Haemophilus influenzae genome evolution during persistence in the human airways in chronic obstructive pulmonary disease

Melinda M. Pettigrew^a, Christian P. Ahearn^{b,c}, Janneane F. Gent^d, Yong Kong^{e,f,g}, Mary C. Gallo^{b,c}, James B. Munro^{h,i}, Adonis D'Mello^{h,i}, Sanjay Sethi^{c,j,k}, Hervé Tettelin^{h,i,1}, and Timothy F. Murphy^{b,c,l,1,2}

^aDepartment of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT 06510; ^bDepartment of Microbiology and Immunology, University at Buffalo, The State University of New York, Buffalo, NY 14203; ^cClinical and Translational Research Center, University at Buffalo, The State University of New York, Buffalo, NY 14203; ^dDepartment of Environmental Health Sciences, Yale School of Public Health, New Haven, CT 06510; ^eDepartment of Biostatistics, Yale School of Public Health, New Haven, CT 06510; ^fDepartment of Molecular Biophysics and Biochemistry, Yale School of Medicine, New Haven, CT 06510; ^gW.M. Keck Foundation Biotechnology Resource Laboratory, Yale School of Medicine, New Haven, CT 06510; ^hInstitute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201; ⁱDepartment of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD 21201; ^jDivision of Pulmonary, Critical Care and Sleep Medicine, Department of Medicine, University at Buffalo, The State University of New York, Buffalo, NY 14203; ^kDepartment of Medicine, Veterans Affairs Western New York Healthcare System, Buffalo, NY 14215; and ^lDivision of Infectious Diseases, Department of Medicine, University at Buffalo, The State University of New York, Buffalo, NY 14203

Edited by Rino Rappuoli, GSK Vaccines, Siena, Italy, and approved February 27, 2018 (received for review November 10, 2017)

Nontypeable *Haemophilus influenzae* (NTHi) exclusively colonize and infect humans and are critical to the pathogenesis of chronic obstructive pulmonary disease (COPD). In vitro and animal models do not accurately capture the complex environments encountered by NTHi during human infection. We conducted whole-genome sequencing of 269 longitudinally collected cleared and persistent NTHi from a 15-y prospective study of adults with COPD. Genome sequences were used to elucidate the phylogeny of NTHi isolates, identify genomic changes that occur with persistence in the human airways, and evaluate the effect of selective pressure on 12 candidate vaccine antigens. Strains persisted in individuals with COPD for as long as 1,422 d. Slipped-strand mispairing, mediated by changes in simple sequence repeats in multiple genes during persistence, regulates expression of critical virulence functions, including adherence, nutrient uptake, and modification of surface molecules, and is a major mechanism for survival in the hostile environment of the human airways. A subset of strains underwent a large 400-kb inversion during persistence. NTHi does not undergo significant gene gain or loss during persistence, in contrast to other persistent respiratory tract pathogens. Amino acid sequence changes occurred in 8 of 12 candidate vaccine antigens during persistence, an observation with important implications for vaccine development. These results indicate that NTHi alters its genome during persistence by regulation of critical virulence functions primarily by slipped-strand mispairing, advancing our understanding of how a bacterial pathogen that plays a critical role in COPD adapts to survival in the human respiratory tract.

Haemophilus influenzae | chronic obstructive pulmonary disease | whole-genome sequencing | genome evolution | candidate vaccine antigens

Nontypeable *Haemophilus influenzae* (NTHi) are pathobionts that exclusively colonize and infect humans and are adapted to survival in the human respiratory tract, their primary ecological niche. NTHi are critical to the course and pathogenesis of chronic obstructive pulmonary disease (COPD). Approximately 65 million people globally have COPD, which is the fourth leading cause of death worldwide and predicted to be third by the year 2030 (1, 2). NTHi persists in the lower airways of individuals with COPD for extended periods of time and causes inflammation, impaired pulmonary function, and tissue damage that leads to progressive loss of lung function (3–6). COPD is also characterized by acute exacerbations, which are intermittent worsening of symptoms that cause enormous morbidity (4). Approximately half of exacerbations of COPD are caused by bacteria, and NTHi is the most common bacterial cause (3).

NTHi has developed mechanisms to survive and persist in the hostile environment of the human airways (7). Mechanisms of

adaptation cannot be accurately studied in vitro or in animal models as a result of the unique physiological and immunological environments encountered by NTHi during colonization and infection in the human host. We conducted a 15-y prospective study of adults with COPD who were followed monthly to collect detailed clinical data and sputum samples that were cultured for bacterial pathogens, including NTHi. We hypothesized that NTHi alters its genome to adapt to survival in the human respiratory tract and that these adaptations facilitate persistence. To test this hypothesis, we conducted whole-genome sequencing (WGS) on this large collection of carefully characterized strains of NTHi.

The goals of the present study were to use our unique set of 269 prospectively collected cleared and persistent NTHi strains with corresponding epidemiologic and clinical data to (i) elucidate the phylogeny of NTHi strains from individuals with COPD relative to other strains with publicly available genomes; (ii)

Significance

Nontypeable *Haemophilus influenzae* (NTHi) exclusively colonize and infect humans and play an important role in the course and pathogenesis of chronic obstructive pulmonary disease (COPD). We conducted whole-genome sequencing of 269 NTHi isolates from a 15-y prospective study of COPD to assess in vivo adaptation of NTHi. NTHi uses slipped-strand mispairing in simple sequence repeats to regulate critical virulence functions as the primary mechanism to adapt to survival in the human airways. Analyses of changes in 12 candidate vaccine antigens during persistence provided data with important implications for guiding vaccine development. These results advance understanding of how an exclusively human pathogen alters its genome to adapt to survival in the hostile environment of the human respiratory tract.

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

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Published under the PNAS license.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database. For a list of accession numbers, see [SI Appendix, Table S2](#).

¹H.T. and T.F.M. contributed equally to this work.

²To whom correspondence should be addressed. Email: murphyt@buffalo.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1719654115/-DCSupplemental.

Published online March 19, 2018.

analyze the genomes of persistent strains of NTHi to identify changes that occur with persistence, including phase variation in simple sequence repeats (SSRs), SNPs, genome rearrangements, and gene loss and gene gain; and (iii) evaluate the effect of immune selective pressure and adaptation to the host airway environment. Overall, genetic variation during persistence in the human respiratory tract occurs via multiple mechanisms, with changes in SSRs being especially frequent. NTHi are naturally transformable; however, surprisingly limited gene gain or loss occurred during long-term persistence in the human host. Several candidate vaccine antigens underwent changes that may mediate immune escape during persistence, whereas others remained stable, an observation that has important implications in guiding vaccine development for NTHi.

Materials and Methods

The institutional review boards of the University at Buffalo and the Veterans Affairs Western New York Healthcare System approved this study; study participants provided written informed consent before enrollment.

COPD Study Clinic. The present study includes NTHi strains from subjects enrolled during the 15-y period from April 1994 through March 2009 taken from a 20-y prospective study of adults with COPD that was conducted at the Buffalo Veterans Affairs Medical Center from 1994 to 2014 (3, 8). Subjects were seen at monthly clinic visits and at unscheduled visits at the onset of suspected acute exacerbations. Detailed clinical data and expectorated sputum samples were collected at each visit.

