INFECTION AND IMMUNITY, June 2006, p. 3366–3374 Vol. 74, No. 6
0019-9567/06/$08.00+0 doi:10.1128/IAI.01904-05

Differential Genome Contents of Nontypeable *Haemophilus influenzae* Strains from Adults with Chronic Obstructive Pulmonary Disease

Matthew M. Fernaays,[1] Alan J. Lesse,[2,3,5] Sanjay Sethi,[2,5] Xueya Cai,[4] and Timothy F. Murphy[1,2,5]*

Departments of Microbiology,[1] Medicine,[2] Pharmacology and Toxicology,[3] and Biostatistics,[4] University at Buffalo, State University of New York, and VA Western New York Healthcare System,[5] Buffalo, New York

Received 21 November 2005/Returned for modification 10 January 2006/Accepted 10 March 2006

Haemophilus influenzae **is an important cause of otitis media in children and lower respiratory infection in adults with chronic obstructive pulmonary disease (COPD). Patients with COPD experience periodic exacerbations that are associated with acquisition of new bacterial strains. However, not every strain acquisition is associated with exacerbation. To test the hypothesis that genetic differences among strains account for differences in pathogenic potential, a microarray consisting of 4,992 random 1.5- to 3-kb genomic fragments of an exacerbation strain was constructed. Competitive hybridization was performed using six strains associated with exacerbation as well as five strains associated with asymptomatic colonization. Seven sequences that were absent in all five colonization strains and present in at least two exacerbation strains were identified. One such sequence was a previously unreported gene with high homology to the meningococcal immunoglobulin A (IgA) protease gene, which is distinct from the previously described *H. influenzae* IgA protease. To assess the distribution of the seven sequences among well-characterized strains of *H. influenzae*, 59 exacerbation strains and 73 asymptomatic colonization strains were screened by PCR for the presence of these sequences. The presence or absence of any single sequence was not significantly associated with exacerbations of COPD. However, logistic regression and subgroup analysis identified combinations of the presence and absence of genes that are associated with exacerbations. These results indicate that patterns of genes are associated with the ability of strains of *H. influenzae* to cause exacerbations of COPD, supporting the concept that differences in pathogenic potential are based in part on genomic differences among infecting strains, not merely host factors.**

Haemophilus influenzae is a gram-negative coccobacillus that has an ecological niche in the human respiratory tract (16). Nonencapsulated or nontypeable *H. influenzae* is an important respiratory pathogen (35). In addition to being a leading cause of otitis media (34, 35), nontypeable *H. influenzae* also plays an important role in the course and pathogenesis of chronic obstructive pulmonary disease (COPD) (2, 3, 35, 37, 44). COPD is the fourth leading cause of death in the United States (4). The natural history of the disease is characterized by intermittent exacerbations, or periods of increased sputum production and dyspnea, which are associated with increased morbidity and mortality (3, 4, 44). Nontypeable *H. influenzae* is the most commonly isolated organism from the lower airways of adults with COPD and is the most common bacterial cause of exacerbations (37, 44, 53).

Acquisition of a strain of nontypeable *H. influenzae* by an adult with COPD can result in a variety of outcomes. Some episodes of acquisition are associated with symptoms of acute exacerbation of COPD, while others are associated with no perceptible worsening of the patient's clinical condition (43). On some occasions, strains are quickly cleared from the respiratory tract in 1 month or less, while at other times, persistent colonization of the respiratory tract is observed for many months to years (36). An immune response is sometimes associated with acquisition of a new strain, while no such association is seen at other times (45). We hypothesize that differences among strains of *H. influenzae* contribute to the diverse presentations of infection in adults with COPD.

One source of diversity among strains of *H. influenzae* is the presence or absence of certain genes that tend to increase the organism's fitness in a specific niche within the human host (16). A recently proposed hypothesis posits that as a species, *Haemophilus influenzae* possesses a supragenome from which individual strains obtain a subset of genes that confer optimal adaptation to a particular environment (46). Associations of certain genes to virulence are clear, as pathogenicity islands that are found in meningitis-causing strains of *H. influenzae* type b have been identified (5, 9). In addition, genomic comparison of nontypeable strains of *H. influenzae* from patients with otitis media and the avirulent sequenced strain, KW20 Rd, have identified several genes associated with otitis media (18, 33). A similar genomic comparison followed by molecular epidemiological examination of strains of *H. influenzae* from the middle ear and the nasopharynx associated the lipooligosaccharide biosynthesis gene *lic2B* with otitis media (39). We hypothesize that, like in otitis media-causing strains of *H. influenzae*, certain sequences in the genomes of *H. influenzae* strains isolated from adults with COPD are associated with the diverse clinical outcomes of infection with *H. influenzae* in these patients.

MATERIALS AND METHODS

Bacterial strains. Strains of nontypeable *H. influenzae* were isolated from the sputum of adults with COPD in an ongoing, prospective study of COPD at the Buffalo VA Medical Center. The Human Studies Subcommittee of the Veteran's Affairs Western New York Healthcare System approved the study protocol. All participants gave written informed consent. Inclusion criteria for study partici-

* Corresponding author. Mailing address: VA Western New York Healthcare System, Medical Research 151, 3495 Bailey Avenue, Buffalo, NY 14215. Phone: (716) 862-7874. Fax: (716) 862-6526. E-mail: murphyt@buffalo.edu.

pants were presence of chronic bronchitis, the absence of asthma or bronchiectasis, an ability to comply with a schedule of monthly clinical visits, and the absence of immunosuppressive or other life-threatening disorders (43).

Patients were seen monthly and at the time of suspected exacerbation. A clinical evaluation was performed to determine whether the patient was experiencing an exacerbation or was clinically stable, as previously described (43). The patients were questioned about the status of their chronic respiratory symptoms (dyspnea, cough, sputum production, viscosity, and purulence), and the responses were graded as 1 (at the usual level), 2 (somewhat worse than usual), or 3 (much worse than usual). A minor worsening of two or more symptoms or a major worsening of one or more symptoms prompted a clinical assessment of the cause. If the patient had fever, appeared ill, or had signs of consolidation on examination of the lungs, a chest X-ray was obtained to rule out pneumonia. If other causes of the worsening of the symptoms, such as pneumonia, upper respiratory tract infection, and congestive heart failure, were ruled out, the patient was considered to be having an exacerbation of COPD. The determination of whether the patient had stable disease or an exacerbation was made before the results of sputum cultures were available.

Sputum samples obtained during these visits were subjected to semiquantitative culture. Strains of *H. influenzae* were identified by standard methods. When *H. influenzae* was present in the sample, 10 colonies were isolated and subjected to typing by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). A strain was considered newly acquired if the unique SDS-PAGE profile was not previously seen in any strains isolated from that patient. Exacerbation strains were defined as newly acquired strains that were initially isolated at the time of an exacerbation of COPD. Colonization strains were defined as those strains that were not associated with symptoms of exacerbation upon initial acquisition of the strain.

The strain used in microarray construction and sequencing was 11P6H. This strain was initially isolated during an acute exacerbation of COPD in a study patient. The patient developed an immune response to surface antigens of the homologous infecting strain as detected in whole-cell enzyme-linked immunosorbent assay and bactericidal assay (45). Whole-cell radioimmunoprecipitation determined that this response was in part directed toward the P2 outer membrane protein (55). This strain was not isolated from sputum in subsequent clinic visits, indicating that the strain was cleared from the patient's respiratory tract.

