Partial Analysis of the Genomes of Two Nontypeable Haemophilus influenzae Otitis Media Isolates


Center for Microbial Pathogenesis, Columbus Children's Research Institute and The Department of Pediatrics, The Ohio State University, Columbus, Ohio 43205, and Laboratory for Microbial Genomics, Department of Microbiology and Immunology, The University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma 73104

Received 7 January 2004/Accepted 3 February 2004

In 1995, The Institute for Genomic Research completed the genomic sequence of a rough derivative of Haemophilus influenzae serotype d, strain KW20. This sequence, though extremely useful in understanding the basic biology of H. influenzae, has yet to provide significant insight into our understanding of disease caused by nontypeable H. influenzae (NTHI), because serotype d strains are not generally pathogens. In contrast, NTHI strains are frequently mucosal pathogens and are the primary pathogens of chronic otitis media as well as a significant cause of acute otitis media in children. Thus, it is of great importance to further understand their biology. We used a DNA-based microarray approach to identify genes present in a clinical isolate of NTHI that were absent from strain Rd. We also sequenced the genome of a second NTHI isolate from a child with chronic otitis media to threefold coverage and then used an array of bioinformatics tools to identify genes present in this NTHI strain but absent from strain Rd. These methods were complementary in approach and results. We identified, in both strains, homologues of H. influenzae lav, an autotransported protein of unknown function; madA, which encodes tryptophanase; as well as a homologue of Pasteurella multocida tspA, which encodes an alkyl peroxidase that may play a role in protection against reactive oxygen species. We also identified a number of putative restriction-modification systems, bacteriophage genes and transposon-related genes. These data provide new insight that complements and extends our ongoing analysis of NTHI virulence determinants.

Otitis media (OM) is a highly prevalent pediatric disease worldwide (9, 26, 31). The most recently available statistics indicate that 24.5 million physician office visits were made for OM in 1990, representing a >200% increase over those reported in the 1980s (41). OM is the most frequently diagnosed illness in children (<15 years) and is the primary cause for emergency room visits (9). While OM is only very rarely associated with mortality, the morbidity associated with OM is significant. Hearing loss is the most common complication of OM (4, 25, 31) with behavioral, educational, and language development delays being additional consequences of early-onset OM with effusion (26, 31, 49). The socioeconomic impact of OM is also great, with direct and indirect costs of diagnosing and managing OM exceeding $5 billion annually in the United States alone (1, 8, 27, 31).

Clearly, there is a tremendous need to develop effective and accepted approaches to the management and, preferably, the prevention of OM. Vaccine development holds the greatest promise and would be the most cost-effective method to accomplish this goal (22, 28). Progress in this area is most advanced for Streptococcus pneumoniae, a causative agent of acute OM, and in fact, capsular-conjugate vaccines are currently licensed for use. Less progress has been made for nontypeable Haemophilus influenzae (NTHI), the gram-negative pathogen that predominates in chronic OM with effusion (13, 32, 46) and is also associated with approximately one-third of all cases of acute OM (30). Hampering development of effective vaccines against NTHI is the as-yet-incomplete understanding of the pathogenesis of NTHI-induced middle ear disease. There are multiple gaps in our understanding of the dynamic interplay between microbe-expressed virulence factors and the host's immune response as the disease progresses from one of host immunological tolerance of a benign nasopharyngeal commensal to that of an active defensive reaction to an opportunistic invader of the normally sterile middle ear space.