Bacterial Strains. NTHi were identified by using standard microbiology techniques, and a P6-specific monoclonal antibody was used to distinguish NTHi from *H. haemolyticus* (9). Strains that were isolated at a single monthly clinic visit and were not isolated again at subsequent monthly clinic visits were classified as cleared. Strains that were isolated from a study participant at more than one monthly clinic visit and that were the same strain by multilocus sequence typing (MLST) were classified as persistent. We estimated the duration of persistence of individual NTHi strains by calculating the number of days between the first visit date when the strain was detected and the last visit date when the strain was detected, which provides a minimum estimate of the duration of persistence. MLSTs were determined as described previously (10). This study included 67 cleared NTHi strains and the first and last detected NTHi isolates of 101 persistent strains as described in *SI Appendix, Table S1*

Genome Sequencing, Assembly, and Annotation. Genomic DNA was extracted from low-passage NTHi strains (three passages from the original isolation), and samples were subjected to 150-bp paired-end sequencing on an Illumina HiSeq 2500 system in the Next-Generation and Expression Analysis Core at the University at Buffalo or 250-bp paired-end sequencing on an Illumina MiSeq system at the Yale Center for Genome Analysis. *Btrim* was used to filter the raw sequencing data to retain high-quality sequence reads (11). De novo assembly was accomplished by using velvet version 1.2.10 with 99 as the kmer size and default parameters (12). Assembled sequences were annotated by using Prokka (13).

The first and last isolates of four persistent strains and one cleared strain (nine NTHi strains in total) were also sequenced at the Yale Center for Genome Analysis by using the PacBio RS II platform. These strains were sequenced by using P6-C4 chemistry and a targeted library size of 10 kb with one strain per single molecule, real-time cell. The Hierarchical Genome Assembly Process was used to assemble the PacBio genomes (14). The consensus sequence was further corrected by using in-house scripts to iteratively map Illumina reads to the assembly until no differences were found between the assembly and the Illumina reads. The nine finished gap-free genome sequences assembled from PacBio data were trimmed of overlaps at the ends of the circular chromosome and then aligned to the *H. influenzae* Rd KW20 reference nucleotide sequence (GenBank accession no. NC_000907.1) using NUCmer (15) and rotated to make the first nucleotide the same as Rd KW20. The nine trimmed and rotated sequences were subjected to the CloVR Microbe automated annotation pipeline (16).

Phylogenetic Analyses. Whole genome-level core SNP-based phylogenetic analysis of our 269 genomes together with 134 publicly available *H. influenzae* genomes [from the National Center for Biotechnology Information (NCBI) as of October 25, 2017] was performed. First, SNPs were identified within the core regions shared by all genomes by using the In Silico Genotyper (ISG) pipeline (17) with default parameters. The ISG output of “clean

unique variants,” a concatenated FASTA file of aligned core SNPs, was then used for phylogenetic analysis by using RAxML v8.2.0 (18). Nondefault parameters include using the -d flag for complete random starting tree instead of the default randomized stepwise addition parsimony starting tree, the -f flag for rapid bootstrap analysis and search for best scoring maximum-likelihood tree, and the -asc-corr=lewis flag to correct the likelihood for ascertainment bias and account for the lack of invariant sites. An initial analysis using the autoMRE option (i.e., extended majority-rule consensus tree criterion) was employed to determine the minimum number of bootstrap replicates needed, which was 50. We then performed three independent analyses with different starting seeds, each with 500 rapid bootstrap inferences followed by a thorough maximum-likelihood search. The SNP matrix consisted of 403 strains, 157,303 positions per strain, and 110,159 distinct alignment patterns. Trees were visualized with Dendroscope v3.4.4 (19) and FigTree v1.4.3.

Genome Alignment and SNP Identification. Reference-free whole-genome multiple alignment-based comparative analyses of diverse subsets (as many as ~150 genomes) of our *H. influenzae* genomes and/or annotated publicly available *H. influenzae* genomes were performed by using the CloVR-Comparative pipeline (20). Among other outputs, this pipeline generates a list of SNPs present within aligned sections of the genomes compared and a list of clusters of syntenic orthologous genes across the genomes (21). It also generates a Sybil interactive Web interface (22) for interrogation of the comparative genomics data. Sybil instances for the comparison of our nine finished genomes among themselves and together with 19 publicly available *H. influenzae* finished genomes (from NCBI as of October 25, 2017). An independent set of clusters of orthologs, Jaccard-filtered ortholog clusters (JOCs) based on reciprocal best-BLAST matches (23) was generated by performing all-vs.-all searches of all of the genes predicted in our 269 NTHi genomes.

SNPs identified in persistent strains with closed genomes are reported in *Dataset S1*. JOCs were used to expand analysis of these robust SNPs to the entire set of 269 genomes. When SNP-containing genes were included in JOCs that harbored one whole gene from each of the 269 genomes, the multiple alignment of the gene nucleotide sequences from each JOC was inspected for variations at the SNP position. Counts for each variant are reported in the column labeled “Orthologs (JOCs) in 269 genomes” in *Dataset S1*. In several cases, SNP-containing genes were undergoing frequent frame shifting (*Identification of SSRs*) or were present in only a small subset of the 269 strains. These cases resulted in JOCs with too few gene members for expanding the SNP analysis.

Identification of SSRs. A Perl program was used to identify tandemly repeated motifs ranging from 1 to 9 nt in length (24). The minimum number of repeat units required to be present in an uninterrupted tandem arrangement was as follows: nine repetitions for homopolymeric tracts, five for 2-nt repeats, four for 3-nt repeats, and three for 4–9-nt repeats (24). The Perl program was set to include a 500-nt region upstream of genes and not allow any mismatches. SSRs that changed between the first and last isolate in at least one of the four persistent strains with closed genomes were manually curated to determine their position within ORFs or in the promoter of ORFs. Orthologous genes across the first and last isolates of the four strains were defined by using the CloVR-based clusters of syntenic orthologous genes.

SSRs in the remaining 97 persistent strains with draft genomes were also identified using the Perl script described here previously. JOCs were defined across the first and last isolates of the persistent strains (97 persistent strains with draft genomes and four persistent strains with closed genomes).

We used an in-house script to identify SSRs that changed between the first and last isolate of a persistent strain. Briefly, we used an iterative process to search for motifs in the first isolate that matched the motifs in the final isolate of a persistent strain; the program was used to search for the original motif and the reverse complement, and also accounted for potential shifts in the starting base pair of a given motif.

Assessment of Competence of Strains for Transformation. Transformation efficiency of persistent strains of NTHi was determined by using a modification of the protocol by the M-IV method of Poje and Redfield (25), using a linearized plasmid p572b that contains the gene that encodes peroxiredoxin-glutaredoxin (26). Results are expressed as a percentage of cells transformed.