Microarray construction. Genomic DNA was purified from 11P6H using the Wizard genomic DNA purification kit (Promega, Madison, WI). Following shearing by sonication, 1.5- to 3-kb fragments were isolated by excision from an agarose gel. DNA was recovered using the QIAGEN gel extraction kit. The DNA was subjected to an end repair reaction, followed by dephosphorylation, phenol-chloroform extraction, and ethanol precipitation. The DNA was then ligated into the pCR4-Blunt TOPO vector (Invitrogen, Carlsbad, CA) following the manufacturer's instructions, and TOP 10 *Escherichia coli* cells were transformed with the resulting constructs by electroporation. Approximately 10,000 transformants were screened by PCR with universal primers that amplified the plasmid insert. Agarose gel electrophoresis was used to identify clones with 1.5- to 3-kb inserts in the vector. A total of 4,992 clones with 1.5- to 3-kb inserts were verified and individually archived.

Plasmid DNA was isolated from each clone using an Autogen 740 automated DNA isolation system. The inserts were amplified from the plasmid DNA by PCR, verified for proper amplification by agarose gel electrophoresis, precipitated with ethanol, dehydrated, and resuspended in 20% dimethyl sulfoxide. The DNA fragments were then spotted on glass slides (type A; Schott Glas) by using a MicrogridII TAS arrayer and MicroSpot 2500 split pins (Apogent Discoveries, Hudson, NH) in triplicate on three different regions of the slide. Irrelevant and unique genes from *Moraxella catarrhalis* were also spotted as negative controls. Quality control was performed on the slides by visual and microscopic inspection followed by test hybridization with reference strain DNA.

Microarray hybridization. Genomic DNA was isolated from the microarray strain, 11P6H, as well as five exacerbation strains and five colonization strains with a Wizard genomic DNA purification kit (Promega). DNA from each of the 10 experimental strains was then competitively hybridized with 11P6H DNA to the microarray in 10 separate experiments at the Roswell Park Cancer Institute Microarray and Genomics Center as previously described (10). Briefly, 1 μg 11P6H DNA was random prime labeled with the fluorescent nucleotide analogue Cy3 dUTP, while 1 μg of each DNA preparation from the competing strains was labeled individually with Cy5 dUTP. The labeled DNA probes were then ethanol precipitated and resuspended in 110 μl SlideHyb buffer 3 (Ambion, Inc., Austin, TX) with 80 μg salmon sperm DNA (Invitrogen). The probe solution was heated to 95°C for 5 min, cooled on ice, and then loaded on the microarray slide. Hybridization was allowed to proceed for 16 h at 55°C in a GeneTAC hybridization station (Genomics Solutions, Ann Arbor, MI), after which the slide was washed with decreasing concentrations of SSC (sodium chloride-sodium citrate) and SDS, ethanol rinsed, and dried by centrifugation. The slides were scanned using a GenePix 4200A scanner (Axon, Inc. [now Molecular Devices Corporation], Sunnyvale, CA) to generate high-resolution images for both the Cy3 and Cy5 channels. The images were analyzed using ImaGene software, version 4.1 (Biodiscovery, El Segundo, CA). Each spot was background corrected for nonspecific probe binding and scored for reliability, as previously described (10). For all spots that passed quality control, the $\log_2$ test/control ratio (Cy5/Cy3) was calculated. The values were then normalized based on the ratios of 50 fragments whose values suggested they were present in all strains examined. The replicate measures were averaged for statistical analysis.

Fragment sequence analysis. For fragments of interest, the insert was amplified by PCR from the original clone and the sequence was determined. The sequences were aligned using Sequencher (version 4.5; Gene Code Company). BLASTN analysis was performed for each of the resulting contigs against the NCBI database. Those sequences that did not have a significant nucleotide match were subjected to open reading frame (ORF) analysis with MacVector (version 7.2.2; Accelrys Inc.). All ORFs greater than 25 amino acids in size were then subjected to BLASTP analysis against the NCBI database.

Screening of strains for homologous sequences. Those sequences that were of interest were used in a PCR-based screening method. Crude lysates were made from strains of *H. influenzae* isolated from the sputum of adults followed in the COPD study clinic using previously described methods (19, 24, 27, 28). Briefly, bacteria grown overnight on chocolate agar plates were harvested with a sterile loop and suspended in 100 μl sterile water by vortexing. The bacteria were incubated at 100°C in a heat block for 5 min, resuspended by vortexing, and incubated at 100°C for an additional 5 min. The samples were centrifuged for 1 min at 16,000 × *g*, and the supernatant was saved for use as the template in PCR. Internal primers were designed for each sequence of interest based on the nucleotide sequence of strain 11P6H at each region. PCRs were carried out under the following conditions: 10 min at 94°C; 30 cycles of 30 seconds at 94°C, 30 seconds at 55°C, and 90 seconds at 72°C; followed by 3 min at 72°C. Samples were prepared identically for the following strains to be used as positive and negative controls in the PCR screening experiments: *H. influenzae* 11P6H (positive control), the sequenced *H. influenzae* strain KW20 Rd (positive or negative control, depending on the sequence), and *Moraxella catarrhalis* strain 43617 (negative control). Water controls were also included in the analysis.

Statistical analysis. Univariate analyses to identify associations between the presence or absence of each sequence and the classification as an exacerbation strain or a colonization strain were performed with Fisher's exact test and chi-square analysis.

Correlations between sequences were analyzed using SAS (version 9.1; SAS Institute Inc.). SAS was also used to perform logistic regression analysis on the entire data set as well as two subsets. This approach disregarded one and then the other of two sequences in two separate analyses, a procedure employed to address the problem of multicolinearity in the data set. Bonferroni's method was used for type I error adjustment in the successive logistic regression analyses that disregarded specific sequences.

Cluster analysis was performed with the JMP statistical package (version 5.0.1a; SAS Institute Inc.). The hierarchical cluster was performed using the average method and two-way clustering without standardizing data. Fisher's exact test comparing specific branches of the cluster with the rest of the population was used to identify clades of interest. Differences between two clades of interest with respect to sequence presence and absence and the association of specific sequences with exacerbation within the 29-member subset of strains from these two clades were analyzed using Fisher's exact test.

RESULTS

Competitive genomic hybridization. To identify sequences associated with exacerbations of COPD, a microarray of 4,992 1.5- to 3-kb fragments was constructed from the genomic DNA of strain 11P6H of *H. influenzae* that caused an exacerbation of COPD. This microarray was used in a series of competitive hybridization experiments in which genomic DNA from the microarray strain was labeled with one fluorescent marker and genomic DNA from 1 of 10 other strains (5 exacerbation strains and 5 colonization strains) was labeled with another marker. For each experiment, normalized $\log_2$ ratios were obtained for all microarray fragments. These ratios indicate the