In 1995, a group from The Institute for Genomic Research sequenced the genome from H. influenzae strain Rd KW20 (20). Strain Rd is a noncapsulated derivative of a serotype d organism. Although strain Rd has some virulence properties, serotype d strains are generally considered to be commensals; they do not frequently cause disease (11). It is therefore important to determine the differences between disease-causing strains of H. influenzae and strain Rd. Since we lack complete genome sequence for pathogenic strains, which would allow an exhaustive comparison, this question may be approached by analyses that indicate which genes are present in disease-causing strains of H. influenzae that are absent from strain Rd. Suppressive subtractive hybridization has been employed to identify genomic differences between isolates of a species. In Haemophilus, this technique was used to identify genomic dif-
ferences between H. influenzae biogroup aegyptius strains of the Brazilian purpuric fever clonal group and conjunctivitis isolates (33, 45). Bergman and Akerley employed a position-based PCR scanning technique to compare an H. influenzae serotype b genome with the strain Rd genome (6). With respect to NTHI, most studies have focused on differences in the sequence of single genes or small gene clusters or the presence or absence of specific genes, although Williams et al. identified and characterized a bacteriophage that was present in an invasive nontypeable isolate (50). From an epidemiologic perspective, Pettigrew et al. employed pulse field gels, ribotyping and enterobacterial repetitive intergenic consensus typing to examine strain differences among NTHI (42). Similarly, Meats et al. have employed multilocus sequence typing for epidemiologic characterization of isolates (36).

In an effort to more broadly approach the identification of virulence determinants in NTHI, we have employed two methods to identify genes in NTHI that are absent from H. influenzae strain Rd. First, we employed a DNA-based microarray approach to identify genes present in NTHI strain 1885MEE that were absent from strain Rd. We also sequenced the genome of a pathogenic NTHI strain, designated 86-028NP, to threefold coverage. Assuming a Poisson distribution, the probability of any given base being sampled at threefold coverage is 95%, although practical limitations, due to the nonrandom nature of biological sequences, cause the actual percentage of sequence identified to be somewhat lower than the theoretical model predicts. With these partial sequence data, a bioinformatics approach was used to identify genes that were not present in strain Rd. Several genes of interest present in both strains were further characterized.

**Materials and Methods**

**Choice of strains for the projects.** Epidemiologic studies of NTHI have indicated that the strains are heterogeneous with respect to outer membrane protein profiles (5), enzyme allotypes (38), and readouts generated via various commonly used epidemiologic tools. There have been several attempts to subtype NTHI, but the methodologies have not been useful with respect to identification of an isolate that is most suitable for both genomic sequencing and genomic comparison with strain Rd. We therefore chose the low-page-number isolates, strains 1885MEE and 86-028NP, that were independently recovered from children with chronic OM. Both have subsequently been well characterized in vitro (3, 24) as well as in chinchilla models of OM, with strain 86-028NP being the better understood (2, 14, 48).

**Library constructions.** Chromosomal DNA was prepared from strains 86-028NP and 1885MEE using Puregene reagents (Gentra Systems, Minneapolis, Minn.). DNA from strain 1885MEE was partially digested with Sau3A, and fragments in the range 0.5 to 1.5 kb were isolated by preparative agarose gel electrophoresis and then ligated into the BamHI site of a pUC18 derivative, modified by substitution of the blu gene with the kanamycin resistance gene of Tn903. The transformation mixture was transformed into Escherichia coli TOP10 cells and clones containing inserts were identified by blue/white selection as outlined above. Six thousand white clones were saved.

Genomic DNA from strain 86-028NP was sheared to 2 to 4 kb with a nebulizer (10). The sheared DNA was ethanol precipitated, its ends were repaired using the End-It Repair Kit (Epicentre), and it was subjected to size selection by agarose gel electrophoresis to obtain 2- to 4-kb fragments (10, 44). These fragments were cloned into the vector pUC18 that had been cleaved with SmaI and was then phosphorylated. The ligation mixture was transformed into E. coli XL-1 Blue MRF and clones that contained inserts were identified by blue-white screening on 2% yeast tryptone (YT)-ampicillin plates containing X-Gal (5-bromo-4-chloro-3-indolyl-β-D-glucopyranoside). Insert-containing clones were transferred into 96-deep-well plates containing 1.5 ml of 2 × YT-ampicillin broth. The deep-well-plate cultures were grown overnight (18 to 22 h) at 37°C with shaking.