Analysis of Candidate Vaccine Antigens. Sequences of 12 NTHi candidate vaccine antigens from the first and final isolates of the 101 persistent strains were analyzed for changes during persistence and positive selection. Reference sequences of the 12 NTHi candidate vaccine antigens were aligned to the *H. influenzae* 86-028NP genome (GenBank accession no. CP000057.2) by using the Burrows–Wheeler Alignment tool (bwa) (27). For a particular

vaccine antigen, the reads that aligned to its genome position were extracted from BAM files using SAMtools (28). An in-house multiple alignment program was used to align the reads and obtain the consensus sequence of the vaccine antigen gene for each strain. The ORFs were translated by using EMBOSS Transeq (European Molecular Biology Laboratory/European Bioinformatics Institute) (29). Multiple sequence alignments of the nucleotide and the amino acid sequences were generated with ClustalW in MacVector using default parameters. The nucleotide and amino acid sequences of persistent strains were analyzed for synonymous and nonsynonymous changes during persistence. The Phylogenetic Analysis by Maximum Likelihood (PAML) codeML program was used to identify amino acids of antigens experiencing positive selection by using previously described methods (*SI Appendix, SI Materials and Methods*) (30, 31).

P2 and P5 Antigen Membrane Topology Prediction. The membrane topology of the P2 protein of strain 86–028NP (GenBank accession no. AAX87199.1) was predicted by Prediction of TransMembrane Beta-Barrel Proteins Viterbi method in combination with the predicted P2 membrane topology described by Bagos et al. (32) and Sikkema and Murphy (33). The P5 protein membrane topology prediction of 86–028NP (GenBank accession no. AAX88164.1) was generated by ClustalW alignment and comparison of the 86–028NP sequence to strain UC19 strain sequence, of which the topology was determined by Webb and Cripps (34). Transmembrane domains of P2 and P5 proteins were further supported by the BOCTOPUS 2 beta barrel prediction program (35).

Analysis of Amino Acid Sequence Haplotypes of Candidate Vaccine Antigens. Unique haplotype sequences of each candidate antigen were extracted from the amino acid sequences of persistent strains with the *uniqHaplo.pl* Perl script. Unique amino acid haplotype sequences for each antigen were aligned, and pairwise sequence percent identity scores were calculated by using MacVector. Mean haplotype pairwise diversity (MHPD) values are the calculated average of all pairwise percent identity values between the unique amino acid sequence haplotypes of each candidate antigen.

Expression and Purification of Recombinant Proteins and Mutant Construction. Stop codon mutations in candidate antigen ORF sequences were identified by translation of the nucleotide sequences. Immunoblot assays with antiserum to recombinant protein E were used to determine whether the protein was expressed in strains with the stop codon. Briefly, cloning, expression, and purification of protein E were accomplished by using pCATCH, allowing expression of recombinant lipoprotein (36). A protein E-KO mutant was generated by overlap extension PCR and allelic exchange in NTHi strain 86–028NP as described previously (37). Whole-cell lysates of WT, KO mutants, and NTHi COPD strains with and without the stop codon were performed by using polyclonal antiserum to recombinant purified protein E.

Results

Phylogeny of NTHi Genomes from COPD Airways. The genome sequences of 67 cleared NTHi strains and the first and final detected NTHi isolates of 101 persistent strains isolated prospectively over 15 y from adults with COPD were determined. Summary assembly data on the 260 draft and 9 closed (gap-free or finished) genomes are provided in *SI Appendix, Table S2*. The median number of contigs of draft genomes is 70.5 (range, 26–968). The median genome size is 1,841,659 (range 1,688,785–2,031,876) and the median number of predicted genes is 1,813 (range, 1,599–2,395). These data substantially increase the number of publicly available *H. influenzae* genomes, including increasing the number of closed genomes from 19 to 28.

We used MLST as an initial step to examine the diversity of our strains. Among the 168 initial acquisition strains, 77 different sequence types (STs) and one (strain 99P15H1) that lacked the *fucK1* gene and could not be assigned an ST were represented (*SI Appendix, Table S1*). The most common STs were 155, 156, and 159, which represented 8, 6, and 11 strains, respectively. Of the 101 persistent strains, two changed ST during persistence (39P25H/39P29H1 and 106P3H1/106P7H1); in both cases, the change resulted from single locus variants caused by base pair changes in the *fucK1* gene.

We generated three maximum-likelihood phylogenetic trees based on core genome SNPs of the 269 COPD genomes and 134 publicly available *H. influenzae* genomes (Fig. 1). Resulting tree

topologies were largely congruent with RAxML-computed average relative Robinson–Foulds distances ranging from 1% to 3.25%. Our 269 genomes are annotated with the ST, designation as cleared or persistent, designation as an exacerbation (infection) or colonization (no symptoms) strain, and year of isolation. The genomes of initial acquisition and final isolates of persistent strains group together, consistent with results from MLST. Moreover, STs generally group together; for example, ST155 strains group together in the same clade along with the single-locus variant ST414. The COPD strains are distributed throughout the tree, indicating that there is no clear genetic clustering based on clinical source of the strain, geography, duration of persistence, exacerbation vs. colonization, or year of isolation.

Changes During Persistence. We focused our detailed analyses of changes in NTHi genomes during persistence in the human respiratory tract initially on four persistent strains by analyzing the closed genomes of the first detected isolate and final detected isolate. Strains 5P28H1/5P54H1, 6P24H2/6P32H1, 48P106H1/48P153H1, and 67P38H1/67P56H1 persisted in four different study participants for 819, 252, 993, and 570 d, respectively. We extended these analyses to the first detected isolate and final detected isolate of 97 additional persistent strains with draft genomes to determine whether key observations regarding SSR repeats and SNPs were consistently observed. The 101 persistent strains persisted for a median of 161 d (range, 2–1,422 d).

SSR Analysis. Phase variation mediated by slipped-strand mispairing is an important regulatory mechanism that allows NTHi to adapt to different or changing environments (38, 39). Changes in the number of SSRs may affect gene expression when the SSRs are located in upstream regulatory regions. Alternatively, SSRs within ORFs may affect protein translation. The genomes of the four persistent strains contained unique SSRs, as designated by the SSR coordinate, ranging in number from 72 (48P153H1) to 92 (5P28H1) among the eight initial and final isolates of the four strains (*Dataset S2*).

Further analysis identified 22 genes in which the number of SSRs changed during persistence in at least one of the four strains, suggesting that NTHi adapts to the environment in the human respiratory tract by regulating the expression of these genes. Each of the SSRs that changed in number was manually evaluated with regard to its position in relation to the ORF, and, in each case, we also determined whether the SSR changed in the other isolate of the persistent strain (*Dataset S3*). All of the genes that change in SSR copy number during persistence encode potential virulence functions, including adhesins, modifications of lipooligosaccharide, and iron uptake (Table 1). For example, one such gene is a DNA methylase that mediates a rapid and reversible change in the expression of many genes throughout the genome (40). The number of repeats changed in all four persistent strains in the HMW1A/HMW2A-related adhesins and in each of the hemoglobin-haptoglobin-binding proteins. We observed an interrupted repeat in the HMW1A/HMW2A adhesin encoding locus (BV121_1850) in strain 5P28H1 and in the hemoglobin-haptoglobin-binding protein loci (BV121_803) in strain 5P28H1. These interrupted repeats may arise from homologous recombination between paralogous loci, which generates imperfect repeats (39).

