INFECTION AND IMMUNITY, Jan. 2006, p. 321–330 Vol. 74, No. 1
0019-9567/06/$08.00+0 doi:10.1128/IAI.74.1.321–330.2006

Characterization, Distribution, and Expression of Novel Genes among Eight Clinical Isolates of *Streptococcus pneumoniae*†

Kai Shen,[1] John Gladitz,[1] Patricia Antalis,[1] Bethany Dice,[1] Benjamin Janto,[1] Randy Keefe,[1] Jay Hayes,[1] Azad Ahmed,[1] Richard Dopico,[1] Nathan Ehrlich,[1]‡ Jennifer Jocz,[1]§ Laura Kropp,[1]¶ Shujun Yu,[1] Laura Nistico,[1] David P. Greenberg,[3] Karen Barbadora,[3] Robert A. Preston,[1] J. Christopher Post,[1,2] Garth D. Ehrlich,[1,2]* and Fen Z. Hu[1,2]

Center for Genomic Sciences, Allegheny-Singer Research Institute, Pittsburgh, Pennsylvania 15212[1]; Department of Microbiology and Immunology, Drexel University College of Medicine, Allegheny Campus, Pittsburgh, Pennsylvania 15212[2]; and Children's Hospital of Pittsburgh, Pittsburgh, Pennsylvania 15213[3]

Received 1 May 2005/Returned for modification 13 July 2005/Accepted 21 September 2005

Eight low-passage-number *Streptococcus pneumoniae* clinical isolates, each of a different serotype and a different multilocus sequence type, were obtained from pediatric participants in a pneumococcal vaccine trial. Comparative genomic analyses were performed with these strains and two *S. pneumoniae* reference strains. Individual genomic libraries were constructed for each of the eight clinical isolates, with an average insert size of ~1 kb. A total of 73,728 clones were picked for arraying, providing more than four times genomic coverage per strain. A subset of 4,793 clones were sequenced, for which homology searches revealed that 750 (15.6%) of the sequences were unique with respect to the TIGR4 reference genome and 263 (5.5%) clones were unrelated to any available streptococcal sequence. Hypothetical translations of the open reading frames identified within these novel sequences showed homologies to a variety of proteins, including bacterial virulence factors not previously identified in *S. pneumoniae*. The distribution and expression patterns of 58 of these novel sequences among the eight clinical isolates were analyzed by PCR- and reverse transcriptase PCR-based analyses, respectively. These unique sequences were nonuniformly distributed among the eight isolates, and transcription of these genes in planktonic cultures was detected in 81% (172/212) of their genic occurrences. All 58 novel sequences were transcribed in one or more of the clinical strains, suggesting that they all correspond to functional genes. Sixty-five percent (38/58) of these sequences were found in 50% or less of the clinical strains, indicating a significant degree of genomic plasticity among natural isolates.

Streptococcus pneumoniae is a gram-positive cocci that is etiologically associated with meningitis as well as numerous infections of the respiratory mucosa, including pneumonia, otitis media (OM), and sinusitis. On a worldwide basis, it is estimated that there are over 10,000 pneumococcus-related deaths per day. *S. pneumoniae* possesses an inducible system (38) for the uptake of DNA from its environment and has served as the model organism for the study of bacterial transformation for nearly 8 decades (4, 23, 28). These autocompetence and autotransformation mechanisms, together with its role as a major human pathogen, make *S. pneumoniae* an ideal model organism for studying the effects of in vivo bacterial strain evolution on pathogenicity and persistence during chronic infection.

Genomic diversity among multiple strains within a pathogenic bacterial species has been proposed to play a key role in virulence by the continual evolution of new strains via horizontal gene transfer (HGT) mechanisms during polyclonal infections (16, 17, 31, 32, 45). There are over 90 catalogued serotypes of *S. pneumoniae*, each of which is distinguished by a unique capsular gene cluster which exists as a cassette at a common site within the genome. Moreover, comparative genomic studies (11) of the sequenced *S. pneumoniae* reference strains, TIGR4 (49) and R6 (27), together with information from the pneumococcal diversity project, in which multiple strains have been sequenced at multiple sites (25), have established that within the pneumococci there exists substantial genetic heterogeneity (allelic differences) as well as genomic plasticity (genic differences). Claverys et al. (13) have theorized that substitutive recombination among DNAs from "other pneumococci" is the primary driver of pneumococcal capsular diversity through the insertion and creation of mosaic genes via an iterative homologous gene recombination process. These investigators also refer to a global pneumococcal genome that is larger than the genome of any single bacterium.

