the strains and contained a majority of the isolates that were neither from Finland nor Spain. Clade VI also included all except one of the NTHi isolates that could be classified as serotype b on the basis of the presence of the capsule locus sequence; these (Fig. 3) formed a monophyletic branch within clade VI, together with the two Brazilian purpuric fever isolates F3031 and F3047. We tested whether there was association between phylogenetic analysis and the clinical source of the strain and found that, in many cases, closely related strains were both from asymptomatic carriage and different diseases (Fig. S3*B*).

Phylogeny, Based on the Dispensable Genome, Separates Isolates into Groups That Closely Correlate with Clades. We predicted a total of 171,607 ORFs, with an average of 1,769 ORFs per genome (Dataset S1). We identified 2,852 gene families that were present in a single genome plus 3,328 gene families containing at least two sequences. Of these, 935 were core: i.e., present in all genomes. If we relaxed the definition of core genome by including also the gene families that were absent in one genome, we found a core genome composed of 1,207 ORFs, larger than what was found using a hybridization array (30). Given the draft status of most of the sequences, we consider this latter figure a better estimate of the size of the core genome of *H. influenzae*.

The distribution of the gene families in the *H. influenzae* population had the characteristic U-shape found in other bacterial species (18, 20), with the majority of the noncore (or dispensable)

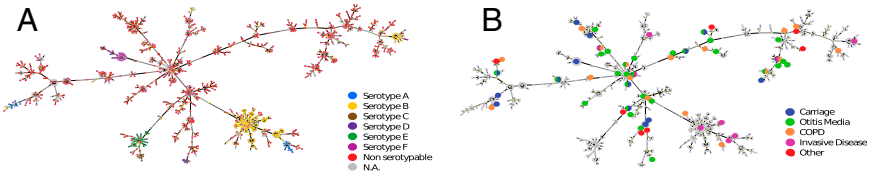

Fig. 1. (*A*) Minimum spanning tree (MST) based on the allelic profiles of all isolates present in the MLST database. Each node is an ST, and it is colored according to the serotypes of the isolates. The size of the nodes is proportional to the number of isolates. By applying the traditional definition of clonal complex (CC) (i.e., six out of seven identical MLST genes), we found 464 different CCs, of which 151 included more than one ST. Most capsulated isolates were concentrated in a small number of CCs whereas the nontypeable STs were dispersed in a high number of CCs, many of which included a single ST. (*B*) STs of the sequenced isolates colored according to disease. Although there is some association between serotype B and invasive disease, carriage, COPD, and otitis media are associated with variable capsular genotype.

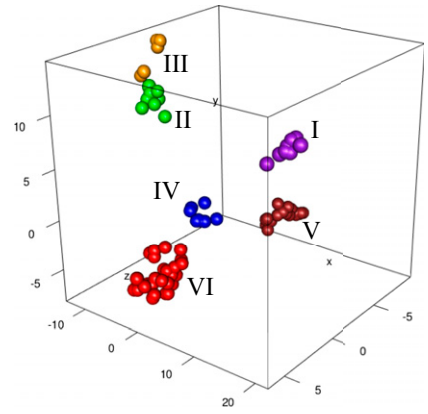

Fig. 2. A 3D scatterplot of the first three principal components of the DAPC of the polymorphic sites. Each dot is one isolate, colored according to the classification into one of the six clusters. Clusters II and III are closely related whereas all other clusters are clearly separated by DAPC.

genes either sporadically absent from the sequenced isolates or present only in a small subset of the isolates. To test whether the composition of the dispensable genome supported the partitioning of the collection in clades I–VI, we performed a DAPC analysis on the presence/absence profiles. The results are reported in Fig. 4, where the dots are colored according to the classification of isolates into clades I–VI. There was a clear separation of the isolates into groups that closely correlated with the clades, with little overlap between the groups. The only exceptions were clades II and III, which were not separated by this analysis. Isolates from clades II and III were the most closely related also in the SNP DPCA analysis (Fig. 2), probably an indication of their relatively recent diversification. The distribution of the families of orthologous genes in the sequenced isolates (Fig. S4) identifies the existence of groups of genes that are specific for clades I–V and that correlate with the phylogenetic reconstruction shown in Figs. 2 and 3 and Dataset S2. Clade VI appears to be less resolved, supporting the fact that this group of isolates contains a substructure that will become evident only when a larger panel of isolates is sequenced.