The number of unique SSR repeats in 101 persistent strains ranged from 30 to 88 per isolate, with a mean of 64.6 (SD, 8.8; *SI Appendix, Table S3*). We also determined whether each SSR was present in the final isolate of the persistent strain and whether the matching SSR repeat changed in number. The SSR repeat coordinate, motif, gene name, and number of repeats are provided for SSRs that changed during persistence (*Dataset S3*). The mean number of repeats that changed in persistent strains was 6.9 (SD, 3.0). The proportion of SSR repeats that changed per persistent strain ranged from 1.9% to 27.4% (*SI Appendix, Table S3*). We examined the correlation between the number of repeats that changed and the duration of persistence. There was a significant positive correlation between the proportion of SSR

Fig. 1. Phylogenetic tree of 269 nontypeable *H. influenzae* COPD genomes from this study together with 134 publicly available *H. influenzae* genomes. The core-genome SNP-based maximum-likelihood phylogenetic tree is presented in its polar tree layout, and nodes are sorted in increasing order. Bootstrap values from 500 iterations are shown on each node, and most show very strong support with values $\geq 90\%$. A branch-length scale bar equivalent to 0.05 substitutions per site is provided at the bottom of the figure. The 269 COPD genomes are labeled with strings that include data separated by underscores in the following order: isolate name, MLST ST, paired isolate information for persistent strains (a number representing the clinic visit of isolation, followed by "perV," followed by the visit number of the paired isolate) or "clr" for cleared strains, first isolate (marked as "1") or last isolate ("0"), exacerbation ("1") or colonization ("0"), and year of isolation. Isolates are color-coded in a gradient of red shades based on grouped dates of isolation, from 1995 to 1996 in light red, then darker for 1997–1999, 2000–2001, 2003–2005, and the darkest for 2006–2009. Closed (gap-free) COPD genomes are colored in green. The 19 closed publicly available genomes (per NCBI as of October 25, 2017) are labeled with their strain name followed by MLST type and colored in blue. Strain names for the remaining 115 draft publicly available genomes are in black.

Table 1. Genes with SSRs that change frame during persistence in the human respiratory tract

Gene function*	Implications for respiratory tract persistence	Region of gene with SSRs	Strains that change no. of SSRs (of 4)
HMW1A/HMW2A adhesin	Adherence to respiratory mucosal surface	Promoter	4
Sialyltransferase	Molecular mimicry to evade host immune response	ORF	4
DNA methylase, type I restriction-modification methylase	Phase variation of multiple virulence genes	ORF	1
Efflux transporter	Resist toxic substances in airways	Promoter	1
Glycosyltransferase family	Modifications of LOS on the bacterial surface	ORF, promoter	4
Hemoglobin-haptoglobin-binding protein	Iron acquisition	ORF	4
Acetyltransferase	Post translational modification and regulation of multiple proteins	ORF	1
Outer membrane autotransporter	Secretion of multiple virulence factors	ORF	1
Phosphotransferase	Transport of carbohydrate and cellular regulatory functions	ORF	3
TonB dependent receptor family protein	Uptake and transport of nutrients and substrates	Promoter, ORF	2

LOS, lipooligosaccharide.

*Some gene functions underwent changes in multiple genes.

repeats that changed during persistence and the duration of colonization in days (Pearson correlation coefficient $r^2 = 0.35$, $P < 0.0001$). Based on the frequency of SSRs that likely result in changes in expression of multiple virulence factors, and the correlation between the number of changes and duration of persistence, we conclude that NTHi uses slipped-strand mispairing as a major mechanism to adapt to microenvironments in the human respiratory tract.

SNPs Occurring During Persistence of NTHi in Human Airways. Analysis of closed genomes of the four persistent strains showed substantial variability in the number of SNPs that occurred in genomic loci during persistence in human airways among the four strains (Table 2 and Dataset S1). For example, strain 6P24H2/6P32H1 underwent 81 SNPs during persistence for 570 d; of these, 11 were intergenic and 70 were in 13 genes, including 59 nonsynonymous and 11 synonymous SNPs. Remarkably, strain 6P24H2/6P32H1 showed a single SNP during persistence for 252 d (Table 2 and Dataset S1). Nonsynonymous SNPs were observed in a wide variety of genes in these four pairs, with the highest number of nonsynonymous SNPs in hemoglobin-haptoglobin-binding protein A, followed by outer membrane protein P2. SNPs in these genes occurred in clusters of nucleotides undergoing variations in regions that we called “hyper-variable” (Dataset S1). The hemoglobin-haptoglobin-binding protein A genes also underwent frequent SSR-induced frame-shifting. The SNP identified in ribosomal protein L22, which resulted in a switch from glycine to aspartic acid at amino acid 91 (G91D) in strain 5P28H1/5P54H1, was previously identified as associated with a

fourfold increase in azithromycin minimum inhibitory concentration during persistence (10). However, this SNP was observed only twice within the 269-genome dataset.

Genome Rearrangements During Persistence in Human Airways. An inversion of ~400 kbp in length occurred during persistence in the human respiratory tract in two of the four persistent strains with closed genomes. The inversions depicted in the Sybil (22) synteny gradient image (Fig. 2) were confirmed via PCR amplification of upstream and downstream nonrepetitive flanking regions of the inversion sites for both persistent strains.

The ends of the 400-kbp inversion are located within the loci that encode HMW1 and HMW2, which are key adhesins that mediate adherence of NTHi to human respiratory epithelial cells. The *hmw1* and *hmw2* structural genes are followed by two additional genes downstream, *hmw1B* and *hmw1C* or *hmw2B* and *hmw2C* (41). The *hmwA* genes encode the surface-exposed nonpilus adhesins, and *hmwB* and *hmwC* genes encode accessory proteins required for processing and secretion of the structural protein (42). HMW1B/1C and HMW2B/2C are interchangeable, which is consistent with the observed high degree of homology between the two accessory proteins (41). Close examination of the inversion sites demonstrates that the HMW operons are not disturbed, such that each HMW locus still contains one A, one B, and one C gene.

Gene Gain and Loss and Competence in Persistent Strains. Based on our analyses of ortholog clusters, there were no major changes in the genome in terms of gene gain or loss. The size differences

Table 2. SNPs that occur in NTHi genomes during persistence in the human respiratory tract

Persistent strain	Persistence, d	Total SNPs	Genes with SNPs	Genes with nonsynonymous SNPs	Intergenic SNPs
5P28H1 5P54H1	819	23	5	5	5
6P24H2 6P32H1	252	1	1	1	0
48P106H1 48P153H1	993	3	2	2	1
6P24H2 6P32H1	570	81	13	13	11

Fig. 2. (A) Sybil synteny gradient image of nine closed genomes compared with reference strain 86_028NP. Clusters of syntenic orthologous genes were used to draw orthologs of genes predicted in the reference genome (86-028NP). Each ortholog is drawn with the color from the gradient corresponding to its position in its own genome. As a result, breaks in the color gradient represent inversions, translocations, or insertions, and white spaces indicate deletions or regions harboring genes that are not included with reference genes in clusters of orthologs. The three genomes containing the inversion are indicated with asterisks. (B) Schematic of upstream and downstream inversion sites in the first isolate of a persistent strain 5P28H1, containing the inversion, and the final isolate of the same strain, 5P54H1, with no inversion. PCR primers were designed to span the flanking regions of the inversion break sites, utilizing a common forward primer for the upstream break site and a common reverse primer for the downstream break site. (C) Agarose gel of PCR products demonstrating the specificity of PCR primers surrounding the inversion sites. Common forward primer for the upstream break site “1” amplifies a product with unique reverse primer “3” using 5P28H1 DNA in lane 1, but not with 5P54H1 DNA in lane 5. Common forward primer “1” amplifies a product with unique reverse primer “4” using 5P54H1 DNA in lane 6, but not with 5P28H1 DNA in lane 2. Forward primer “5” with common reverse primer “2” for the downstream break site amplifies a product with 5P28H1 DNA in lane 3, but not with 5P54H1 DNA in lane 7. Forward primer “6” with common reverse primer “2” amplifies a product with 5P54H1 DNA in lane 8, but not with 5P28 DNA in lane 4.