We have previously posited that the reassortment of genes among strains is a supravirulence factor (one that exists above the individual bacterial level, or at the population level) and is one of the major contributing factors to the morbidity and mortality with which pneumococci are associated (17, 45). Phylogenetic studies performed with serogroup 6 subtypes have demonstrated that HGT plays a dominant role in strain evolution and the development of invasiveness and that the major type 6 subtypes have evolved several times through recombination (34, 43).

* Corresponding author. Mailing address: Center for Genomic Sciences, Allegheny-Singer Research Institute, Allegheny General Hospital, 320 East North Ave., 11th Floor South Tower, Pittsburgh, PA 15212. Phone: (412) 359-4228. Fax: (412) 359-6995. E-mail: gehrlich@wpahs.org.

† Supplemental material for this article may be found at http://iai.asm.org/.

‡ Present address: University of Pittsburgh, Pittsburgh, PA 15213.
§ Present address: Carnegie-Mellon University, Pittsburgh, PA 15213.
¶ Present address: Pennsylvania State University, State College, PA 16802.

S. pneumoniae is one of the principle pathogens isolated from OM effusions, and the treatment of OM is estimated to exceed 5 billion dollars a year in the United States (30). Chronic OM with effusion is the most common cause of conductive hearing loss in children; its treatment with antibiotics, although common, is largely ineffective and is believed to be one of the major evolutionary drivers in the development of antibiotic-resistant *S. pneumoniae.* The resistance to antibiotics of chronic OM with effusion stems from the fact that the major OM pathogens, including *S. pneumoniae*, form biofilms on the middle-ear mucosa (15, 18, 22), which are nearly impossible to eradicate using conventional antimicrobial therapy. Bacterial biofilms have been increasingly recognized as playing an important role in mucosal diseases of the human aerodigestive tree (12, 14, 15, 16, 18, 37, 40), and biofilm bacteria are demonstrably more resistant to antibiotics and host defense mechanisms than are free-swimming bacteria (9, 14). These observations are important, as rates of HGT have been observed to be several orders of magnitude higher for biofilm bacteria than for their planktonic counterparts (36). This elevated level of gene transfer helps to explain the very rapid spread of antibiotic resistance genes among chronic pathogens and is likely the same evolutionary engine that drives the development of strains that can persist in the face of the host's adaptive immune response.

The current investigation is one in a series of studies designed to gauge the extent of genomic plasticity among isolates within a pathogenic species of bacteria. Recently, Bruckner et al. performed a whole-genome comparison between the two fully sequenced pneumococcal isolates (TIGR4 and R6) and demonstrated that approximately 10% of each strain's genes are unique (11). In order to estimate the genetic diversity among clinical pneumococcal strains, we selected eight isolates, each of a different serotype, recovered from pediatric participants in an *S. pneumoniae* vaccine trial and constructed individual genomic libraries from these strains. A survey of a subset of the clones was used to estimate the genomic differences among the clinical isolates and the two reference strains, TIGR4 (49) and R6 (27).

MATERIALS AND METHODS

Isolation, growth, and storage of bacterial strains. Pneumococcal isolates were obtained as nasal washes from symptomatic pediatric participants at Children's Hospital of Pittsburgh who were enrolled in a pneumococcal vaccine trial. All *S. pneumoniae* strains were isolated and restreaked for single colonies on Trypticase soy agar supplemented with 5% sheep's blood (Becton Dickinson, Sparks, MD) and then cultured in Todd-Hewitt broth (Sigma, St. Louis, MO) at 37°C in a humidified 5% CO_2 atmosphere for one passage, followed by division into aliquots and cryopreservation in 22% glycerol at −80°C. Pneumococcal isolates were typed by the Quellung reaction, using polyvalent serogroup- and serotype-specific antisera (Statens Seruminstitut, Copenhagen, Denmark). Isolates of serotypes 4, 6B, 9V, 14, 18C, 19F, and 23F were considered vaccine types. Isolates of *S. pneumoniae* were considered vaccine related if they were in the same serogroup as the vaccine types (e.g., 6A). The eight strains, designated BS68 to -75 in this study, were serotyped as types 9F, 14, 11, 3, 23F, 6A, 18C, and 19F, respectively. All eight strains were specifically chosen because of their sensitivities to both penicillins and cephalosporins, as we did not want to bias our analyses toward the identification of mobile genetic elements associated with antibiotic resistance.