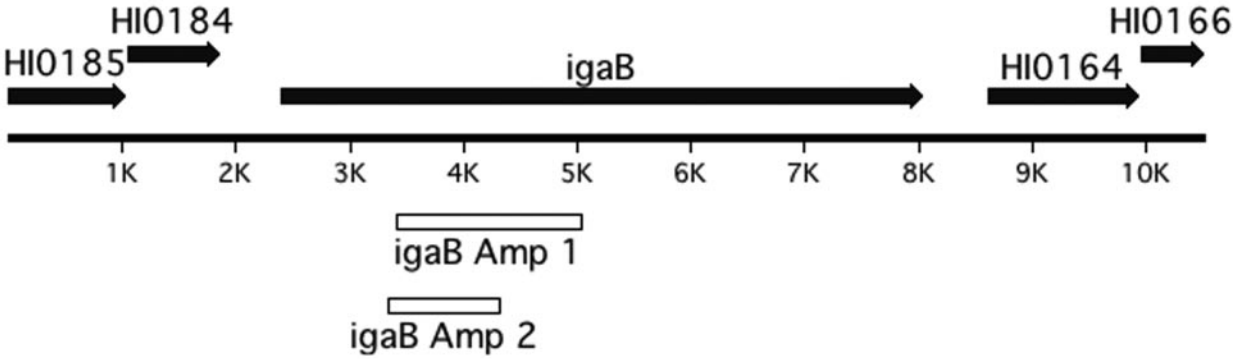


FIG. 1. Diagram of the locus containing the novel IgA protease gene, *igaB*, in *H. influenzae* strain 11P6H. On either end of the region are known genes HI0164, HI0166, HI0184, and HI0185, identified in strain KW20 Rd. Empty boxes depict the amplicons used to screen multiple strains.

level of hybridization of the experimental strain and the reference strain to each fragment. That is, for a $\log_2$ ratio of 0, both strains hybridized to the fragment on the slide with equal efficiency. Thus, the sequence is expected to be found in both strains. As the ratio decreases from zero, this indicates that DNA from the experimental strain hybridizes less efficiently, indicating that there is lower homology for that sequence of DNA in the experimental strain, or, if the ratio is low enough, that the sequence is absent. The ratios should not rise significantly above zero, because the fragments on the slide came from the reference strain; thus, the sequences are all present and identical in labeled genomic DNA from this strain.

A control hybridization in which two aliquots of 11P6H DNA were labeled with different fluorescent markers was performed to define cutoff values for significance of the experimental hybridization ratios. In this experiment, the $\log_2$ ratios clustered tightly about the mean value of 0.003894 with a standard deviation of 0.10351. Nearly all values for the 4,992 fragments fell within three standard deviations of the mean. Based on this experiment, a fragment was considered likely present in the experimental strain if the competitive hybridization ratio was within three standard deviations of this mean or had a ratio above −0.306636. Further, a sequence was considered likely absent in the strain if the hybridization ratio was below eight standard deviations from the mean, or below −0.824186. This low value was chosen to avoid fragments whose ratios were below that seen in the control experiment because of decreased homology in a sequence that was present in the experimental strains. This stringent criterion could miss single gene differences in a fragment that spans multiple genes. This effect was minimized by constructing a microarray of 4,992 fragments, a number estimated to include a fivefold coverage of the genome. This approach, combined with the use of variable fragment sizes, increased the likelihood that single gene differences would be detected.

To identify fragments that were present in exacerbation

TABLE 1. Sequences identified in microarray experiments and used as targets in screening with PCR, with known homologies after NCBI database search

Sequence name and GenBank accession no.	G+C content (%)	Gene content and homology	e value (GenBank accession no. for match)		Size of gene or sequence (bp)
			BLASTn	BLASTp	
HI0568, DQ423213	41.7	Known *H. influenzae* gene; transcription accessory protein (tex)	0.0 (L42023)		2,313
HI0696, DQ423215	38.3	Known *H. influenzae* gene; "conserved hypothetical protein"	0.0 (L42023)		3,937
HI0698, DQ423215	39.8	Known *H. influenzae* gene; "conserved hypothetical protein"	0.0 (L42023)		1,752
HC, DQ423214	33.8	Known *H. influenzae* gene; hypothetical protein	0.0 (AY599455)		759
igaB, DQ423203	42.4	IgA protease; high homology to *N. meningitidis*	0.0 (AF012206)	0.0 (YP_207437)	5,664
183UM, DQ423205	29.7	10 ORFs with >25 amino acids (aa); none with homology			785
553UM, DQ423212	30.6	227-aa ORF; hypothetical *H. influenzae* protein		10^{-125} (ZP_00349643)	1,027
707UM, DQ423216	34.8	>364-aa ORF; conserved hypothetical protein, many organisms, highest is *Oceanicola batsensis*		9×10^{-40} (ZP_01000803)	1,485
1069-1070UM, DQ423219	31.8	294-aa ORF; *Streptococcus mitis* phage SM1 gp111		10^{-8} (AAP81893)	1,556
89-90UM, DQ423204	32.1	>120-aa ORF; conserved hypothetical *N. meningitidis* protein		3×10^{-32} (AAF41647)	574

strains and absent in colonization strains, fragments that had ratios that were below the absence cutoff in all five colonization strains but above the presence cutoff in at least one of the five exacerbation strains were identified. These criteria would identify fragments in at least two exacerbation strains, as the strain used to construct the microarray (11P6H) is an exacerbation strain as well. This approach proved necessary because the diversity of the exacerbation strains resulted in no strains meeting the absence criterion in all colonization strains and the presence criterion in all exacerbation strains. A total of 110 of 4,992 fragments met these inclusion criteria of being absent in five of five colonization strains and present in at least two of six exacerbation strains.

These fragments were amplified by PCR from the strain 11P6H library, and their sequences were determined. The sequences were aligned into 24 contigs. The sequences for these contigs are available in the GenBank database under accession numbers DQ423203 to DQ423226. BLASTN searches were performed, comparing the sequences to known genes in the NCBI database. There were 57 genes represented on the contigs either in part or in whole (as defined by homology to known *H. influenzae* genes in the NCBI database). Additionally, of note was the identification of a gene, previously unreported in *H. influenzae*, with >90% nucleotide identity to the secreted portion of the IgA protease (*iga*) gene in *Neisseria meningitidis*. This gene is distinct from previously described IgA proteases in *H. influenzae* (40) and is not associated with *H. influenzae* in the NCBI database. Preliminary experiments have shown that this gene is present in addition to the previously identified IgA protease (called *iga* or *iga1*) known in *H. influenzae* (7). Figure 1 displays the layout of the locus of the novel *iga* protease gene, in relation to known *H. influenzae* genes. We propose the name *igaB* for this novel IgA protease gene.

Further, five sequences of at least 574 bp with no significant nucleotide database match were identified. These were subjected to ORF analysis using the MacVector program. ORFs that were greater than 25 amino acids in length were subjected to BLASTP analysis. Four sequences (89-90UM, 553UM, 707UM, and 1069-1070UM) showed homology to known translated nucleotide sequences, while one (183UM) did not (Table 1).

The G+C content of these unmatched sequences was also determined (Table 1). The G+C content of the sequenced strains of *H. influenzae* (http://www.ncbi.nlm.nih.gov/) ranges from 38.0% to 38.2% (14, 15, 18). The G+C contents of most of the sequences identified in this study range from 29.7% to 34.8%. The exception is the *igaB* gene, whose G+C content is 42.4%. The G+C content of a sequenced *Neisseria meningitidis* strain is 51.6%, and the content of the *iga* gene of this strain is 46.8% (52).