between the acquisition and final isolates of persistent strains with closed genomes were minimal; -39 bp, -25 bp, -127 bp, and $+76$ bp for strains 5P28H1/5P54H1, 6P24H2/6P32H1, 48P106H1/48P153H1, and 67P38H1/67P56H1, respectively (*SI Appendix, Table S2*). We evaluated the transformation frequency of the acquisition and final isolates of each of the four persistent strains with closed genomes, given the limited gene gain and loss observed during persistence. The average transformation frequencies were 8.55×10^{-7} for 5P28H1, 3.27×10^{-6} for 5P54H1, 1.10×10^{-5} for 48P106H1, 9.61×10^{-5} for 48P153H1, 6.29×10^{-6} for 67P38H1, and 5.8×10^{-7} for 67P56H1. 6P24H2 and 6P32H1 were not transformable by using laboratory conditions that readily resulted in transformation of the other six isolates. The genome of this strain underwent only one coding-region SNP in the coding regions during 272 d of persistence in the human respiratory tract.

Efficient transformation of *H. influenzae* requires the presence of 9-mer (aagtcggt) DNA uptake signal sequences (USSs) (43). We reasoned that a noncompetent lineage would experience a divergence or gradual loss of USSs over time. Inspection of the number of uptake sequences in the four persistent strains revealed

no difference for 6P24H2 and 6P32H1 compared with the other persistent strains. 6P24H2 and 6P32H1 each had 1,504 uptake sequences, 5P28H1 and 5P54H1 each had 1,508, 48P106H1 and 48P153H1 each had 1,478, and 67P38H1 and 67P56H1 each had 1,479 uptake sequences. Thus, the number of DNA USSs in each isolate of persistent strains remains constant. However, USS sequences are stable for hundreds of millions of years, and loss of USS likely occurs over large time scales (44).

In *H. influenzae*, competence for transformation is controlled by the Sxy-dependent cAMP receptor protein regulon, which is composed of 26 genes that are controlled by the CRP and Sxy regulators (45, 46). We identified orthologs of each of the 17 required competence genes from *H. influenzae* strain Rd (46) in each isolate of our four persistent strains. There were no obvious mutations that explain the lack of transformability of 6P24H2 and 6P32H1 in vitro. Although isolates 6P24H2 and 6P32H1 were not competent under laboratory conditions, they may be competent under natural conditions.

Evaluation of NTHi Candidate Vaccine Antigens. Vaccine development for NTHi is undergoing exciting new developments with the licensing of a pneumococcal conjugate vaccine that contains the conserved NTHi protein, protein D (47). Recognizing the limited efficacy of a vaccine with a single NTHi antigen, this proof-of-principle observation has stimulated active preclinical and clinical development of several additional NTHi vaccine antigens (48, 49).

We conducted two separate analyses of candidate vaccine antigens in 101 persistent strains. Complete ORF sequence data were not available for some antigens in a few isolates; in these instances, they were not included in the analyses (Table 3). We compared multiple sequence alignments of each of the 12 vaccine antigens in the first and final isolates of each strain to identify positively selected codons. We also assessed the extent to which 12 candidate vaccine antigens of NTHi underwent changes in the airways of adults with COPD. Positive selection describes selection in the population of persistent strains. In contrast, our second analysis examined how genes that encode antigens in individual strains change during persistence.

Positive Selection in Vaccine Candidate Antigens. We investigated 12 candidate vaccine antigens for positively selected codons by using the PAML-codeml program based on the sequence alignments of the first and final isolates in the population of the 101 persistent strains. Because a majority of codons in a protein-coding sequence are under physiochemical constraint based on protein structure and function (50), we applied selection models in the codeml program that identified individual codons in antigen sequences under positive selection.

Several amino acids of four antigens, P2, P5, Hap, and D15, experienced positive selection in the persistent strains (Table 3 and *SI Appendix, Table S4*). The surface exposed loops of the outer membrane proteins P2 and P5 are variable among strains and during persistent infection (33, 51, 52). Based on the P2 and P5 predicted membrane topology, positively selected amino acids were located in the surface-exposed regions of both proteins. Fig. 3 highlights the amino acids experiencing positive selection in reference to the 86-028NP antigen sequences.

Changes in Candidate Vaccine Antigens During Persistence in COPD Airways. In addition to sequence diversity identified in the population of strains, we also assessed the extent to which 12 candidate vaccine antigens of NTHi underwent changes during persistence in the airways of adults with COPD. All 101 persistent strains contained the ORFs of the 12 candidate antigens. Of the 12 candidate vaccine antigens analyzed, 36 amino acid sequence changes were observed between the first and final isolates of 32 persistent strains. Two strains contained multiple changes in different antigens. Changes occurred during persistence in eight antigens: P2, P5, PD, Hap, D15, HtrA, P4, and PE (Table 3 and *SI Appendix, Table S5*). The remaining four antigens

Table 3. Candidate vaccine antigens that undergo selection and change during persistence of NTHi in the human respiratory tract

Candidate vaccine antigen	Biological function	Codons experiencing selection (sequences analyzed) [†]	No. of strains that change during persistence [‡]
P6	Peptidoglycan-associated lipoprotein	None (201)	0
PilA	Type IV pilus	None (200)	0*
OMP 26	Translocase	None (199)	0*
Protein F	ABC transporter	None (200)	0*
P4	Acid phosphatase	None (200)	1
Protein E	Adhesin	None (200) [§]	1
HtrA	Heat shock protein	None (199)	2**
Protein D	Phosphodiesterase	None (195)	4***
D15 (BamA-like protein)	Nucleotidyltransferase	Multiple (196)	2**
Hap	Adhesin	Multiple (186)	3*** [#]
P5 fimbrin	Fimbrin	Multiple (200) [§]	5*
P2	Porin	Multiple (187)	21

[†]Positive selection analysis was performed on the first and final isolate (202) antigen sequences of the 101 persistent strains. The analysis was performed on the initially extracted ORFs from the genome sequences using bwa.

* $n = 100$, ** $n = 99$, *** $n = 95$.

[‡]Analysis of 101 strains except where noted.

[§]The nucleotides encoding the PE and P5 stop codons were removed so to include these strain sequences for detection of positive selection.

^{||}Two of the four changes were nonsynonymous SNPs.

[#]One of the three changes was a nonsynonymous SNP.

showed no changes during persistence (Protein F, P6, OMP26, and PilA).

Outer-membrane protein P2 underwent amino acid sequence changes during persistence in 21 of 101 strains. Insertions, deletions, and SNPs all occurred in predicted surface-exposed loops, with loops 4, 5, and 6 having the most frequent changes (Fig. 3). Despite these insertions, deletions, and SNPs, the gene remained in frame in all strains, consistent with the observation that, as the major porin protein of NTHi, P2 is important for viability.