Escherichia coli TOP10 cells were grown in Luria-Bertani broth (Becton Dickinson) at 37°C. Kanamycin (Invitrogen, Carlsbad, CA) was added to a final concentration of 50 μg/ml when necessary for selecting the transformed *E. coli* cells.

Extraction of bacterial genomic DNA. Genomic DNA from each pneumococcal strain was extracted using a modification of the method described by Ausubel et al. (3). Bacterial cells were grown overnight in 100 ml of Todd-Hewitt broth, centrifuged at 5,500 × *g*, and resuspended in 1× TE buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, pH 8.0). Sodium dodecyl sulfate (SDS) (Invitrogen) was added to a final concentration of 0.5% to lyse the cells, followed by an incubation at 37°C for 1 h with 50 μg/ml RNase A (Gentra Systems, Inc., Minneapolis, MN). Proteinase K (Invitrogen) was added to a concentration of 100 μg/ml, followed by an additional incubation at 37°C for 1 h. Cetyltrimethylammonium bromide (CTAB; Sigma) was added to a concentration of 1%, and the mixture was incubated at 65°C for 20 min, followed by chloroform-isoamyl alcohol (24:1) extraction of the DNA. The DNA was precipitated with 0.6 volume of isopropanol, centrifuged, and washed with 70% ethanol. The DNA was air dried, resuspended in 1× TE buffer at 65°C for 1 h, and analyzed by UV spectrophotometry and agarose gel electrophoresis.

Construction of genomic libraries. Individual genomic libraries for each of the eight clinical pneumococcal strains were constructed as previously described, with modifications (20). Briefly, genomic DNA from each clinical strain was isolated and hydrodynamically sheared to give a mean fragment size of 1.2 kb, using HydroShear (GeneMachines, San Carlos, CA). For each library, 3 μg of the sheared bacterial genomic DNA was end repaired using T4 and Klenow DNA polymerases (Invitrogen), ligated into the plasmid pCR4Blunt-TOPO, and transformed into *E. coli* TOP10 cells according to the instructions of the manufacturer (Invitrogen). A Q-bot multitasking robot (Genetix Limited, United Kingdom) was used to construct an addressable array for each pneumococcal strain of 9,216 transformants, providing more than four times coverage of the genome. The transformants were replica plated and stored in 10% glycerol at −80°C. A total of 73,728 clones were arrayed and replica plated from the eight libraries to provide an addressable storage system for future retrieval. Representative clones from each library were randomly chosen for initial analysis, with the expectation that the entire genomes will be sequenced in the future.

DNA sequencing. Plasmid DNA templates for sequencing were prepared from the cryopreserved pneumococcal genomic libraries by using a RevPrep Orbit robot (GeneMachines, San Carlos, CA) after overnight growth at 37°C in a robotic HiGro plate incubator-shaker (GeneMachines) according to the manufacturer's instructions. All plasmids were digested with EcoRI (Invitrogen) and analyzed in ethidium bromide-stained 1% agarose gels in TAE (0.04 M Tris-acetate, 0.001 M EDTA) buffer. Only constructs with inserts of >0.5 kb were used as sequencing templates. Sequencing reaction mixtures were prepared in a 3-μl volume consisting of the following reagents: 1.4 μl of plasmid template (approximately 100 ng DNA), 0.5 μl of 10-pmol/μl primer, and 1.1 μl of BigDye Terminator v3.1 cycle sequencing kit mix (Applied Biosystems Inc., Foster City, CA). Reaction aliquots of 0.5 μl were then subjected to thermal cycling and purified using the nano-pipetter of a Parallab 350 nanoliter genomic workstation (Brooks Automation, Inc., Chelmsford, MA). Parallab cycling conditions were as follows: 35 cycles with a 0-s denaturation step at 96°C, a 0-s annealing step at 50°C, and a 45-s extension step at 60°C. The purified samples were loaded into an ABI 3730 DNA analyzer (Applied Biosystems) for sequence analysis using ABI analysis software, v.5.1.1.