**Plasmid preparation and DNA microarray analysis.** Plasmid DNA from the strain 1885MEE library was purified by alkaline lysis in a 96-well format on a Beckman Biomek 2000 automated robotics workstation, using a typical sodium dodecyl sulfate (SDS)-NaOH lysis protocol (10). The final ethanol-precipitated templates (in a 96-well plate) were each dissolved in 100 μl of double-distilled H2O. DNA from random clones was analyzed by gel electrophoresis to determine the size of the insert. After demonstrating that the library was comprised of clones with inserts of the appropriate size, plasmids were denatured by boiling in 3× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate)–400 mM NaOH–10 mM EDTA (pH 8.0) and spotted onto SuperChip I, aminopropylsilane coated slides (Perkin-Elmer Life Sciences Inc., Boston, Mass.) using an Affymetrix 417 four-pin arrayer (MWG Biotech Inc., High Point, N.C.). To generate fluorescent targets, genomic DNA was directly labeled with either Cy5-dUTP or Cy3-dUTP using a BioPrime DNA labeling kit that utilized a random primed Klenow method (Amersham Biosciences [Little Chalfont, Buckinghamshire, United Kingdom] and Invitrogen Corporation [Carlsbad, Calif.]). For hybridization, each slide was blocked at 42°C with 5× SSC–0.5% SDS–1% bovine serum albumin, and then a LifterSlip (Erie Scientific Co., Portsmouth, N.H.) was placed over the arrayed DNA; the denatured probe was pipetted under it, and the slide was incubated at 50°C for 16 h. Posthybridization, the slide was washed with 0.5× SSC–0.1% SDS and then 0.06× SSC–0.01% SDS followed by 0.06× SSC alone, and then the slide was dried and scanned using an Affymetrix 428 Array Scanner (MWG Biotech Inc.). Using a GenePix Pro3 analysis package (Axon Instruments, Union City, Calif.), each wavelength image was assigned a color and a ratio image of the two was built. Inserts from positive clones were sequenced using ABI Big-Dye terminators and M13 universal primers and then analyzed on an Applied Biosystems (Foster City, Calif.) 3100 capillary electrophoresis DNA sequencer.

**Genomic sequencing and assembly.** Template from the 86-028NP library was prepared as above using the alkaline lysis procedure. Cycle-sequencing reactions were run in 96-well format using PE Big-Dye terminators and universal primers (M13 forward and reverse). Excess dye terminators were removed with Sephadex G-50 columns and analyzed on an ABI 3700 capillary electrophoresis DNA sequencer. Phred/Phrap was used for data assembly, employing the default assembly parameters (17, 18, 25).

**Data analysis.** Using the BLAST algorithms, the sequences from clones identified as strain 1885MEE-specific, were searched against the NCBI nucleotide and protein databases and the previously published strain Rd genome. In order to determine whether there was a high level of synteny between the Rd and 86-028NP genomes, the consensus sequences (contigs) from our focal assembly was cross-compared to the strain Rd genome sequence, using a two-dimensional version of tricross (43). The tricross ordering for the 86-028NP contigs was based on the Rd chromosome coordinates, with each contig anchored at the start coordinate of the earliest Rd gene hit.