NTHi Population Structure Correlates with the Presence of Lipooligosaccharide Biosynthesis Genes and Surface-Associated Proteins Known to Contribute to Virulence. Virulence and colonization factors are important determinants of disease caused by *H. influenzae*, but, apart from capsule, little is known about the association between population structure and their expression (31). We searched through the available genomes for the presence of selected virulence genes, to assess whether the classification of strains into clades could be correlated to the "infectious potential" of the respective isolates. We identified *Haemophilus* adhesion and penetration protein (Hap), high molecular weight proteins 1 and 2 (HMW1/HMW), *Haemophilus influenzae* adhesin/*Haemophilus* surface fibrils (Hia/hsf), and the pili proteins Hif as good candidates for this analysis, as they are expressed in different combinations by encapsulated and noncapsulated isolates (32, 33). The *hap* gene was present in all isolates, with an identity always over 97%. The *hia*/*hsf*, *hmw1*/*hmw2*, and *hif* genes were present in a subset of the isolates, and their distribution was correlated with the clades (Fig. 5 and Table S1). *hia*/*hsf* was associated with clades VI and I (log odds ratio = 4.53, *P* value = 0.0007); the four strains with remnants of the serotype b capsule locus contained the *hsf* gene (*hsf* is an allelic variant of *hia*). *hmw1*/*hmw2* were associated with clades II, III, IV, and V (log odds ratio = 7.05, *P* value = 1.0×10^{-11}). *hif* was associated with clade VI (log odds ratio = 3.38, *P* value = 6.36×10^{-8}). A peculiar *hif* locus showing the large deletion typical of serotype f strains (32) was present in the f* isolate in clade I.

A second form of the immunoglobulin A1 peptidase gene with high homology to the meningococcal *igaA* gene, and known as *igaB,* is found in a subset of NTHi isolates (34) and has a reported association with COPD (35, 36). The *igaB* gene was strongly associated with clades II and IV (log odds ratio = 7.61, *P* value = 2.07×10^{-6}) whereas we could not confirm the correlation with COPD (log odds ratio=-0.39, *P* value = 0.27). Outer-membrane protein P6 is a peptidoglycan-associated lipoprotein. Since its discovery, it has been considered a vaccine candidate for a number of features like its immunogenicity (37) and its conservation and ubiquity (38). The OMP P6 protein was present in all isolates of our collection and showed a high degree of conservation. The most common peptide was shared by 81 out of 97 strains. We found 4 variants distributed in 16 strains. The mutations were as follows: A11T (3 strains); A32V (2 strains); A114T (9 strains); and T125A (2 strains). All of the mutations from the dominant allele were segregated within single clades and shared by closely related strains. The A11T and A32V mutations were found in clade VI whereas the T125A mutation was in clade V. The A114T mutation was specific for all except one (Hi16) isolate of clade I.

Lipooligosaccharide (LOS) is the major glycolipid on the *H. influenzae* cell surface and plays multiple roles in host interaction and disease. There is significant heterogeneity of LOS structure across NTHi strains dependent upon the biosynthetic gene repertoire. In Table 1, we summarized the information on the loci encoding genes involved in LOS biosynthesis. The *hmg* locus, encoding a sialylated four-sugar unit that enhances resistance to host immune killing, was associated with clades II, III, IV, and VI. The *lpt3* gene, required for phase variable expression of a phosphoethanolamine (PEtn) linked to the LOS inner core that alters bacterial serum resistance, was associated with clade V.

Thus, for both surface-associated proteins and LOS, some correlation between population structure and the presence of selected antigens can be identified.

Discussion

Early studies demonstrated that capsulated *H. influenzae* strains belong to a small number of nearly clonal lineages (12), each expressing a specific type of capsule and each genetically distinct from NTHi, the nonencapsulated family of this organism (11, 39). Due to their variability and the higher impact of recombination, molecular typing methods have so far failed to identify a population structure for NTHi strains (15). This difficulty may account for the slow development of an efficacious vaccine against NTHi-associated disease, which, despite a number of promising candidates (40–44), has not been supported by robust clinical data (45). Knowledge of the population structure is instrumental to generate preclinical data supporting the capacity of selected antigens to induce a cross protecting immune response against epidemiologically relevant strains.