Selection of genes for screening. Twenty of 63 sequences identified in the previously described sequence analysis were chosen for further study because a majority of the gene sequence was found on the 24 contigs (Fig. 2). Forty-three sequences were not studied further because a large proportion of their expected sequences was not on the contigs identified in the microarray analysis. That is, only a small portion of each of these genes was included on the fragments that met the inclusion criteria (Fig. 3). The 20 sequences chosen for further study underwent a small pilot screen with six exacerbation strains and five colonization strains from the COPD study clinic (those used in the microarray analysis) as well as the sequenced strain, KW20 Rd, to preliminarily identify genes that are present in some strains but not in others. Ten of these sequences were present in all 12 strains and were thus not studied further. The 10 sequences that were present in some strains but not others were selected to undergo a large-scale screen of strains of

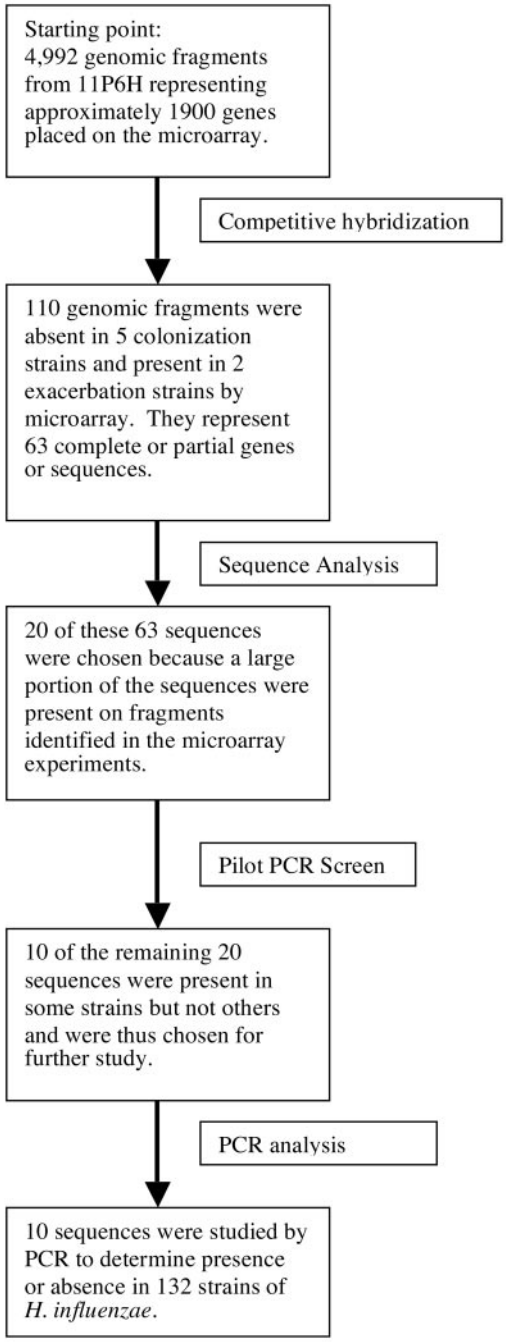


FIG. 2. Flow diagram outlining the process of selection of sequences for screening of 132 strains of *H. influenzae* from adults with COPD.

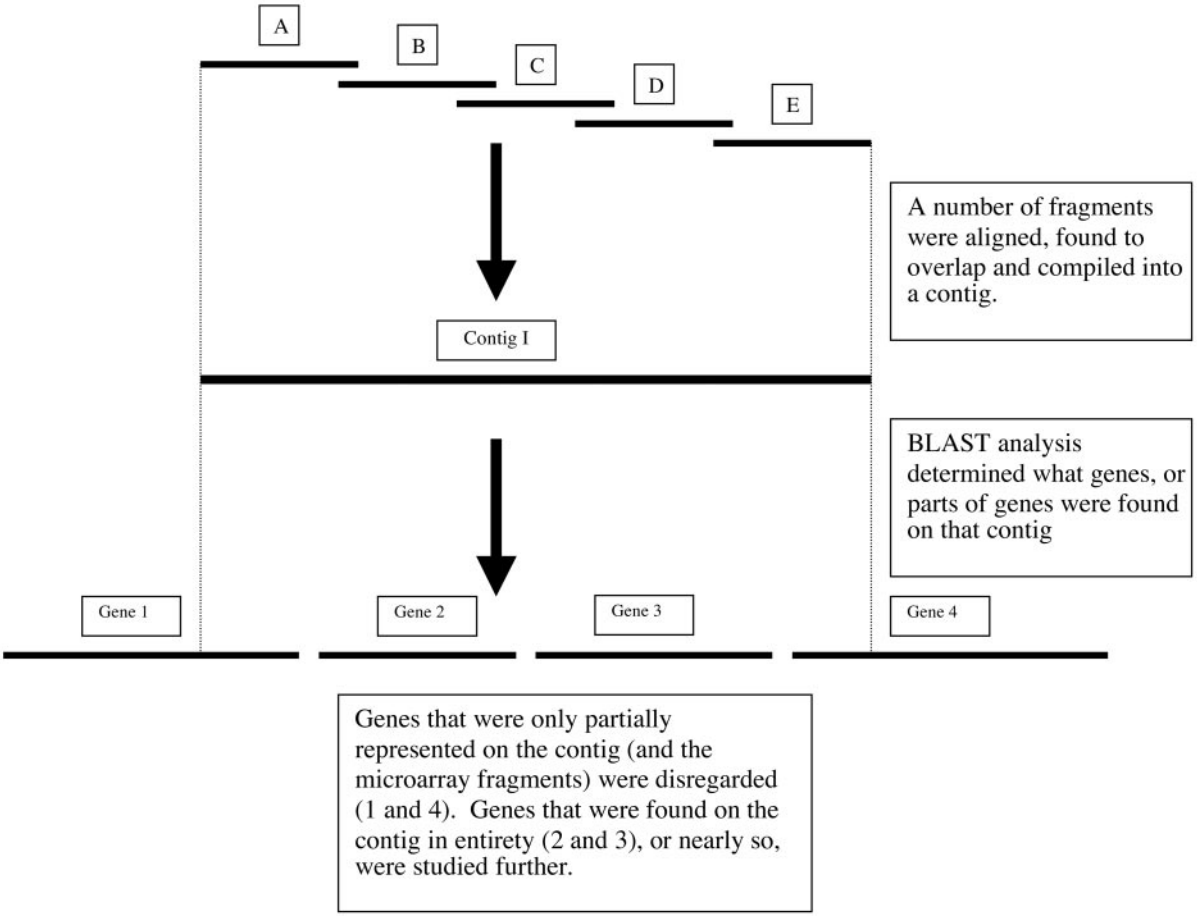


FIG. 3. Diagram outlining the process of identifying hypothetical sequences for further study.