To generate the list of genes observed in strain 86-028NP which are not found in strain Rd, the following procedure was used. The strain 86-028NP threefold-coverage contigs were analyzed by BLAST (BLASTX) against the NCBI’s non-redundant (NR) database (data current as of 8 May 2003). This was accomplished using a Perl wrapper for NCBI’s blastcl3 interface. The results were accumulated in a relational database (MySQL). The top 100 hits to NR (10–9 or better) were layered over the contig sequence in a best-score first fashion. For each BLAST hit, contig bases that were not covered by a better-scoring hit were marked as belonging to this hit in the relational database. For each hit that claimed more than 10 unique bases to itself, the translation of the entire region that generated the hit was analyzed by BLAST (TBLASTX, using NCBI’s blaststall software, to a locally generated strain Rd database) against the H. influenzae strain Rd gene set (built from NC_000907.8n). Regions that found hits to strain Rd sequence at an E score of 0.01 or better were marked as strain Rd genes and excluded. All hits to NCBI’s NR database that claimed more than 10 unique bases to itself found no strain Rd hits, are reported as genes potentially unique to strain 86-028NP compared to strain Rd. The BLAST parameters for analysis of the contig set versus the NR database were chosen to accommodate the partial sequence and low-quality sequencing regions, and the presence of false positives in the analysis is expected. Likewise, the BLAST parameters for excluding genes present in Rd were chosen to eliminate sequences with even cursory similarity to Rd and are expected to limit the result set to only those sequences for which we have a high confidence in their uniqueness.

The strain 1885MEE data set was compared by BLAST with the threefold-contig data from strain 86-028NP using the BLAST program implemented at http://www.microbial-pathogenesis.org.
RESULTS AND DISCUSSION

Genomic differences between NTHI strain 1885MEE and H. influenzae strain Rd. In order to develop a methodology that could be used to compare the genomes of a panel of NTHI isolates, a DNA microarray approach was adopted as this approach has provided useful data in other systems (16). Chromosomal DNA from strain 1885MEE was labeled with Cy5, and chromosomal DNA from H. influenzae strain Rd was labeled with Cy3, using a Klenow DNA polymerase-based direct incorporation method. Equal amounts of labeled genomic DNA were then pooled and used to hybridize to the strain 1885MEE genomic DNA array. The fluorescence signals from Cy5 and Cy3 were measured independently for each spot, using an Affymetrix 428 Array Scanner. Using GenePix Pro 3.0, the two channels were combined and a ratio image was generated. By comparing the relative fluorescence values for the Cy5 channel and the Cy3 channel, it was then possible to determine which clones contained DNA that was specific for strain 1885MEE, relative to strain Rd.

There is little precedent for the analysis of data derived from genomic DNA arrays, although work with Campylobacter jejuni indicates that a ratio of signal intensities between the two channels of twofold or higher, could be indicative of a clone containing a strain-specific insert (16). Thus, using the signal generated by labeled strain 1885MEE genomic DNA hybridizing with strain 1885MEE arrayed DNA as the control signal, a minimum fluorescence ratio of 2.5 (1885MEE/Rd) was used to select clones for further analysis (Fig. 1A). First, as the library was spotted in duplicate, only clones in which both replicates met this threshold were further characterized. Of 268 clones that met this criterion, viable sequence was obtained from 263 clones, vector sequence was trimmed and batch-wise analysis was conducted using the BLAST programs (http://www.microbial-pathogenesis.org). Thirty-two percent of the clones had homology to genes in both strains Rd and 1885MEE. Reanalysis of the data set, choosing clones having a fluorescence ratio of 5.0 (1885MEE/Rd), reduced the number of false positives from 32 to 8%, with the loss of only five strain 1885MEE-specific clones (Fig. 1A).

A total of 55 loci were identified as being unique to strain 1885MEE. A large number of these were conserved hypothetical genes. Some of the identified genes were of immediate interest due to their putative functions (Table 1). The translated sequence from clone 32-G11 had homology to hypothetical proteins from Pasteurella multocida and Neisseria meningitidis, both known pathogens. Similarly, the translated sequence from clone 62-G04 had homology to a hypothetical protein from Pseudomonas aeruginosa and a hypothetical membrane-associated transport protein from Helicobacter pylori, organisms both associated with human disease. Translated sequence from clone 17-H05 was 100% identical to the sequence of HiA from strain INT1, an invasive NTHI isolate from a meningitis patient. The translated amino acid sequence from clone 66-E09 was 80% identical to a portion of the amino acid sequence of Lav from H. influenzae strain INT1. Translated sequence from clone 17-H02 had 95% amino acid identity with a portion of the amino acid sequence of Tnase (TnA) from H. influenzae serotype b strain Eagan (34). Translated sequence from clone 14-A12 was 85% identical to the amino acid sequence of TsaA from P. multocida. TsaA is an alkyl peroxidase, which has a role in protection against reactive oxygen species and is potentially important in survival during phagocytosis (21, 35). These data, and the database reference of their respective homologue, are shown in Table 1.