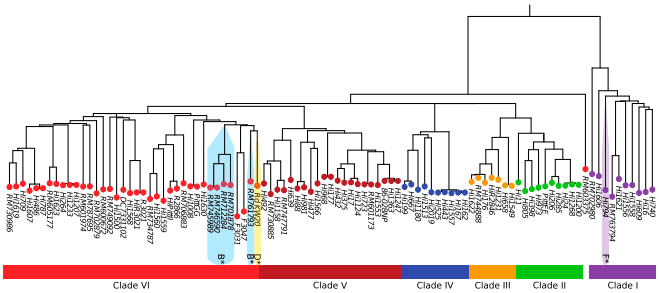

Fig. 3. Neighbor-joining phylogenetic tree of the sequenced strains. The tree was rooted using *Haemophilus haemoliticus* as outgroup. The tips are colored according to the assigned cluster in the DAPC analysis. The DAPC classification identifies six monophyletic clades. Note the position of the strain RM603375, the only discordance between the DAPC and the phylogenetic analysis. Shades indicate isolates that have most of the capsule biosynthetic locus intact although they do not express capsule. The serotype inferred by homology of the capsule operon is indicated (B*, D*, F*).

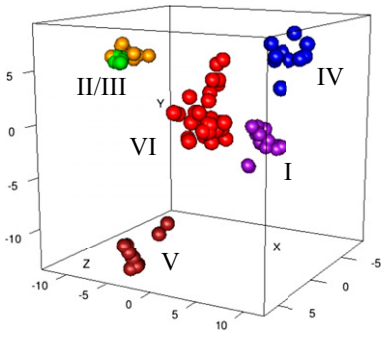

Fig. 4. A 3D scatterplot of the first three principal components of the DAPC of the presence/absence profiles. Points are colored according to the clades defined using the core SNPs (Figs. 2 and 3). Although clades I and IV–VI are clearly separated by the analysis of the presence/absence profiles, isolates of clades II and III are not separable.

To address this problem and explore whether it is possible to identify groups of isolates that are ancestrally related, and the variability within these groups in terms of virulence-related factors and potential antigens, we sequenced the genomes of a collection of NTHi clinical isolates. The sequence data of core and accessory genome gave us the means to reduce the background noise in the phylogenetic reconstruction and to discriminate the different lineages. We identified six distinct clades supported by population genetics and molecular phylogenetic analysis based on the core alignment, and correlated with the distribution of the noncore gene families. This concordance points toward a prevalently clonal nature of the NTHi population (and indeed of the entire species), which is made up of well-defined lineages, or clades. The distribution of the virulence factors also identifies examples of horizontal gene transfer, highlighting the role of interstrain exchanges of DNA in the genetic variability of this naturally competent species (46).

One of the fundamental questions in molecular epidemiology is the definition of typing methods that correlate with the disease induced by a pathogen. In some cases, this result is attained by using known virulence factors as molecular markers, as in the case of the capsular serotype for *Streptococcus pneumoniae* (47). In others, like the meningococcus, certain lineages (hypervirulent clusters) associate with disease (48). We asked whether, as in the case of the Brazilian purpuric fever, which is caused by a well-defined *H. influenzae* clone (biogroup Aegyptius) (49), there was a correlation between the clades and the pathologies represented in our strain collection. Although we found a clear correlation between known virulence factors and clades, COPD, OM, and asymptomatic carriage isolates were present in most of the clades, and no clear-cut association was identified with disease. This result supports the importance of host-related factors interacting with the genotype of the infecting pathogen.

The six clades are a good representation of the known diversity of NTHi. However, we expect that more exhaustive sampling will reveal substructures within them, possibly related to the dynamics of spread of individual clones within well-defined geographical locations, similarly to what has been found for other pathogens (21, 50). Geographical structuring of the population was not evident in our collection. The identification of clades that define a discrete population structure is a first step toward characterizing NTHi species diversity and, despite the issues on intra- and interclade genetic variations, opens the door to rational strategies for the prevention, containment, and treatment of important human diseases, including OM and COPD.

Materials and Methods

Bacterial Strains, DNA Extraction, and Genome Sequencing. NTHi strains selected for genome sequencing (Dataset S1) were obtained from a collection of isolates archived in Oxford. Four main NTHi strain sets used in this study were (*i*) isolates from children with OM from Finland; (*ii*) carriage strains from healthy children in Finland; (*iii*) adults with COPD from Spain; and (*iv*) strains from diverse clinical situations and geographical origin isolated over a 20-y time period (46). The disease state of the host, dates, and geographical sites of isolation are included in Dataset S1.