H. influenzae from adults with COPD. These sequences are listed in Table 1.

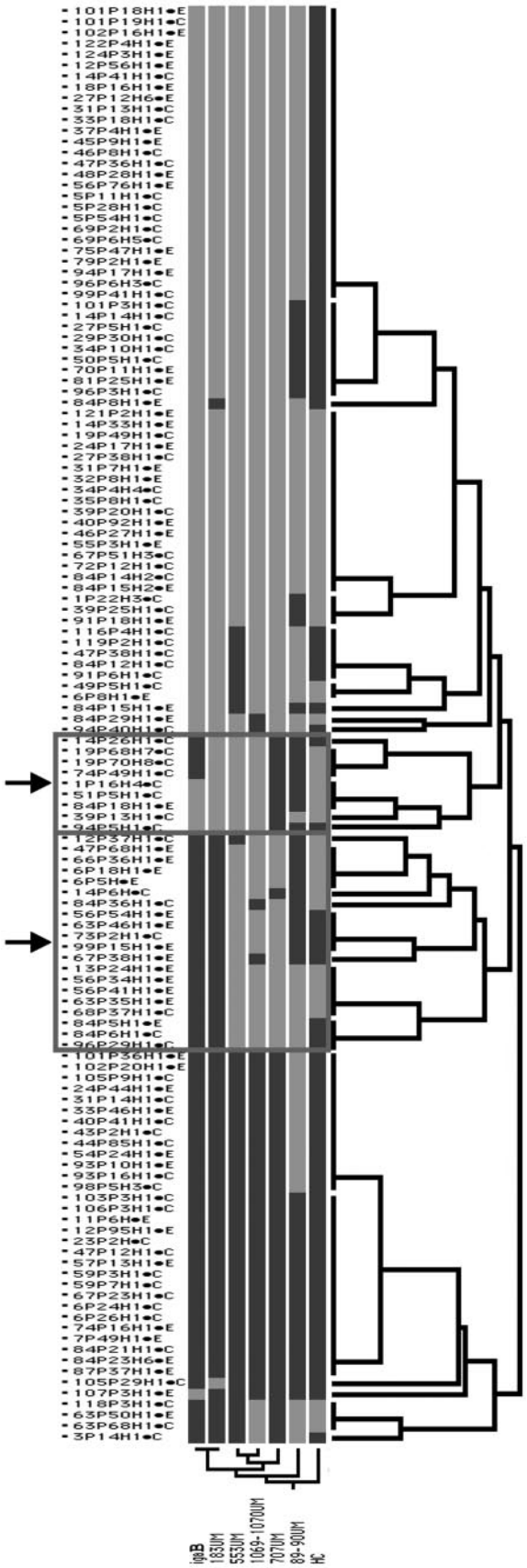
Screening of strains from adults with COPD. A total of 132 consecutive well-characterized isolates of *H. influenzae* collected from the COPD study clinic from 1994 to 2003 were screened for the presence or absence of 10 selected genes or sequences using PCR. Ten sets of internal primers were designed for the selected sequences. Each assay contained control amplification reactions using template DNA from strain 11P6H (positive), the sequenced strain KW20 Rd (positive or negative, depending on the sequence in question), *Moraxella catarrhalis* sequence strain 43617 (negative), and water (negative). The screen was performed in duplicate on separate days, with all strains scored for the presence or absence of amplification of the predicted size fragments detected by agarose gel electrophoresis. When replicates of a strain-primer combination did not agree (less than 2% of all reactions), additional reactions were carried out with either purified genomic DNA or a second lysate preparation. The final determination was made based on these repeated reactions. In this screen, three sequences were present in all strains tested. The rest of the sequences were variably present in the population of strains used in the screen (Table 2).

To increase the sensitivity of the assay for these targets, an additional set of primers from different regions of each sequence was designed for the remaining seven target sequences. The screen was performed again in duplicate with these additional primers. For those sequences that were studied with two primer sets, a strain was considered positive if there was amplification with either primer set and negative if both primer sets resulted in no amplification.

TABLE 2. Results of screen of 132 strains of *H. influenzae* from COPD patients for selected genes

Gene or sequence name	Amplicon size(s) (bp)	No. of positive strains (n = 132)	% Positive		
			All strains tested	Exacerbation strains	Colonization strains
igaB[a]	1,650, 1,011	59	44.7	45.8	43.8
183UM[a]	723, 521	56	42.4	49.2	37.0
553UM[a]	807, 657	45	34.1	28.8	38.4
HC[a]	686, 489	87	65.9	62.7	68.5
707UM[a]	934, 1,045	46	34.8	27.1	41.1
1069-1070UM[a]	799, 455	36	27.3	27.1	27.4
89-90UM[a]	351, 266	52	39.4	35.6	42.5
HI0568	981	132	100	100	100
HI0696	948	132	100	100	100
HI0698	852	132	100	100	100

[a] A second set of primers was designed, and the screen was repeated for these sequences. The two resulting amplicon sizes are reported.



Analysis of screening results. Results of PCR screening for each gene or sequence were compared to clinical data to determine if there was an association between the presence or absence of individual genes and exacerbations of COPD. Using Fisher's exact test and chi-square analysis, no significant association of the presence or absence of any of these seven sequences individually with exacerbation was observed, although a trend toward association of the absence of 707UM with exacerbation was observed ($P = 0.094$).

Linkage between sequences was analyzed using Spearman correlation coefficients (from SAS, version 9.1). There were multiple significant correlations in the data set. The most significant was between *igaB* and 183UM, with a coefficient of 0.893 (95% confidence interval, 0.815 to 0.970). In 94.7% of strains, these sequences were either both present or both absent.

Logistic regression analysis conducted on the data set indicated no significant associations between individual sequences and exacerbation. This result was likely due to the high degree of multicolinearity between many of the genes in the data set. The analysis was repeated twice, disregarding first *igaB* and then 183UM, the most highly correlated sequences in the data set. The first of these showed significant association of 183UM with exacerbation ($P = 0.0177$) and a trend toward significance between the absence of 707UM and exacerbation ($P = 0.0601$). The second analysis showed an association between the absence of 707UM and exacerbation ($P = 0.0461$). Following type I error adjustment using Bonferroni's method, it was concluded that 183UM is associated with exacerbation ($P = 0.035$), and the absence of 707UM shows a trend toward significant association with exacerbation ($P = 0.092$).

To further investigate whether patterns of genes were associated with exacerbations, data from the seven sequences were subjected to a cluster analysis using the JMP software package. This application arranged each of the strains into trees based on the relatedness of patterns of sequence presence or absence (Fig. 4). Within this cluster, one clade ($n = 20$) had a predominance of exacerbation strains ($P = 0.05$), while another ($n = 9$) was predominantly composed of colonization strains ($P = 0.04$). These subgroups were statistically compared to each other to determine if there were differences between the two with respect to sequence presence. The exacerbation clade was associated with the presence of *igaB* ($P = 0.001$) and 183UM ($P < 0.001$), while the colonization clade was associated with the presence of 707UM ($P < 0.001$). The two-way clustering component of this analysis confirmed the previously discussed correlation between 183UM and *igaB*. Further, when these 29 strains were analyzed separately, 183UM was associated with exacerbation ($P = 0.014$) and the absence of 707UM was associated with exacerbation ($P = 0.005$) with Fisher's exact test.

FIG. 4. Cluster alignment showing 132 strains of *H. influenzae* grouped according to patterns of sequence presence and absence. Numbers along the left represent the 132 strains. The sequences are noted at the bottom. Dark rectangles indicate a sequence is present, and light rectangles indicate a sequence is absent. Arrows denote two clades in boxes. The upper box is associated with colonization. The lower box is associated with exacerbation.

DISCUSSION

This study utilized competitive hybridization microarrays followed by PCR screening to identify several genes and sequences that were variably present in strains of *H. influenzae* from adults with COPD. Some of these sequences showed a high degree of homology to previously identified *H. influenzae* genes, including what is referred to in this study as HC (homologous to GenBank AY599455, 99% nucleotide identity) and 553UM (homologous to GenBank ZP_00349643, 94% amino acid identity and 96% similarity). Both of these entries are classified hypothetical proteins. Four additional sequences identified in this study are not present in the four strains of *H. influenzae* whose genomes have been sequenced, including 183UM, 89-90UM, 707UM, and 1069-1070UM. All but the first contain ORFs that show amino acid homology to known conserved hypothetical genes. These sequences have lower G+C content than the sequenced *H. influenzae* strains, suggesting that they may be associated with horizontal transfer events. Finally, this study identified a novel IgA protease gene that has not been described in *H. influenzae*. This gene, which is present in addition to the known IgA protease of this organism, shows a high degree of homology, particularly in the secreted protease domain, with the *iga* gene of *Neisseria meningitidis*. Further, the G+C contents of these genes indicate that horizontal transfer from *Neisseria* to *H. influenzae* is likely. Transfer from *Haemophilus* to *Neisseria* has been described in previous studies (11, 31), but transfer in the opposite direction has not been described previously.