Sequences in 29 clones had homology to phage genes, many of which were similar to genes from Haemophilus phages HP1 and HP2. Of more interest, was the identification of strain 1885MEE-specific homologues of genes from the Gifsy 1 and 2 and Fels-2 prophages. Gifsy 1 and 2 were originally identified in S. enterica serovar Typhimurium and have been implicated in Salmonella virulence (19). Fels-2 has been identified in S. enterica (7). Sequences with homology to a number of other hypothetical genes were observed. The database reference for each hit as well as each alignment is available on our Web site.

Genomic differences between two NTHI clinical isolates analyzed by microarray. To determine further the genetic content of strain 1885MEE compared to strain 86-028NP, genomic DNA was labeled with either Cy5 (1885MEE) or Cy3 (86-028NP), and equal amounts of labeled genomic DNA were pooled and used to hybridize to the strain 1885MEE genomic DNA array. The relative fluorescence values for the Cy3 channel and the Cy5 channel were then compared to determine which clones contained DNA that was specific for strain 1885MEE, relative to strain 86-028NP. Again, a fluorescence ratio of 5.0 (1885MEE/86-028NP) was used to choose putative strain 1885MEE-specific clones, relative to strain 86-028NP (Fig. 1B). Plasmid DNA was purified and valid sequence obtained from 72 of the 83 clones thought to contain sequence unique to strain 1885MEE. Sixty-eight of the clones had homology to genes in strain 86-028NP. The remaining four clones had homology to two hypothetical genes from H. influenzae strain Rd (H11029 and H11317), one from Vibrio vulnificus and one from P. multocida, respectively.

Comparison of the partial genome sequence of strain 86-028NP with the strain Rd genome sequence. The genome of strain 86-028NP was sequenced to threefold coverage. Sequencing reads (8219) from the strain 86-028NP clones were assembled into 576 contigs. Seventy-five of the eighty-eight contigs of $\leq 5000$ bp showed significant similarity via BLASTN to genes in H. influenzae strain Rd. To visualize the relationship between the gene order in H. influenzae strain 86-028NP and H. influenzae strain Rd, the strain 86-028NP threefold-contig set and the Rd gene set were compared by BLASTN using tricross. The results were plotted by sorting the contigs based on gene coordinates of the Rd genes hit, then anchoring each contig at the smallest Rd coordinate found (Fig. 2) (43). When compared in this fashion, if the 86-028NP genome had the same gene order as strain Rd and the contig assemblies were completely correct, the plot would appear as an increasing staircase. At the current level of resolution, a high level of synteny is observed between the two genomes. This is in contrast to a comparison of the H. ducreyi and H. influenzae genomes, where little synteny is observed (Munson, Ray, and coworkers, unpublished data).

BLASTX was used to identify 161 genes that had homologues in GenBank but were not found in H. influenzae strain Rd. We identified sequences related to known Haemophilus genes not present in strain Rd. The genes that encode the high-molecular weight-adhesive proteins and their associated...
processing proteins, the \textit{lav} gene, the \textit{tnaA} gene, the \textit{hicAB} genes and restriction systems, among others, were identified in this data set (Table 2). Sequences with homology to transposases, phage genes and hypothetical genes were also identified.