Outer-membrane protein P5 experienced nonsynonymous insertions, deletions, and SNPs during persistence in five of 101 persistent strains. Each of the variations observed during persistence resulted in amino acid changes in regions that are predicted to be surface exposed loops, in particular loops 2 and 3 (Table 3 and Fig. 3). No changes in P5 loops 1 and 4 were observed during persistence, yet sequence diversity was present among the strains. Nonsynonymous SNPs occurred in PD, Hap, D15, HtrA, P4, and PE during persistence in one or two strains each among the 101 persistent strains.

Stop Codons in Candidate Vaccine Antigens. Analysis of genes that encode candidate vaccine antigens of NTHi from 168 COPD strains (67 cleared and 101 initial acquisitions of persistent strains) revealed the presence of stop codons in three genes. The protein E gene of 15 of 101 persistent strains (14.9%) and 12 of 67 cleared strains (17.9%) contained a G169A amber mutation. The mutation was present upon acquisition in all but one strain that had the mutation; the persistent strain 59P3H1/59P6H1 acquired the mutation during persistence in the human respiratory tract. Immunoblot assays with antiserum to recombinant protein E confirmed that the protein was not detected in strains with the stop codon but was detected in other strains (SI Appendix, Fig. S1). Thus, the protein E gene of 27 of 168 strains (16.1%) contained an amber mutation that resulted in the absence of expression of protein E. This amber mutation was observed by Singh et al. (53) in 1.6% of NTHi strains.

The gene that encodes the adhesin Hap in 2 of 168 strains had deletions in the ORF of the strains, both resulting in downstream stop codons. In persistent strain 31P7H7/31P14H6, the Hap gene contained a one-base deletion resulting in a frame-shift mutation and several downstream stop codons. In persistent strain 102P20H1/102P32H1, the final isolate contained a 22-bp deletion in the Hap gene resulting in downstream stop codons.

The P5 gene of strain 18P16H1/18P17H2 had an opal mutation (GAA to TGA) upon acquisition. Another strain, 48P28H1/48P45H1, acquired a 4-bp insertion during persistence that resulted in a stop codon. Immunoblot with antiserum to P5 confirmed that P5 was not expressed in the strains that contained the stop codon but was expressed in other strains.

Amino Acid Sequence Haplotypes of Candidate Vaccine Antigens. We assessed amino acid sequence conservation of the 12 candidate vaccine antigens by determining the pairwise percent identity scores of the unique haplotypes for each antigen and determining the prevalence of each haplotype among the persistent strains. Eight of the 12 antigens had greater than a 95% MHPD of percent identity (Fig. 4A). In addition to the high MHPD for these antigens, they had the fewest number of unique haplotypes, with the exception of the D15 antigen, which had 48 haplotypes (Fig. 4B). The P6 antigen had the fewest unique haplotypes with eight; 92% of persistent strains were represented by one P6 haplotype (Fig. 4B).

PilA, P5, P2, and Hap showed the most sequence heterogeneity among strains with an MHPD of percent identity between 90% and 71% and a large number of haplotypes (SI Appendix, Table S6). The P2 antigen had the greatest number of haplotypes ($n = 103$) and a range of pairwise percent identity comparisons from 99.7% to as low as 42.1%. The Hap antigen had the lowest MHPD of percent identity at 71.4%. This analysis has important implications in assessing the conservation of candidate vaccine antigens to prioritize them further as NTHi vaccine formulations are developed.

Discussion

H. influenzae was the first bacterial species to have its genome sequenced, in 1995 (54). WGS has been used to describe the population structure of *H. influenzae* (55), provide insight into genomic diversity (56), identify targets for strain differentiation and diagnostics (57), and characterize virulence properties (58–61). Our study is unique in that we examined a large collection of prospectively collected *H. influenzae*, including strains that persisted for months to years in the human respiratory tract, to answer questions related to in vivo adaptation of this exclusively human pathogen in the airways of its human host. Our analyses indicate that NTHi uses slipped-strand mispairing in SSRs as a major mechanism for survival and pathogenesis in the human airways. Moreover, the propensity for SNPs during persistence

Pettigrew et al.

We observed evidence of a relatively large-scale genome rearrangement occurring during persistence; a 400-kbp inversion occurred between the loci that encode HMW1 and HMW2 in two of the four persistent strains with closed genomes. Inversions may change bacterial phenotypes by altering gene expression patterns or disrupting genes and are therefore thought to help bacteria persist and adapt to changing environments (63, 64). Three of the 19 publicly available closed genomes also have the same inversion: 477 (GenBank accession no. NZ_CP007470.1), 723 (GenBank accession no. NZ_CP007472.1), and C486 (GenBank accession no. NZ_CP007471.1). Our analyses indicate that the genes in the HMW operon are swapped, but operon structures are not disrupted by the inversion. The two swapped HMW operons may change the expression of the adhesins.

Future experimental work should examine the role of this inversion in altering the phenotype and/or transcriptional profile of NTHi strains.

A key observation from our study indicates that SSRs are a main driver of change during persistence in the respiratory tract. We identified and manually confirmed changes in 22 genes in four persistent strains, and essentially every gene with changing SSRs has virulence implications. In a human challenge model, phase variation of 13 selected NTHi genes was examined in vivo during experimental human colonization of 15 subjects over 6 d (65). Genes encoding a choline kinase (*licA*) and IgA protease (*igaB*) were generally phase-off in the inoculated strain and switched to phase-on in a significant number of samples. Power et al. (39) evaluated SSR-mediated phase variation in *H. influenzae* and determined that there were 56, 60, 53, and 54 SSRs in strains 86–028NP, Rd KW20, R2846, and R2866, respectively. We identified a greater number of SSRs, between 72 and 92, in our closed genomes by using similar bioinformatics thresholds for SSR identification. Our research expands upon the observations from these two studies because we were able to examine the potential for genome-wide phase variation during long periods of in vivo persistence in the human airways by using prospectively collected strains.

We also observed that, as the length of time a patient carried a strain increased, the number of SSRs that changed also increased. This important observation suggests that changes in SSRs, and thus regulation of expression genes by slipped-strand mispairing, are driven by selective pressure in the human airways. This observation is consistent with recent work that shows that passage of NTHi strains through human respiratory epithelial cells selects for expression of IgA proteases that are regulated by slipped-strand mispairing (66).

In several instances, including the LOS sialyltransferase and the hemoglobin-haptoglobin-binding proteins, the SSR was present near the beginning of the ORF such that a change in copy number would engender an early frame shift, resulting in the translation of a very short peptide. It should be noted that most of these N-terminal peptides (short ORFs) were not predicted by the annotation system in our genomes; therefore, we performed manual curation of these peptides by using existing non-frame-shifted protein sequences as a guide (Dataset S2). It should also be noted that our detailed determinations were based on whole-genome sequences of four strains (eight isolates) and may therefore be underestimates of the number of potentially phase-variable genes given that we likely detected the most prevalent variants. Phase variants may exist in mixed populations, and ours were not based on uncultured samples, nor were they confirmed by additional methods (e.g., fragment length analysis of mixed populations) (40, 65). Data from the 97 persistent strains with draft genomes provide valuable information regarding the potential extent of phase variation in NTHi. However, these data should be interpreted with caution; repeat regions are challenging to sequence and assemble, and some of the SSR repeats were near the end of contigs.