No statistically significant association between individual genes and exacerbation of COPD was observed in the univariate analysis. However, logistic regression as well as cluster and subgroup analyses indicated that the specific combination of presence of 183UM, presence of *igaB*, and absence of 707UM was associated with exacerbation of COPD.

These findings are not surprising, given the degree of redundancy in *H. influenzae* and other bacterial systems. For example, *H. influenzae* expresses a number of adhesins, including pilus, Hia, Hap, HMW1, HMW2 (41, 48–50), and P5 fimbrin (32), as well as other molecules that promote adhesion to host tissue surfaces (41, 48–50). Each of these products recognizes different host factors and can be variably present among strains. Yet, all of these factors contribute to the vital role of adherence in pathogenesis. The results of the present study resemble this model in that while single genes were not associated with clinical outcomes, groups of genes were associated.

That there were no univariate associations between genes and exacerbation is not an unexpected result, as there are, undoubtedly, more genes involved in exacerbation that were not accounted for in this analysis. Furthermore, several other factors are known to be important determinants of clinical exacerbation and thus are potential confounding variables in the present study, blunting differences between exacerbation and colonization strains. For example, several host factors are associated with exacerbation, including underlying lung function (8, 30, 47), lymphocyte responsiveness (1), levels of cytokines, including interleukin-6 and interleukin-8 (6), and levels of a variety of host proteins, including mannose binding lectin 2 (54), secretory leukocyte protease inhibitor (17), and salivary lysozyme (51). Acquisition of a bacterial strain is only one cause of exacerbation in patients with COPD. Viral infection as well as environmental insults can also trigger a worsening of disease symptoms and, thus, may have confounded the results of this study to some extent by allowing for the designation of a strain of *H. influenzae* that was concomitantly acquired as an exacerbation strain when, on its own, it may have merely colonized the host. In spite of these confounders, inherent in studies of COPD, statistically significant differences in patterns of gene content in exacerbation and colonization strains were observed.

An interesting component of these data is the observation that the absence of some sequences as part of a combination of genes was associated with exacerbation, despite the acceptance criteria in analysis of microarray data that were designed to identify sequences positively associated with exacerbation. This result was likely due to the small sample size of strains studied in the microarray experiments ($n = 11$) and the fact that accepted fragments needed to be present in only two exacerbation strains, for reasons previously outlined. We speculate that these gene products may be associated with enhanced immune responses resulting in clearance of strains with these genes. Alternatively or additionally, those sequences may encode products that cause interference with other virulence factors or may contain control elements that negatively regulate other genes. Analysis of the functions of these novel sequences in future work will test these hypotheses.

An interesting observation was the association of 183UM with exacerbation, despite the absence of ORFs greater than 41 amino acids in size on this fragment. This fragment may contain genes that encode small peptides that contribute to pathogenesis. This sequence could also contain promoter regions or control elements that affect transcription of other genes. We propose an alternate hypothesis that this sequence serves as a marker for the acquisition of the novel IgA protease gene. Examination of the genomic milieu of the IgA protease gene reveals that it is located between genes designated HI0164 and HI0184. In the sequenced strain KW20 Rd, these two genes are located more than 17 kb from each other and transcribed in opposite directions. By contrast, in strain 11P6H, the genes are less than 7 kb apart and transcribed in the same direction. It is likely that the integration of this gene was associated with a large genomic inversion event that left HI0183 in another part of the genome with a 750-bp novel DNA segment adjacent to it. This hypothesis is supported by the observation that 183UM and the novel IgA protease gene were either both present or both absent in most strains.

The present study used PCR to screen 132 strains of *H. influenzae* for the presence or absence of sequences, a method that has been effective in multiple species of bacteria (12, 13, 20–23, 25–27, 29, 38, 42). Potential limitations of this methodology should be considered, including lack of sensitivity due to single nucleotide polymorphisms causing a lower affinity of primer binding, as well as the potential for small amounts of DNA to contaminate reactions, resulting in lowered specificity. To maximize specificity, duplicate reactions were performed on different days and controls with irrelevant DNA and with water in place of template were included with each assay. Also, for sequences of interest, a duplicate set of primers was designed and the screen was repeated to minimize the impact of

primer site single nucleotide polymorphisms on the sensitivity of the screen.

In summary, this study used competitive hybridization of genomic DNA with a microarray and PCR screening of well-characterized strains of *H. influenzae* to identify genes associated with exacerbations of COPD. The approach was based on the hypothesis that isolates of *H. influenzae* differ in their ability to cause infections in the airways of adults with COPD. The results showed that patterns of sequences are associated with the ability of strains to cause exacerbations, supporting the hypothesis that differences in pathogenic potential are based on genetic differences among strains. Future studies will be directed toward characterizing the expression patterns of these genes and identifying the mechanisms by which these gene products participate in the pathogenesis of infection.

ACKNOWLEDGMENTS

We recognize the valuable help of Norma Nowak, Jeffrey Conroy, Devin McQuaid, Paul Quinn, W. Michael Henry, and Michael Bianchi of the Roswell Park Cancer Institute Microarray and Genomics Facility in the planning, execution, and analysis of the microarray experiments. James Johnson contributed valuable advice concerning techniques of molecular epidemiology. We also thank Brydon Grant for his help with statistical analysis and Thomas Russo for reviewing the manuscript.

This work was supported by NIH grant AI19641 and by the Department of Veterans Affairs.