DNA sequence analysis. Sequence quality checking and vector trimming were performed using a Center for Genomic Sciences (CGS) custom-designed sequence analysis software package (J. Gladitz and S. Yu, unpublished data), which rapidly and automatically prepares finished sequences for automated homology searches. Sequences were also analyzed for the assembly of contigs using Sequencher (v. 4.1.4; Gene Codes Corporation, Ann Arbor, MI). DNA and hypothetical protein sequence homology searches were performed on the CGS high-speed BLAST cluster server (Erdos et al., unpublished data), using BLASTn and BLASTx (2), respectively. The BLAST server is automatically updated using downloads from the NCBI website (http://www.ncbi.nlm.nih.gov/). We compared our sequences with the TIGR4 and R6 genomic sequences and posted all of the non-TIGR4 and non-R6 DNA sequences which we used in the distribution study on the CGS website (www.centerforgenomicsciences.org/) as public documents.

PCR-based gene distribution studies. PCR primers for each novel clone were designed within the most likely open reading frame (ORF) of each clone or contig, using InforMax Vector NTI Advance software v.9.0 (Invitrogen). The primer sequences are available at the CGS website. Amplification of the *S. pneumoniae* cell division protein gene *pbp2x* was used as a positive control for all strains. An Eppendorf Master*Taq* kit (Brinkman Instruments, Inc., Westbury, NY) was used to amplify the target sequences. All amplifications were performed in a suite of Perkin-Elmer 9600 thermal cyclers (Applied Biosystems Inc.) in 25-μl reaction mixtures made up of 0.6 units of *Taq* DNA polymerase, 50 ng of template DNA, 10 pmol of each primer, 1.5 mM $MgCl_2$, and a 0.2 mM concentration of each deoxynucleoside triphosphate. PCR conditions were as follows:

10 min of denaturation at 95°C; 35 cycles of 30 s at 94°C, 1 min at 55°C, and 1 min at 72°C; a final extension of 7 min at 72°C; and a 4°C hold. The PCR products were loaded into 1.7% agarose gels, stained with ethidium bromide, and photographed with a Kodak Image Station 440 UV light scanner (Kodak Inc.).

RT-PCR detection of transcripts encoded by novel ORFs. RNA transcription was assessed to determine if each of the novel sequences corresponded to functional gene units. Using the hot-phenol method (Invitrogen), total RNAs for each of the eight *S. pneumoniae* clinical isolates and the reference strains were extracted from mid-exponential-phase cultures grown in Todd-Hewitt broth. After treatment with Turbo DNase (Ambion, Inc., Austin, TX), the RNA quality was checked on an Agilent 2100 bioanalyzer using an RNA 6000 nano assay kit (Agilent Technologies, Palo Alto, CA) to ensure that there was no degradation of RNA. RNAs were then tested in both reverse transcriptase PCR (RT-PCR)-based assays and sham (no RT) PCR-based assays, with the latter being performed to ensure that there was no DNA contamination. RT-PCR and PCR analyses were performed as follows. An RNA premix was prepared by mixing random hexamers (Invitrogen) with 4 μg of total RNA from each strain, heat denaturing the mixture, quenching it on ice, and dividing it into two tubes. Two master mixes were also prepared, with one containing all of the RT components, including Moloney murine leukemia virus (Invitrogen) reverse transcriptase (+RT), and the other one lacking the RT enzyme (−RT). Each pair of RNA specimens received an aliquot of the +RT mixture in one tube and the −RT mixture in the other. PCR was carried out using 2.5 μl of the first-strand cDNA from each RT reaction. The *S. pneumoniae* cell division protein *ftsI* mRNA was amplified as a positive control for each strain. Negative controls had the same reaction mixtures prepared but with no template nucleic acid added.