Bacterial isolates were cultured from frozen stock, plated on brain heart infusion (BHI) medium supplemented with 10% (vol/vol) Levinthal base and 1% agar, and incubated at 37 °C. For DNA preparation, bacteria were cultured on BHI liquid supplemented with haemin (10 μg/mL) and NAD (2 μg/mL).

Strains were grown on BHI broth, and chromosomal DNA was isolated from bacteria using Qiagen columns as described by the supplier. The genomic DNA was sequenced using multiplex (12 separately indexed DNAs per lane) Illumina sequencing as described previously (22).

Genome Alignments and SNP Selection. We aligned all of the sequenced genomes to the reference complete genome of the strain 86-028NP using the program Nucmer (51). From these alignments, we generated a list of polymorphic sites. The list was filtered to include only sites in the core genome of NTHi: i.e., those regions of the reference strain 86-028NP that could be aligned against all other strains, yielding a set of 149,214 SNPs.

Identification of Genetic Clusters and Statistical Analysis. All statistical analysis was conducted using R, version 2.15.3 (52). The SNPs extracted from the genome alignment where filtered to extract the biallelic loci. This procedure yielded a sample of 134,703 SNPs that were analyzed using the DAPC technique (29). The analysis was performed using the *adegenet* 1.3-6 package (53, 54) of the statistical software R. DAPC implements a statistical method to describe genetic clusters in a way that maximizes the between-clusters variance while minimizing the within-cluster variance. The number of components retained in the initial principal component analysis to identify the number of clusters was 60, accounting for >90% of the sample variability. The number of genetic clusters was inferred using BIC (55). Statistical association tests were performed using the package vcd, version 1.2–13.

Phylogenetic Analysis. Phylogenetic analysis on the SNPs was conducted using Mega, version 5 (56). The evolutionary distances were computed using the maximum composite likelihood method (57). The evolutionary history was inferred using the neighbor-joining method (58). The tree was rooted using Path-O-Gen (http://tree.bio.ed.ac.uk/software/pathogen/), and the position of the root was confirmed using *Haemophilus haemoliticus* as outgroup.

Genome Assembly and Annotation. The genome sequences were assembled using Celera Assembler 7 (59). Each genome was assembled seven times using a different number of reads, from 400,000 reads up to a maximum of 1,000,000 total reads, and the assembly with the lowest number of separate contigs and the highest total number of assembled bases was chosen. The resulting coverage ranged from 30× to 80×, with an average of 50×. The draft genomes were annotated using a hybrid approach. First, the annotation of the complete and draft genomes downloaded from the web was transferred onto the newly sequenced genomes using RATT (60). To identify ORFs in regions that had no close homology to already-annotated strains, we performed a de novo ORFs prediction using Glimmer3 (61). Overlapping predictions from the two methods were merged if the ORFs shared the same

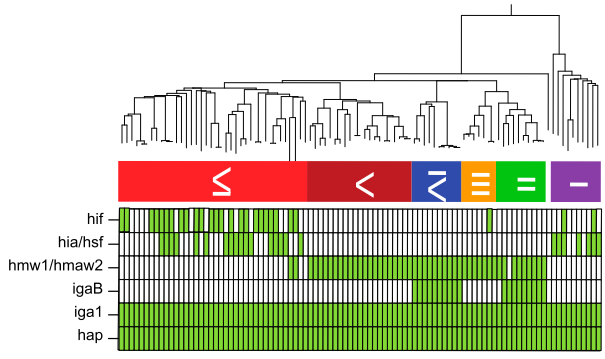

Fig. 5. Distribution of the *hif*, *hia/hsf*, *hmw1/hmw2*, *igaB*, *iga1*, and *hap* genes. Green marks isolates where the gene is present, white where it is absent. There is a good correlation between presence/absence of the *hif*, *hia/hsf*, *hmw1/hmw2*, and *igaB* genes and the six clades. Clades I and VI are relatively more heterogeneous whereas clades II–V are composed of a very homogeneous set of isolates.