REFERENCES

1. **Abe, Y., T. F. Murphy, S. Sethi, H. S. Faden, J. Dmochowski, Y. Harabuchi, and Y. M. Thanavala.** 2002. Lymphocyte proliferative response to P6 of *Haemophilus influenzae* is associated with relative protection from exacerbations of chronic obstructive pulmonary disease. Am. J. Respir. Crit. Care Med. **165:**967–971.
2. **Bandi, V., M. A. Apicella, E. Mason, T. F. Murphy, A. Siddiqi, R. L. Atmar, and S. B. Greenberg.** 2001. Nontypeable *Haemophilus influenzae* in the lower respiratory tract of patients with chronic bronchitis. Am. J. Respir. Crit. Care Med. **164:**2114–2119.
3. **Bandi, V., M. Jakubowycz, C. Kinyon, E. O. Mason, R. L. Atmar, S. B. Greenberg, and T. F. Murphy.** 2003. Infectious exacerbations of chronic obstructive pulmonary disease associated with respiratory viruses and nontypeable *Haemophilus influenzae*. FEMS Immunol. Med. Microbiol. **37:**69–75.
4. **Barnes, P. J.** 2000. Chronic obstructive pulmonary disease. N. Engl. J. Med. **343:**269–280.
5. **Bergman, N. H., and B. J. Akerley.** 2003. Position-based scanning for comparative genomics and identification of genetic islands in *Haemophilus influenzae* type b. Infect. Immun. **71:**1098–1108.
6. **Bhowmik, A., T. A. Seemungal, R. J. Sapsford, and J. A. Wedzicha.** 2000. Relation of sputum inflammatory markers to symptoms and lung function changes in COPD exacerbations. Thorax **55:**114–120.
7. **Bricker, J., M. H. Mulks, A. G. Plaut, E. R. Moxon, and A. Wright.** 1983. IgA1 proteases of *Haemophilus influenzae*: cloning and characterization in Escherichia coli K-12. Proc. Natl. Acad. Sci. USA **80:**2681–2685.
8. **Burge, P. S., P. M. Calverley, P. W. Jones, S. Spencer, J. A. Anderson, and T. K. Maslen.** 2000. Randomised, double blind, placebo controlled study of fluticasone propionate in patients with moderate to severe chronic obstructive pulmonary disease: the ISOLDE trial. BMJ **320:**1297–1303.
9. **Chang, C. C., J. R. Gilsdorf, V. J. DiRita, and C. F. Marrs.** 2000. Identification and genetic characterization of *Haemophilus influenzae* genetic island 1. Infect. Immun. **68:**2630–2637.
10. **Cowell, J. K., S. Matsui, Y. D. Wang, J. LaDuca, J. Conroy, D. McQuaid, and N. J. Nowak.** 2004. Application of bacterial artificial chromosome array-based comparative genomic hybridization and spectral karyotyping to the analysis of glioblastoma multiforme. Cancer Genet. Cytogenet. **151:**36–51.
11. **Davis, J., A. L. Smith, W. R. Hughes, and M. Golomb.** 2001. Evolution of an autotransporter: domain shuffling and lateral transfer from pathogenic *Haemophilus* to *Neisseria*. J. Bacteriol. **183:**4626–4635.
12. **De, K., T. Ramamurthy, S. M. Faruque, S. Yamasaki, Y. Takeda, G. B. Nair, and R. K. Nandy.** 2004. Molecular characterisation of rough strains of *Vibrio cholerae* isolated from diarrhoeal cases in India and their comparison to smooth strains. FEMS Microbiol. Lett. **232:**23–30.
13. **Dore, N., D. Bennett, M. Kaliszer, M. Cafferkey, and C. J. Smyth.** 2003. Molecular epidemiology of group B streptococci in Ireland: associations between serotype, invasive status and presence of genes encoding putative virulence factors. Epidemiol. Infect. **131:**823–833.
14. **Erwin, A. L., K. L. Nelson, T. Mhlanga-Mutangadura, P. J. Bonthuis, J. L. Geelhood, G. Morlin, W. C. Unrath, J. Campos, D. W. Crook, M. M. Farley, F. W. Henderson, R. F. Jacobs, K. Muhlemann, S. W. Satola, L. van Alphen, M. Golomb, and A. L. Smith.** 2005. Characterization of genetic and phenotypic diversity of invasive nontypeable *Haemophilus influenzae*. Infect. Immun. **73:**5853–5863.
15. **Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, et al.** 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269:**496–512.
16. **Gilsdorf, J. R., C. F. Marrs, and B. Foxman.** 2004. *Haemophilus influenzae*: genetic variability and natural selection to identify virulence factors. Infect. Immun. **72:**2457–2461.
17. **Gompertz, S., D. L. Bayley, S. L. Hill, and R. A. Stockley.** 2001. Relationship between airway inflammation and the frequency of exacerbations in patients with smoking related COPD. Thorax **56:**36–41.
18. **Harrison, A., D. W. Dyer, A. Gillaspy, W. C. Ray, R. Mungur, M. B. Carson, H. Zhong, J. Gipson, M. Gipson, L. S. Johnson, L. Lewis, L. O. Bakaletz, and R. S. Munson, Jr.** 2005. Genomic sequence of an otitis media isolate of nontypeable *Haemophilus influenzae*: comparative study with *H. influenzae* serotype d, strain KW20. J. Bacteriol. **187:**4627–4636.
19. **Johnson, J. R., and J. J. Brown.** 1996. A novel multiply primed polymerase chain reaction assay for identification of variant *papG* genes encoding the Gal(α1–4)Gal-binding PapG adhesins of *Escherichia coli*. J. Infect. Dis. **173:**920–926.
20. **Johnson, J. R., S. Jelacic, L. M. Schoening, C. Clabots, N. Shaikh, H. L. Mobley, and P. I. Tarr.** 2005. The IrgA homologue adhesin Iha is an *Escherichia coli* virulence factor in murine urinary tract infection. Infect. Immun. **73:**965–971.
21. **Johnson, J. R., M. A. Kuskowski, A. Gajewski, D. F. Sahm, and J. A. Karlowsky.** 2004. Virulence characteristics and phylogenetic background of multidrug-resistant and antimicrobial-susceptible clinical isolates of *Escherichia coli* from across the United States, 2000–2001. J. Infect. Dis. **190:**1739–1744.
22. **Johnson, J. R., M. A. Kuskowski, T. O'Bryan, R. Colodner, and R. Raz.** 2005. Virulence genotype and phylogenetic origin in relation to antibiotic resistance profile among *Escherichia coli* urine sample isolates from Israeli women with acute uncomplicated cystitis. Antimicrob. Agents Chemother. **49:**26–31.
23. **Johnson, J. R., A. C. Murray, A. Gajewski, M. Sullivan, P. Snippes, M. A. Kuskowski, and K. E. Smith.** 2003. Isolation and molecular characterization of nalidixic acid-resistant extraintestinal pathogenic *Escherichia coli* from retail chicken products. Antimicrob. Agents Chemother. **47:**2161–2168.
24. **Johnson, J. R., T. T. O'Bryan, P. Delavari, M. Kuskowski, A. Stapleton, U. Carlino, and T. A. Russo.** 2001. Clonal relationships and extended virulence genotypes among *Escherichia coli* isolates from women with a first or recurrent episode of cystitis. J. Infect. Dis. **183:**1508–1517.
25. **Johnson, J. R., T. A. Russo, P. I. Tarr, U. Carlino, S. S. Bilge, J. C. Vary, Jr., and A. L. Stell.** 2000. Molecular epidemiological and phylogenetic associations of two novel putative virulence genes, *iha* and *iroN* (*E. coli*), among *Escherichia coli* isolates from patients with urosepsis. Infect. Immun. **68:**3040–3047.
26. **Johnson, J. R., F. Scheutz, P. Ulleryd, M. A. Kuskowski, T. T. O'Bryan, and T. Sandberg.** 2005. Host-pathogen relationships among *Escherichia coli* isolates recovered from men with febrile urinary tract infection. Clin. Infect. Dis. **40:**813–822.
27. **Johnson, J. R., and A. L. Stell.** 2000. Extended virulence genotypes of *Escherichia coli* strains from patients with urosepsis in relation to phylogeny and host compromise. J. Infect. Dis. **181:**261–272.
28. **Johnson, J. R., A. L. Stell, F. Scheutz, T. T. O'Bryan, T. A. Russo, U. B. Carlino, C. Fasching, J. Kavle, L. Van Dijk, and W. Gaastra.** 2000. Analysis of the F antigen-specific *papA* alleles of extraintestinal pathogenic *Escherichia coli* using a novel multiplex PCR-based assay. Infect. Immun. **68:**1587–1599.
29. **Johnson, J. R., C. van der Schee, M. A. Kuskowski, W. Goessens, and A. van Belkum.** 2002. Phylogenetic background and virulence profiles of fluoroquinolone-resistant clinical *Escherichia coli* isolates from the Netherlands. J. Infect. Dis. **186:**1852–1856.
30. **Kanner, R. E., N. R. Anthonisen, and J. E. Connett.** 2001. Lower respiratory illnesses promote FEV(1) decline in current smokers but not ex-smokers with mild chronic obstructive pulmonary disease: results from the lung health study. Am. J. Respir. Crit. Care Med. **164:**358–364.
31. **Kroll, J. S., K. E. Wilks, J. L. Farrant, and P. R. Langford.** 1998. Natural genetic exchange between *Haemophilus* and *Neisseria*: intergeneric transfer of chromosomal genes between major human pathogens. Proc. Natl. Acad. Sci. USA **95:**12381–12385.
32. **Miyamoto, N., and L. O. Bakaletz.** 1996. Selective adherence of non-typeable *Haemophilus influenzae* (NTHi) to mucus or epithelial cells in the chinchilla eustachian tube and middle ear. Microb. Pathog. **21:**343–356.