\textbf{Further characterization of genes identified in both data sets but absent from strain Rd.} Several potential genes of interest were observed in data sets from both isolates. The \textit{lav} gene encodes a member of the AIDA-I/VirG/PerT family of virulence-associated autotransporters and was reported to be
restricted in distribution to a small number of pathogenic strains, including *H. influenzae* biogroup aegyptius and Brazilian purpuric fever isolates. It has previously been reported as absent from strain Rd (12). We completed the sequence for this gene from strain 86-028NP (GenBank accession number AY559035). The derived amino acid sequence of the strain 86-028NP *Lav* protein was 88% identical to the derived amino acid sequence of *Lav* from INT1. Interestingly, *Lav* from INT1, *H. influenzae* biogroup aegyptius, *N. meningitides* and *N. gonorrhoeae* have a series of GCAA repeats 10 bp downstream of the start codon, which are thought to be involved in slip-strand mispairing during DNA replication, giving rise to phase variation of this surface-exposed molecule. The coding sequence of *Lav* in strain 86-028NP has 20 GCAA repeats in the same location. The comparative arrangement of the genes in the *Lav* region for 86-028NP and strain Rd are shown in Fig. 3A.

The *tnaA* gene encodes tryptophanase and is part of the *tna* gene cluster. This gene cluster has been widely studied in *E. coli*. In both *E. coli* and *H. influenzae* serotype b strain Eagan, the other components of the gene cluster are *tnaB*, which encodes a low-affinity tryptophan permease, and *tnaC*, which encodes a small peptide involved in regulation of expression of *tnaAB*. The *tnaA* gene cluster was located between the *nlpD* and *mutS* genes and is absent from the strain Rd genome (34). In the 86-028NP threefold-contig set, fragments of *tnaA* were identified spanning the ends of two contigs with the same organization as that found in strain Eagan. We completed the sequence from strain 86-028NP (GenBank accession number AY559033). The derived amino acid sequence of *TnaA* from strain 86-028NP was 99% identical to the sequence from strain Eagan. The comparative arrangement of the genes in the *tna* region for 86-028NP and strain Rd are shown in Fig. 3B.

Nearly all pathogenic isolates of *H. influenzae* test positive for indole production, one of the phenotypes used in the *H. influenzae* biotyping scheme (29). Tryptophanase catabolizes tryptophan to indole, pyruvate and ammonia, allowing tryptophan to be used as both a carbon and a nitrogen source (39). Tryptophanase may also play a role in both buffering against

![FIG. 2. Contigs larger than 5 kb were aligned to the strain Rd genomic sequence as described in the text. Each point represents a hit to a strain Rd gene or several adjacent hits to closely linked strain Rd genes. The scatter plot shown here indicates that there are some differences between the gene arrangement in the strain 86-028NP genome (or assembly errors) compared to the gene arrangement in strain Rd, though an analysis of the data on a contig-by-contig basis supports the hypothesis that rearrangements represent a small portion of the genome.](http://www.microbial-pathogenesis.org/hinf/files/20030917_Fig2.html)
extremes in pH and in epithelial cell adherence and biofilm formation (15, 47).

Although not initially found in our BLAST analysis of the strain 86-028NP contig set, the tsaA gene was identified in strain 86-028NP contig set when the strain 1885MEE sequence was used to query the threefold database. The sequence of the strain 86-028NP gene was completed (GenBank accession number AY559033). The derived amino acid sequence was 89% identical to the P. multocida homologue. The comparative arrangement of the genes in the tsaA region for strains 86-028NP and Rd are shown in Fig. 3C.

The hicAB genes were identified in the strain 86-028NP data set and the hicA gene was identified in the strain 1885MEE data set. In serotype b organisms, these genes are in close proximity to the hemagglutinating pilus (LKP) genes. Mhlanga-Mutangadura and coworkers characterized the LKP genes as well as the pepN genes, as depicted in Fig. 3D. The serotype b strain Eagan contains the hifaABCDE gene cluster and produces hemagglutinating pilus. Strain Rd lacks the hicAB genes as well as the hifaABCDE gene cluster. In general, the nontypeable strains examined by Mhlanga-Mutangadura and coworkers contained the hicAB genes but not the hifa genes that encode the hemagglutinating pilus.