We did not observe instances of large-scale gene acquisitions or deletions during persistence of NTHi in human airways in the strains with closed genomes. This is in contrast to *Streptococcus pneumoniae*, another exclusively human pathogen. Hiller et al. (67) sequenced sequential *S. pneumoniae* isolates colonizing a

child over a 7-mo period and determined that ~156 kb of the genomic content (7.8% of the genome) was exchanged during multiple horizontal gene-transfer events.

Vaccine development for NTHi is an active, evolving area of research. Immune escape has received little attention in studies of NTHi vaccine development because strains of NTHi that have persistently colonized the human respiratory tract have not been studied. Amino acids of 4 of 12 candidate vaccine antigens in various stages of development experienced positive selection in our 101 persistent strains. Furthermore, several candidate vaccine antigens experienced nonsynonymous changes during persistence in the lower airways, all in predicted surface-exposed regions of the P2 and P5 molecules analyzed. Intriguingly, amino acids experiencing positive selection also underwent changes in surface exposed loops of the P2 and P5 antigens during persistence. This observation supports that antigens with surface exposed epitopes are under immune selective pressure and that the resulting antigenic variation that occurs during persistence confers immune evasion and contributes to NTHi persistence. Although recognition by the host immune system is driving diversity in antigens, adaptation to the host lower airway environment is also likely contributing to such changes.

The observation of sequence variations in the final isolate compared with the first isolate of a persistent strain raises the question of whether these mutations were deleterious and enabled clearance of the strain. To this end, we sequenced the P2 gene in intermediate isolates of the patient 19 strain in which the P2 antigen incurred an 18-bp deletion during persistence. This deletion was a sequential process, but more importantly, the strain persisted for several months after the final deletion event occurred. This provides proof of principle that not all mutations observed in final isolates of persistent strains are deleterious and result in clearance of the strain. In future studies, we will investigate the intermediate isolates of persistent strains to assess the extent to which mutations are advantageous, deleterious, or neutral with regard to persistence.

In summary, our results advance the understanding of how an exclusively human pathogen alters its genome to adapt to survival in the human respiratory tract. Identifying changes that occur in the genomes of a unique set of strains of a human pathogen in its natural biological niche advances our understanding of the pathogenesis of bacterial infection in COPD substantially. The observations regarding the impact of selective pressure on candidate vaccine antigens in the human respiratory tract has immediate implications in prioritizing antigens for vaccine development.

ACKNOWLEDGMENTS. We thank Shaun Adkins, Jonathan Crabtree, James Matsumura, Suvarna Nadendla, the Genomics Resource Center, the Bioinformatics Resource Center, and the IT Department at Institute for Genome Sciences, University of Maryland School of Medicine, for their help with data storage, management, analysis, and visualization; Charmaine Kirkham, Aimee Brauer and Antoinette Johnson (Clinical and Translational Research Center, University at Buffalo) for expert technical assistance in the laboratory; Dr. Lauren Bakaletz (Nationwide Children's Hospital) and Dr. Joseph St. Geme III (Children's Hospital of Philadelphia) for their donations of P5 and Hap antiserum as well as KO mutants of genes encoding each antigen. This work was supported by National Institute of Allergy and Infectious Diseases of the National Institutes of Health Grant R01AI019641 (to T.F.M. and M.M.P.), the Department of Veterans Affairs (S.S. and T.F.M.), and National Center for Advancing Translational Sciences Award UL1TR001412 to the University at Buffalo.

- Mathers CD, Loncar D (2006) Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med* 3:e442.
- Kim V, Criner GJ (2013) Chronic bronchitis and chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 187:228–237.
- Murphy TF, Brauer AL, Schiffmacher AT, Sethi S (2004) Persistent colonization by *Haemophilus influenzae* in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 170:266–272.
- Sethi S, Murphy TF (2008) Infection in the pathogenesis and course of chronic obstructive pulmonary disease. *N Engl J Med* 359:2355–2365.
- White AJ, et al. (2003) Resolution of bronchial inflammation is related to bacterial eradication following treatment of exacerbations of chronic bronchitis. *Thorax* 58:680–685.
- Desai H, et al. (2014) Bacterial colonization increases daily symptoms in patients with chronic obstructive pulmonary disease. *Ann Am Thorac Soc* 11:303–309.
- Wong SM, Akerley BJ (2012) Genome-scale approaches to identify genes essential for *Haemophilus influenzae* pathogenesis. *Front Cell Infect Microbiol* 2:23.
- Sethi S, Evans N, Grant BJ, Murphy TF (2002) New strains of bacteria and exacerbations of chronic obstructive pulmonary disease. *N Engl J Med* 347:465–471.
- Murphy TF, et al. (2007) *Haemophilus haemolyticus*: A human respiratory tract commensal to be distinguished from *Haemophilus influenzae*. *J Infect Dis* 195:81–89.
- Pettigrew MM, et al. (2016) Effect of fluoroquinolones and macrolides on eradication and resistance of *Haemophilus influenzae* in chronic obstructive pulmonary disease. *Antimicrob Agents Chemother* 60:4151–4158.
- Kong Y (2011) Btrim: A fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics* 98:152–153.
- Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829.