33. **Munson, R. S., Jr., A. Harrison, A. Gillaspy, W. C. Ray, M. Carson, D. Armbruster, J. Gipson, M. Gipson, L. Johnson, L. Lewis, D. W. Dyer, and L. O. Bakaletz.** 2004. Partial analysis of the genomes of two nontypeable *Haemophilus influenzae* otitis media isolates. Infect. Immun. **72:**3002–3010.
34. **Murphy, T. F.** 2000. Bacterial otitis media: pathogenetic considerations. Pediatr. Infect. Dis. J. **19:**S9–S15.
35. **Murphy, T. F.** 2003. Respiratory infections caused by non-typeable *Haemophilus influenzae*. Curr. Opin. Infect. Dis. **16:**129–134.
36. **Murphy, T. F., A. L. Brauer, A. T. Schiffmacher, and S. Sethi.** 2004. Persistent colonization by *Haemophilus influenzae* in chronic obstructive pulmonary disease. Am. J. Respir. Crit. Care Med. **170:**266–272.
37. **Murphy, T. F., and S. Sethi.** 2002. Chronic obstructive pulmonary disease: role of bacteria and guide to antibacterial selection in the older patient. Drugs Aging **19:**761–775.
38. **Okeke, I. N., I. C. Scaletsky, E. H. Soars, L. R. Macfarlane, and A. G. Torres.** 2004. Molecular epidemiology of the iron utilization genes of enteroaggregative *Escherichia coli*. J. Clin. Microbiol. **42:**36–44.
39. **Pettigrew, M. M., B. Foxman, C. F. Marrs, and J. R. Gilsdorf.** 2002. Identification of the lipooligosaccharide biosynthesis gene *lic2B* as a putative virulence factor in strains of nontypeable *Haemophilus influenzae* that cause otitis media. Infect. Immun. **70:**3551–3556.
40. **Poulsen, K., J. Reinholdt, and M. Kilian.** 1992. A comparative genetic study of serologically distinct *Haemophilus influenzae* type 1 immunoglobulin A1 proteases. J. Bacteriol. **174:**2913–2921.
41. **Rao, V. K., G. P. Krasan, D. R. Hendrixson, S. Dawid, and J. W. St. Geme III.** 1999. Molecular determinants of the pathogenesis of disease due to non-typable *Haemophilus influenzae*. FEMS Microbiol. Rev. **23:**99–129.
42. **Russo, T. A., U. B. Carlino, and J. R. Johnson.** 2001. Identification of a new iron-regulated virulence gene, *ireA*, in an extraintestinal pathogenic isolate of *Escherichia coli*. Infect. Immun. **69:**6209–6216.
43. **Sethi, S., N. Evans, B. J. Grant, and T. F. Murphy.** 2002. New strains of bacteria and exacerbations of chronic obstructive pulmonary disease. N. Engl. J. Med. **347:**465–471.
44. **Sethi, S., and T. F. Murphy.** 2001. Bacterial infection in chronic obstructive pulmonary disease in 2000: a state-of-the-art review. Clin. Microbiol. Rev. **14:**336–363.
45. **Sethi, S., C. Wrona, B. J. Grant, and T. F. Murphy.** 2004. Strain-specific immune response to *Haemophilus influenzae* in chronic obstructive pulmonary disease. Am. J. Respir. Crit. Care Med. **169:**448–453.
46. **Shen, K., P. Antalis, J. Gladitz, S. Sayeed, A. Ahmed, S. Yu, J. Hayes, S. Johnson, B. Dice, R. Dopico, R. Keefe, B. Janto, W. Chong, J. Goodwin, R. M. Wadowsky, G. Erdos, J. C. Post, G. D. Ehrlich, and F. Z. Hu.** 2005. Identification, distribution, and expression of novel genes in 10 clinical isolates of nontypeable *Haemophilus influenzae*. Infect. Immun. **73:**3479–3491.
47. **Spencer, S., P. M. Calverley, P. S. Burge, and P. W. Jones.** 2004. Impact of preventing exacerbations on deterioration of health status in COPD. Eur. Respir. J. **23:**698–702.
48. **St. Geme, J. W., III.** 2002. Molecular and cellular determinants of nontypeable *Haemophilus influenzae* adherence and invasion. Cell Microbiol. **4:**191–200.
49. **St. Geme, J. W., III.** 2000. The pathogenesis of nontypable *Haemophilus influenzae* otitis media. Vaccine **19**(Suppl. 1)**:**S41–S50.
50. **St. Geme, J. W., III., V. V. Kumar, D. Cutter, and S. J. Barenkamp.** 1998. Prevalence and distribution of the *hmw* and *hia* genes and the HMW and Hia adhesins among genetically diverse strains of nontypeable *Haemophilus influenzae*. Infect. Immun. **66:**364–368.
51. **Taylor, D. C., A. W. Cripps, and R. L. Clancy.** 1995. A possible role for lysozyme in determining acute exacerbation in chronic bronchitis. Clin. Exp. Immunol. **102:**406–416.
52. **Tettelin, H., N. J. Saunders, J. Heidelberg, A. C. Jeffries, K. E. Nelson, J. A. Eisen, K. A. Ketchum, D. W. Hood, J. F. Peden, R. J. Dodson, W. C. Nelson, M. L. Gwinn, R. DeBoy, J. D. Peterson, E. K. Hickey, D. H. Haft, S. L. Salzberg, O. White, R. D. Fleischmann, B. A. Dougherty, T. Mason, A. Ciecko, D. S. Parksey, E. Blair, H. Cittone, E. B. Clark, M. D. Cotton, T. R. Utterback, H. Khouri, H. Qin, J. Vamathevan, J. Gill, V. Scarlato, V. Masignani, M. Pizza, G. Grandi, L. Sun, H. O. Smith, C. M. Fraser, E. R. Moxon, R. Rappuoli, and J. C. Venter.** 2000. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science **287:**1809–1815.
53. **Wilson, R.** 2001. Bacteria, antibiotics and COPD. Eur. Respir. J. **17:**995–1007.
54. **Yang, I. A., S. L. Seeney, J. M. Wolter, E. M. Anders, J. G. McCormack, A. M. Tunnicliffe, G. C. Rabnott, J. G. Shaw, A. G. Dent, S. T. Kim, P. V. Zimmerman, and K. M. Fong.** 2003. Mannose-binding lectin gene polymorphism predicts hospital admissions for COPD infections. Genes Immun. **4:**269–274.
55. **Yi, K., S. Sethi, and T. F. Murphy.** 1997. Human immune response to nontypeable *Haemophilus influenzae* in chronic bronchitis. J. Infect. Dis. **176:**1247–1252.

Editor: D. L. Burns