Table 1. LOS biosynthetic loci

Strain	Clade	*hmg*	HepIV	*lic2C*	PCho	*lpsA*	*pgt1*	*lpt3*
Hi740	I	–	–	+	HepII	β1–2 Glc	–	–
Hi973	II	–	–	+	HepI	β1–2 Glc	+	GT
Hi285	II	+	–	+	HepI	Tr(β1–3 Glc)	+	GT
Hi206	II	+	–	+	HepI	β1–3 Glc	+	GT
Hi1268	II	+	–	+	HepI	β1–2 Glc	–	GT
Hi1200	II	+	–	+	HepII	β1–2 Glc	–	GT
R2846	III	+	HepIV DD	–	HepI	β1–2 Gal	+	GT
Hi658	III	+	–	+	HepIII	β1–2 Gal	+	GT
Hi176	III	+	HepIV DD	–	HepIII/IV	β1–2 Gal	+	GT
Hi1231	III	+	HepIV DD	–	HepIII	β1–2 Gal	+	GT
Hi667	IV	+	–	+	HepI	β1–3 Gal	+	–
Hi2019	IV	+	–	+	HepI	Tr(β1–3 Gal)	+	lpt3
Hi199	IV	+	–	+	HepI	β1–3 Gal	+	–
Hi167	IV	+	–	–	HepI	β1–2 Glc	+	GT
Hi162	IV	+	–	–	HepI	β1–2 Glc	+	GT
Hi1180	IV	+	–	+	HepI	β1–3 Gal	+	–
Hi981	V	–	HepIV DD	–	HepI	Tr(β1–2 Glc)	+	lpt3
Hi723	V	–	–	–	HepI	β1–2 Glc	+	lpt3
Hi639	V	–	hepIV (B1/B2)	–	HepIII	β1–2 Glc	+	lpt3
Hi477	V	–	HepIV DD	–	HepI	Tr(β1–2 Glc)	+	lpt3
Hi432	V	–	–	+	HepI	β1–2 Glc	+	lpt3
Hi375	V	–	–	–	HepI	β1–2 Glc	+	lpt3
Hi1363	V	–	–	–	HepI	β1–2 Glc	+	GT
Hi1247	V	–	–	–	HepI	β1–2 Glc	+	GT
Hi1158	V	–	HepIV DD	+	HepIII/IV	β1–2 Glc	+	lpt3
Hi1124	V	–	–	–	HepI	β1–2 Glc	+	lpt3
HiR3021	VI	+	hepIV(B1)	–	HepII	β1–2 Glc	–	–
HiPittII	VI	+	hepIV(B2)	–	HepII	β1–2 Glc	–	–
Hi486	VI	–	–	+	HepII	β1–3 Glc	–	–
Hi1233	VI	+	HepIV LD	–	HepI	β1–2 Glc	–	–
Hi1207	VI	+	HepIV LD	–	HepI	β1–2 Glc	–	–
Hi1008	VI	+	–	+	HepI	β1–2 Glc	–	–

The LOS genes/epitopes in this table are focused upon those that can distinguish between NTHi strains either through presence/absence or allelic polymorphisms that affect LOS phenotype. +, presence, –; absence; *hmg* +, represents the presence of representative genes from this nine-gene locus, consistent with expression of a four-sugar unit as part of an R1 extension from HepI. *HepIV* occurs in some strains as part of the oligosaccharide extension R1 and can be either an L,D- or D,D-heptose. *lic2C* initiates R2 elongation so, if present (+), then the LOS includes an extension from HepII. *PCho* is typically added to a hexose as the first sugar in oligosaccharide R1 (from HepI), R2 (from HepII), or R3 (from HepIII). In a minority of strains, a second PCho is present linked directly on HepIV. *lpsA* initiates R3 extension from HepIII; this sugar can be either a glucose (Glc) or galactose (Gal) in either one of two linkages. If Gal is added, then this sugar is terminal; Glc allows further chain elongation. *pgt1* encodes a glycosyltransferase of unknown function. For *lpt3*, Lpt3 and GT are alternative products encoded at the same genomic locus. Lpt3 adds phosphoethanolamine to HepIII, and the function of GT remains unknown.

stop codon, and the starting site of the longer ORF was retained. The annotated assembled draft genomes are available at ftp://ftp.sanger.ac.uk/pub/project/pathogens/Haemophilus/influenzae/NT_strains/.