<table>
<thead>
<tr>
<th>Contig</th>
<th>E value</th>
<th>Hit</th>
<th>Database reference</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>153</td>
<td>1.00E-129</td>
<td>Adhesin</td>
<td>gb</td>
<td>AAA20524.1</td>
</tr>
<tr>
<td>502</td>
<td>1.00E-104</td>
<td>Adhesin</td>
<td>gb</td>
<td>AAA20524.1</td>
</tr>
<tr>
<td>153</td>
<td>2.00E-07</td>
<td>Outer membrane protein A</td>
<td>pirs</td>
<td>T30852]</td>
</tr>
<tr>
<td>270</td>
<td>0</td>
<td>HmwA</td>
<td>gb</td>
<td>AAD56660.1</td>
</tr>
<tr>
<td>1</td>
<td>1.00E-30</td>
<td>HmwC</td>
<td>gb</td>
<td>AAFO0477.1</td>
</tr>
<tr>
<td>502</td>
<td>0</td>
<td>HmwC</td>
<td>gb</td>
<td>AAFO0477.1</td>
</tr>
<tr>
<td>545</td>
<td>0</td>
<td>HmwC</td>
<td>gb</td>
<td>AAFO0477.1</td>
</tr>
<tr>
<td>1</td>
<td>5.00E-37</td>
<td>Putative accessory processing protein</td>
<td>gb</td>
<td>AAA20525.1</td>
</tr>
<tr>
<td>421</td>
<td>2.00E-57</td>
<td>HicB</td>
<td>gb</td>
<td>AAC53810.1</td>
</tr>
<tr>
<td>421</td>
<td>3.00E-39</td>
<td>HicA</td>
<td>gb</td>
<td>AAC53809.1</td>
</tr>
<tr>
<td>505</td>
<td>0</td>
<td>Lava</td>
<td>gb</td>
<td>AAK76425.1</td>
</tr>
<tr>
<td>201</td>
<td>1.00E-20</td>
<td>Peroxiredoxin</td>
<td>ref</td>
<td>NP_759448.1</td>
</tr>
<tr>
<td>363</td>
<td>6.00E-05</td>
<td>Putative protein phosphatase</td>
<td>ref</td>
<td>NP_711609.1</td>
</tr>
<tr>
<td>398</td>
<td>1.00E-23</td>
<td>Putative cytoplasmic protein</td>
<td>ref</td>
<td>NP_463063.1</td>
</tr>
<tr>
<td>417</td>
<td>1.00E-05</td>
<td>Repressor protein-related protein</td>
<td>ref</td>
<td>NP_745902.1</td>
</tr>
<tr>
<td>118</td>
<td>1.00E-07</td>
<td>ATP-dependent DNA helicase recG</td>
<td>ref</td>
<td>NP_813441.1</td>
</tr>
<tr>
<td>501</td>
<td>7.00E-31</td>
<td>ATP-dependent DNA helicase recG</td>
<td>ref</td>
<td>NP_813441.1</td>
</tr>
<tr>
<td>110</td>
<td>5.00E-12</td>
<td>Type I restriction-modification system methyltransferase subunit</td>
<td>gb</td>
<td>EAA25229.1</td>
</tr>
<tr>
<td>118</td>
<td>8.00E-45</td>
<td>DNA methylase HsdM,putative</td>
<td>ref</td>
<td>NP_231401.1</td>
</tr>
<tr>
<td>118</td>
<td>1.00E-18</td>
<td>Probable restriction-modification system protein</td>
<td>ref</td>
<td>NP_251425.1</td>
</tr>
<tr>
<td>138</td>
<td>1.00E-148</td>
<td>Type II restriction enzyme HaeII(R. HaeII) (endonuclease HaeII)</td>
<td>sp</td>
<td>O07674</td>
</tr>
<tr>
<td>240</td>
<td>8.00E-27</td>
<td>Modification methylase Bepl (cytosine-specific methyltransferase Bepl)(M. Bepl)</td>
<td>sp</td>
<td>P0283</td>
</tr>
<tr>
<td>297</td>
<td>1.00E-158</td>
<td>DNA methylase HsdM,putative</td>
<td>ref</td>
<td>NP_231404.1</td>
</tr>
<tr>
<td>417</td>
<td>5.00E-06</td>
<td>Adenine-specific DNA methylase</td>
<td>ref</td>
<td>NP_292798.1</td>
</tr>
<tr>
<td>452</td>
<td>2.00E-98</td>
<td>Type I restriction-modification system methyltransferase subunit</td>
<td>gb</td>
<td>EAA23174.1</td>
</tr>
<tr>
<td>501</td>
<td>0</td>
<td>Type I restriction enzyme HsdR,putative</td>
<td>ref</td>
<td>NP_231400.1</td>
</tr>
</tbody>
</table>