13. Seemann T (2014) Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30: 2068–2069.
14. Chin CS, et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 10:563–569.
15. Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* 30:2478–2483.
16. Angiuoli SV, et al. (2011) CloVR: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics* 12:356.
17. Sahl JW, et al. (2015) The in silico genotyper (ISG): An open-source pipeline to rapidly identify and annotate nucleotide variants for comparative genomics applications. *bioRxiv*, 10.1101/015578.
18. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.
19. Huson DH, et al. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 8:460.
20. Agrawal SV, et al. (2017) CloVR-comparative: Automated, cloud-enabled comparative microbial genome sequence analysis pipeline. *BMC Genomics* 18:332.
21. Angiuoli SV, Dunning Hotopp JC, Salzberg SL, Tettelin H (2011) Improving pan-genome annotation using whole genome multiple alignment. *BMC Bioinformatics* 12:272.
22. Riley DR, Angiuoli SV, Crabtree J, Dunning Hotopp JC, Tettelin H (2012) Using Sybil for interactive comparative genomics of microbes on the web. *Bioinformatics* 28: 160–166.
23. Crabtree J, Angiuoli SV, Wortman JR, White OR (2007) Sybil: Methods and software for multiple genome comparison and visualization. *Methods Mol Biol* 408:93–108.
24. Siena E, et al. (2016) In-silico prediction and deep-DNA sequencing validation indicate phase variation in 115 *Neisseria meningitidis* genes. *BMC Genomics* 17:843.
25. Poje G, Redfield RJ (2003) Transformation of *Haemophilus influenzae*. *Haemophilus influenzae Protocols*, eds Herbert MA, Hood DW, Moxon ER (Humana Press, Totowa, NJ), Vol 71, pp 57–70.
26. Murphy TF, Kirkham C, Sethi S, Lesse AJ (2005) Expression of a peroxiredoxin-glutaredoxin by *Haemophilus influenzae* in biofilms and during human respiratory tract infection. *FEMS Immunol Med Microbiol* 44:81–89.
27. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
28. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079.
29. Li W, et al. (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res* 43:W580–W584.
30. Suyama M, Torrents D, Bork P (2006) PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34: W609–W612.
31. Jeffares DC, Tomiczek B, Sojo V, dos Reis M (2015) A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. *Methods Mol Biol* 1201:65–90.
32. Bagos PG, Liakopoulos TD, Spyropoulos IC, Hamodrakas SJ (2004) PRED-TMBB: A web server for predicting the topology of beta-barrel outer membrane proteins. *Nucleic Acids Res* 32:W400–W404.
33. Sikkema DJ, Murphy TF (1992) Molecular analysis of the P2 porin protein of nontypeable *Haemophilus influenzae*. *Infect Immun* 60:5204–5211.
34. Webb DC, Cripps AW (1998) Secondary structure and molecular analysis of interstrain variability in the P5 outer-membrane protein of non-typable *Haemophilus influenzae* isolated from diverse anatomical sites. *J Med Microbiol* 47:1059–1067.
35. Hayat S, Peters C, Shu N, Tsigoris KD, Elofsson A (2016) Inclusion of dyad-repeat pattern improves topology prediction of transmembrane β -barrel proteins. *Bioinformatics* 32:1571–1573.
36. Yang M, Johnson A, Murphy TF (2011) Characterization and evaluation of the *Moraxella catarrhalis* oligopeptide permease A as a mucosal vaccine antigen. *Infect Immun* 79:846–857.
37. Murphy TF, Brauer AL (2011) Expression of urease by *Haemophilus influenzae* during human respiratory tract infection and role in survival in an acid environment. *BMC Microbiol* 11:183.
38. Moxon R, Bayliss C, Hood D (2006) Bacterial contingency loci: The role of simple sequence DNA repeats in bacterial adaptation. *Annu Rev Genet* 40:307–333.
39. Power PM, et al. (2009) Simple sequence repeats in *Haemophilus influenzae*. *Infect Genet Evol* 9:216–228.
40. Attack JM, et al. (2015) A biphasic epigenetic switch controls immunoevasion, virulence and niche adaptation in non-typeable *Haemophilus influenzae*. *Nat Commun* 6: 7828.
41. St Geme JW, 3rd, Grass S (1998) Secretion of the *Haemophilus influenzae* HMW1 and HMW2 adhesins involves a periplasmic intermediate and requires the HMWB and HMWC proteins. *Mol Microbiol* 27:617–630.
42. Cholon DM, et al. (2008) Serial isolates of persistent *Haemophilus influenzae* in patients with chronic obstructive pulmonary disease express diminishing quantities of the HMW1 and HMW2 adhesins. *Infect Immun* 76:4463–4468.
43. Smith HO, Tomb JF, Dougherty BA, Fleischmann RD, Venter JC (1995) Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. *Science* 269:538–540.
44. Redfield RJ, et al. (2006) Evolution of competence and DNA uptake specificity in the Pasteurellaceae. *BMC Evol Biol* 6:82.
45. Redfield RJ, et al. (2005) A novel CRP-dependent regulon controls expression of competence genes in *Haemophilus influenzae*. *J Mol Biol* 347:735–747.
46. Sinha S, Mell JC, Redfield RJ (2012) Seventeen Sxy-dependent cyclic AMP receptor protein site-regulated genes are needed for natural transformation in *Haemophilus influenzae*. *J Bacteriol* 194:5245–5254.
47. Prymula R, et al. (2006) Pneumococcal capsular polysaccharides conjugated to protein D for prevention of acute otitis media caused by both *Streptococcus pneumoniae* and non-typable *Haemophilus influenzae*: A randomised double-blind efficacy study. *Lancet* 367:740–748.
48. Murphy TF (2015) Vaccine for nontypeable *Haemophilus influenzae*: The future is now. *Clin Vaccine Immunol* 22:459–466.
49. Khan MN, et al. (2016) Developing a vaccine to prevent otitis media caused by nontypeable *Haemophilus influenzae*. *Expert Rev Vaccines* 15:863–878.
50. Yang Z, Swanson WJ (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol* 19:49–57.
51. Hiltke TJ, Schiffmacher AT, Dagonese AJ, Sethi S, Murphy TF (2003) Horizontal transfer of the gene encoding outer membrane protein P2 of nontypeable *Haemophilus influenzae*, in a patient with chronic obstructive pulmonary disease. *J Infect Dis* 188:114–117.
52. Duim B, et al. (1997) Molecular variation in the major outer membrane protein P5 gene of nonencapsulated *Haemophilus influenzae* during chronic infections. *Infect Immun* 65:1351–1356.
53. Singh B, Brant M, Kilian M, Hallström B, Riesbeck K (2010) Protein E of *Haemophilus influenzae* is a ubiquitous highly conserved adhesin. *J Infect Dis* 201:414–419.
54. Fleischmann RD, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496–512.
55. De Chiara M, et al. (2014) Genome sequencing of disease and carriage isolates of nontypeable *Haemophilus influenzae* identifies discrete population structure. *Proc Natl Acad Sci USA* 111:5439–5444.
56. 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.
57. Coughlan H, et al. (2015) Comparative genome analysis identifies novel nucleic acid diagnostic targets for use in the specific detection of *Haemophilus influenzae*. *Diagn Microbiol Infect Dis* 83:112–116.
58. Zhang L, et al. (2012) Nontypeable *Haemophilus influenzae* genetic islands associated with chronic pulmonary infection. *PLoS One* 7:e44730.
59. Strouts FR, et al. (2012) Lineage-specific virulence determinants of *Haemophilus influenzae* biogroup aegyptius. *Emerg Infect Dis* 18:449–457.
60. Su YC, Resman F, Hörhold F, Riesbeck K (2014) Comparative genomic analysis reveals distinct genotypic features of the emerging pathogen *Haemophilus influenzae* type f. *BMC Genomics* 15:38.
61. Spencer-Smith R, Varkey EM, Fielder MD, Snyder LA (2012) Sequence features contributing to chromosomal rearrangements in *Neisseria gonorrhoeae*. *PLoS One* 7: e46023.
62. LaCross NC, Marrs CF, Gilsdorf JR (2013) Population structure in nontypeable *Haemophilus influenzae*. *Infect Genet Evol* 14:125–136.
63. Kresse AU, Dinesh SD, Larbig K, Römling U (2003) Impact of large chromosomal inversions on the adaptation and evolution of *Pseudomonas aeruginosa* chronically colonizing cystic fibrosis lungs. *Mol Microbiol* 47:145–158.
64. Cui L, Neoh HM, Iwamoto A, Hiramatsu K (2012) Coordinated phenotype switching with large-scale chromosome flip-flop inversion observed in bacteria. *Proc Natl Acad Sci USA* 109:E1647–E1656.
65. Poole J, et al. (2013) Analysis of nontypeable *Haemophilus influenzae* phase-variable genes during experimental human nasopharyngeal colonization. *J Infect Dis* 208: 720–727.
66. Murphy TF, et al. (2017) Immunoglobulin A protease variants facilitate intracellular survival in epithelial cells by nontypeable *Haemophilus influenzae* that persist in the human respiratory tract in chronic obstructive pulmonary disease. *J Infect Dis* 216: 1295–1302.
67. Hiller NL, et al. (2010) Generation of genetic diversity among *Streptococcus pneumoniae* strains via horizontal gene transfer during a chronic polyclonal pediatric infection. *PLoS Pathog* 6:e1001108.