Orthologous Genes. An all-against-all alignment search was performed using FASTA, v. 3.4t25 (62). Orthologous genes were identified using the reverse best-hit algorithm. Two genes were considered to be orthologous if they were reciprocally the best hit and they shared at least 50% of identity on the 50% of the length of the longer one. Families of orthologous genes were defined by determining the connected components in a graph where each individual protein was a node and the orthologous relationships were the links.

MLST. The public MLST database (http://haemophilus.mlst.net/) was downloaded (April 2013), providing 1,213 MLST profiles and the ST data for 1,983 strains. For the sequenced genomes, MLST sequences were extracted from the assemblies. The goeBurst algorithm (28) was used to generate an MST from the allelic profiles using Phyloviz (63). In this tree, each vertex represents a specific ST.

Genes Encoding for Virulence Factors. Sequences encoding known virulence factors were downloaded from GenBank and aligned against the genome sequences using FASTA, v. 3.4t25. The query sequence used for *hap* was U11024.1. For *hia/hsf*, both allelic variants were used (U38617.2/AY823627.1). For *hif*, the full locus (U19730.1) was used. *hap*, *hia/hsf*, *hif*, P6, and *hmw1/hmw2* were considered present if the alignment had sequence identity >70% on at least 60% of the query length. Isolate Hi740, which has remnants of a capsule biosynthetic locus of serotype f, had the typical serotype f *hif* locus containing a long deletion of *hifB* gene as well as promoter and the *hifC* 5′-end. For outer-membrane protein P6, we used the sequence NTHI0501 from the strain 86-028NP as query. Alleles were identified using the translated peptide sequences. HMW1 and HMW2 proteins are coded by different chromosomal loci, which are usually present in the same isolates, and show high degree of similarity. In particular, *hmw1A* and *hmw2A* genes are identical for the first 1,259 positions and 98% identical in the first 1,685 bases. A high level of conservation is also found in the last 740 bases (83% of identity). The overall identity is about 80% (64). The product of *hmw1B* and *hmw2B* is 99% identical, and the products of *hmw1C* and *hmw2C* show 97% of identity (31). For this reason, assembly of the *hmw1/hmw2* loci from short reads was problematic, and they were often found interrupted by the end of a contig. We used the full *hmw1* locus (U08876.1) as query. In case of sequence interruption caused by the end of the contig, the strain was considered hmw^+ if the aligned sequence was at least 2,000 bp.