*Table 2. Protein similarities identified in the strain 86-028NP three-fold contig set*
glutinating pilus. The strain 86-028NP sequence was identical in this region to the sequence in NTHI strain R3001, a strain that lacks the hif gene cluster (Fig. 3D).

The rfaD gene encodes an enzyme involved in the biosynthesis of endotoxin. Nichols and coworkers characterized the rfaD gene from NTHI strain 2019 (40). In strain 2019, the rfaD gene is immediately upstream of the rfaF gene, a gene that encodes an additional enzyme involved in endotoxin biosynthesis. The gene arrangement in strain Rd is different; the rfaD and rfaF genes are separated by approximately 11 kb of se-
quene. Nichols and coworkers noted that most nontypeable strains examined contained the gene arrangement found in strain Rd 2019. In contrast, strain 86-028NP had a gene arrangement identical to that found in strain Rd (Fig. 3E).

Both strains had genes homologous to the HaeII restriction system found in H. influenzae biogroup aegyptius. A number of additional sequences with homology to different restriction systems were identified in both gene sets (Tables 1 and 2).

In summary, using microarray analysis, the genome of a NTHI isolate was analyzed and several genes were identified that were not present in strain Rd. We also partially sequenced the genome of a second strain. The assembly currently contains 576 contigs, many with regions of low coverage. A global analysis of the current assembly indicated that the gene content and order were similar to that in strain Rd. A more detailed analysis revealed that there were a substantial number of genes not previously found in the Pasteurellaceae and some regions where the gene content and order were different than that found in strain Rd. These data suggested that the strain 86-028NP genome would contain a complex mosaic of strain Rd and non-strain-Rd-like features.

Both approaches to comparative genome analysis identified genes present in NTHI strains relative to strain Rd, demonstrating the complementary nature of these approaches, as well as the genetic similarity of the two strains under investigation. These approaches are both survey methods, limited in precision by the physical process of microarray analysis and by the statistical parameters chosen for the bioinformatics analysis. Nonetheless, as survey methods they have identified a number of verifiable differences between the genomes under investigation, while not forecasting an unmanageable level of false positives. They require neither exhaustive sequencing of the genomes nor massive replication of the microarray experiments. Moreover, the specificity and selectivity of the methods were both sufficiently good that putative virulence genes were identified in the NTHI strains, genes that may be important in understanding the pathogenesis of OM and perhaps other NTHI-induced diseases of the respiratory tract.

The threefold-contig data are available on our Web site at http://www.microbial-pathogenesis.org. Updated data will be made available as the genome sequence is completed.

ACKNOWLEDGMENTS

We thank Huachun Zhong for excellent technical assistance. This work was supported by National Institutes of Health grants R01 DC03915 (to L.O.B.), R01 P20-RR15564 (to D.W.D.), and P20-RR16478 (to D.W.D.). The DNA Sequencing Core at Columbus Children’s Research Institute was supported in part by the National Institutes of Health grant HD34615.

REFERENCES