1. Agrawal A, Murphy TF (2011) Haemophilus influenzae infections in the H. influenzae type b conjugate vaccine era. *J Clin Microbiol* 49(11):3728–3732.
2. Pittman M (1930) The "S" and "R" forms of Haemophilus influenzae. *Exp Biol Med* 27(4):299–301.
3. Pittman M (1931) Variation and type specificity in the bacterial species Hemophilus influenzae. *J Exp Med* 53(4):471–492.
4. Lim CT, Parasakthi N, Puthucheary SD (1994) Neonatal meningitis due to non-encapsulated Haemophilus influenzae in a set of twins: A case report. *Singapore Med J* 35(1):104–105.
5. Gkentzi D, Slack MP, Ladhani SN (2012) The burden of nonencapsulated Haemophilus influenzae in children and potential for prevention. *Curr Opin Infect Dis* 25(3): 266–272.
6. Moghaddam SJ, Ochoa CE, Sethi S, Dickey BF (2011) Nontypeable Haemophilus influenzae in chronic obstructive pulmonary disease and lung cancer. *Int J Chron Obstruct Pulmon Dis* 6:113–123.
7. Bolduc GR, et al. (2000) Variability of outer membrane protein P1 and its evaluation as a vaccine candidate against experimental otitis media due to nontypeable Haemophilus influenzae: An unambiguous, multifaceted approach. *Infect Immun* 68(8): 4505–4517.
8. Webb DC, Cripps AW (1999) Immunization with recombinant transferrin binding protein B enhances clearance of nontypeable Haemophilus influenzae from the rat lung. *Infect Immun* 67(5):2138–2144.
9. Webb DC, Cripps AW (2000) A P5 peptide that is homologous to peptide 10 of OprF from Pseudomonas aeruginosa enhances clearance of nontypeable Haemophilus influenzae from acutely infected rat lung in the absence of detectable peptide-specific antibody. *Infect Immun* 68(1):377–381.
10. Bruce KD, Jordens JZ (1991) Characterization of noncapsulate Haemophilus influenzae by whole-cell polypeptide profiles, restriction endonuclease analysis, and rRNA gene restriction patterns. *J Clin Microbiol* 29(2):291–296.
11. Musser JM, Barenkamp SJ, Granoff DM, Selander RK (1986) Genetic relationships of serologically nontypable and serotype b strains of Haemophilus influenzae. *Infect Immun* 52(1):183–191.
12. Musser JM, Granoff DM, Pattison PE, Selander RK (1985) A population genetic framework for the study of invasive diseases caused by serotype b strains of Haemophilus influenzae. *Proc Natl Acad Sci USA* 82(15):5078–5082.
13. Maiden MC, et al. (1998) Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA* 95(6):3140–3145.
14. Maiden MC (2006) Multilocus sequence typing of bacteria. *Annu Rev Microbiol* 60: 561–588.
15. Meats E, et al. (2003) Characterization of encapsulated and noncapsulated Haemophilus influenzae and determination of phylogenetic relationships by multilocus sequence typing. *J Clin Microbiol* 41(4):1623–1636.
16. Musser JM, et al. (1990) Global genetic structure and molecular epidemiology of encapsulated Haemophilus influenzae. *Rev Infect Dis* 12(1):75–111.
17. LaCross NC, Marrs CF, Gilsdorf JR (2013) Population structure in nontypeable Haemophilus influenzae. *Infect Genet Evol* 14:125–136.
18. Touchon M, et al. (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet* 5(1):e1000344.
19. Muzzi A, Donati C (2011) Population genetics and evolution of the pan-genome of Streptococcus pneumoniae. *Int J Med Microbiol* 301(8):619–622.
20. Donati C, et al. (2010) Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biol* 11(10):R107.
21. Harris SR, et al. (2010) Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327(5964):469–474.
22. Croucher NJ, et al. (2011) Rapid pneumococcal evolution in response to clinical interventions. *Science* 331(6016):430–434.
23. Harrison A, et al. (2005) 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.
24. Bakaletz LO, Leake ER, Billy JM, Kaumaya PT (1997) Relative immunogenicity and efficacy of two synthetic chimeric peptides of fimbrin as vaccinogens against nasopharyngeal colonization by nontypeable Haemophilus influenzae in the chinchilla. *Vaccine* 15(9):955–961.
25. Strouts FR, et al. (2012) Lineage-specific virulence determinants of Haemophilus influenzae biogroup aegyptius. *Emerg Infect Dis* 18(3):449–457.
26. Hogg JS, et al. (2007) Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol* 8(6):R103.
27. Fleischmann RD, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269(5223):496–512.
28. Francisco AP, Bugalho M, Ramirez M, Carriço JA (2009) Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics* 10:152.
29. Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet* 11:94.
30. Eutsey RA, et al. (2013) Design and validation of a supragenome array for determination of the genomic content of Haemophilus influenzae isolates. *BMC Genomics* 14:484.
31. St Geme JW, 3rd (2002) Molecular and cellular determinants of non-typeable Haemophilus influenzae adherence and invasion. *Cell Microbiol* 4(4):191–200.
32. Rodriguez CA, et al. (2003) Prevalence and distribution of adhesins in invasive non-type b encapsulated Haemophilus influenzae. *Infect Immun* 71(4):1635–1642.
33. St Geme JW, 3rd, Kumar VV, Cutter D, Barenkamp SJ (1998) Prevalence and distribution of the hmw and hia genes and the HMW and Hia adhesins among genetically diverse strains of nontypeable Haemophilus influenzae. *Infect Immun* 66(1):364–368.
34. Fernaays MM, Lesse AJ, Cai X, Murphy TF (2006) Characterization of igaB, a second immunoglobulin A1 protease gene in nontypeable Haemophilus influenzae. *Infect Immun* 74(10):5860–5870.
35. Fernaays MM, Lesse AJ, Sethi S, Cai X, Murphy TF (2006) Differential genome contents of nontypeable Haemophilus influenzae strains from adults with chronic obstructive pulmonary disease. *Infect Immun* 74(6):3366–3374.
36. Murphy TF, et al. (2011) A clonal group of nontypeable Haemophilus influenzae with two IgA proteases is adapted to infection in chronic obstructive pulmonary disease. *PLoS ONE* 6(10):e25923.
37. Munson RS, Jr., Granoff DM (1985) Purification and partial characterization of outer membrane proteins P5 and P6 from Haemophilus influenzae type b. *Infect Immun* 49(3):544–549.
38. Nelson MB, et al. (1991) Molecular conservation of the P6 outer membrane protein among strains of Haemophilus influenzae: Analysis of antigenic determinants, gene sequences, and restriction fragment length polymorphisms. *Infect Immun* 59(8):2658–2663.
39. Porras O, et al. (1986) Difference in structure between type b and nontypable Haemophilus influenzae populations. *Infect Immun* 53(1):79–89.
40. Noda K, et al. (2010) Nasal vaccination with P6 outer membrane protein and alpha-galactosylceramide induces nontypeable Haemophilus influenzae-specific protective immunity associated with NKT cell activation and dendritic cell expansion in nasopharynx. *Vaccine* 28(31):5068–5074.
41. Kaur R, Casey JR, Pichichero ME (2011) Serum antibody response to three non-typeable Haemophilus influenzae outer membrane proteins during acute otitis media and nasopharyngeal colonization in otitis prone and non-otitis prone children. *Vaccine* 29(5):1023–1028.
42. Winter LE, Barenkamp SJ (2010) Construction and immunogenicity of recombinant adenovirus vaccines expressing the HMW1, HMW2, or Hia adhesion protein of nontypeable Haemophilus influenzae. *Clin Vaccine Immunol* 17(10):1567–1575.
43. Winter LE, Barenkamp SJ (2009) Antibodies specific for the Hia adhesion proteins of nontypeable Haemophilus influenzae mediate opsonophagocytic activity. *Clin Vaccine Immunol* 16(7):1040–1046.
44. Hong W, Peng D, Rivera M, Gu XX (2010) Protection against nontypeable Haemophilus influenzae challenges by mucosal vaccination with a detoxified lipooligosaccharide conjugate in two chinchilla models. *Microbes Infect* 12(1):11–18.
45. Prymula R, et al. (2009) Effect of vaccination with pneumococcal capsular polysaccharides conjugated to Haemophilus influenzae-derived protein D on nasopharyngeal carriage of Streptococcus pneumoniae and H. influenzae in children under 2 years of age. *Vaccine* 28(1):71–78.
46. Power PM, Bentley SD, Parkhill J, Moxon ER, Hood DW (2012) Investigations into genome diversity of Haemophilus influenzae using whole genome sequencing of clinical isolates and laboratory transformants. *BMC Microbiol* 12:273.
47. Hausdorff WP, Feikin DR, Klugman KP (2005) Epidemiological differences among pneumococcal serotypes. *Lancet Infect Dis* 5(2):83–93.
48. Caugant DA (2008) Genetics and evolution of Neisseria meningitidis: Importance for the epidemiology of meningococcal disease. *Infect Genet Evol* 8(5):558–565.
49. Brenner DJ, et al. (1988) Biochemical, genetic, and epidemiologic characterization of Haemophilus influenzae biogroup aegyptius (Haemophilus aegyptius) strains associated with Brazilian purpuric fever. *J Clin Microbiol* 26(8):1524–1534.
50. Croucher NJ, et al. (2013) Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet* 45(6):656–663.
51. Kurtz S, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5(2):R12.
52. R Core Team (2013) R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria).
53. Jombart T (2008) adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* 24(11):1403–1405.
54. Jombart T, Ahmed I (2011) adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* 27(21):3070–3071.
55. Lee C, Abdool A, Huang CH (2009) PCA-based population structure inference with generic clustering algorithms. *BMC Bioinformatics* 10(Suppl 1):S73.
56. Tamura K, et al. (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28(10):2731–2739.
57. Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci USA* 101(30):11030–11035.
58. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4(4):406–425.
59. Myers EW, et al. (2000) A whole-genome assembly of Drosophila. *Science* 287(5461): 2196–2204.
60. Otto TD, Dillon GP, Degrave WS, Berriman M (2011) RATT: Rapid annotation transfer tool. *Nucleic Acids Res* 39(9):e57.
61. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23(6):673–679.
62. Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol* 183:63–98.
63. Francisco AP, et al. (2012) PHYLOViZ: Phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics* 13:87.
64. Giufrè M, et al. (2006) Conservation and diversity of HMW1 and HMW2 adhesin binding domains among invasive nontypeable Haemophilus influenzae isolates. *Infect Immun* 74(2):1161–1170.