Southern blot analysis. Genomic DNA of each pneumococcal strain was isolated from planktonically grown cultures as described previously (45), digested with EcoRI, and electrophoresed in a 1% agarose gel. Each gel also contained one lane with a pool of the unique plasmid clones being probed to serve as a positive control. The DNAs were transferred to positively charged nylon membranes by capillary blotting using 0.4 M NaOH (Amersham Pharmacia Biotech, Buckinghamshire, United Kingdom) according to the method of Southern (48). Probes were produced by PCR-based amplification of the plasmid inserts corresponding to the unique genes under study, followed by purification of the amplimers with a QIAquick PCR purification kit (QIAGEN). Radioactive labeling of the probes was performed using a random primer DNA labeling system (Invitrogen) according to the manufacturer's instructions. Probes were purified via gel exclusion chromatography (G-50 Sephadex columns; Roche Diagnostics, Indianapolis, IN). Specific activity was measured in a Bioscan QC4000XER counter (Bioscan, Washington, DC). A 30-min prehybridization of the filters was carried out at 42°C (5× SSC [1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate], 5× Denhardt solution, 50% [wt/vol] formamide, 1% [wt/vol] SDS), with heat-denatured sheared salmon sperm DNA (Sigma) added immediately before incubation. Hybridization of the novel gene probes to the transferred DNAs was accomplished by adding $\sim 2 \times 10^7$ dpm of heat-denatured probe to each prehybridization reaction and incubating the mixture at 42°C overnight. Following hybridization, the membranes were washed in 2× SSC–0.1% (wt/vol) SDS for 5 min at room temperature three times, in 0.2× SSC–0.1% (wt/vol) SDS at room temperature for 15 min, in 0.2× SSC–0.1% (wt/vol) SDS at 42°C for 15 min, and finally in 0.1× SSC–0.1% SDS at 68°C for 15 min. Following the washing steps, the membranes were autoradiographed using Kodak XAR film.

Multilocus sequence typing and construction of phylogenetic trees. Using the method described by Enright and Spratt (19; www.mlst.net), we performed DNA sequencing of seven *S. pneumoniae* housekeeping genes, namely, *aroE*, *gdh*, *gki*, *recP*, *spi*, *xpt*, and *ddl*, from each of the eight clinical strains. PCRs using Eppendorf Master*Taq* kits were performed for each gene, using published primers, in a 96-well microtiter plate format. Thermal cycling was performed in a suite of Perkin-Elmer 9600 thermal cyclers as follows: an initial denaturation at 95°C for 4 min; 30 cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 30 s; and then 72°C for 10 min followed by a 4°C hold. The PCR products were purified using a QIAquick PCR purification kit (QIAGEN) and sequenced bidirectionally using an ABI 3730 DNA analyzer. The sequences were analyzed, and alignments of multiple sequences were done with Clustal X. The sequence of each locus in each strain was then compared with the multilocus sequence type (MLST) database to obtain an allelic profile. The combined allelic profiles of the seven gene loci for each strain resulted in an ST that was then compared with the existing ST profiles in the MLST database.

Nucleotide sequence accession numbers. The non-TIGR4 and non-R6 DNA sequences obtained for this study have been deposited in GenBank (accession no. CZ693485 and CZ693542).

TABLE 1. Analysis of sequences from eight individual *S. pneumoniae* libraries

Strain (serotype)	No. of sequences analyzed	No. of non-TIGR4 sequences[a]	% Non-TIGR4 sequences	No. of non-strep-related sequences	% Non-strep-related sequences
BS 68 (9)	880	189	21.5	71	8.1
BS 69 (14)	384	47	12.2	10	2.6
BS 70 (11)	545	71	13.0	22	4.0
BS 71 (3)	758	82	10.8	25	3.3
BS 72 (23)	318	61	19.2	29	9.1
BS 73 (6)	515	74	14.4	33	6.4
BS 74 (18)	883	156	17.7	51	5.8
BS 75 (19)	510	70	13.7	22	4.3
Total	4,793	750	15.6	263	5.5

[a] Non-TIGR4 sequences are not from TIGR4 but can be from other *Streptococcus* strains, phages, or other bacteria.

RESULTS AND DISCUSSION

Identification of novel sequences from clinical isolates. DNA sequencing was performed on 768 clones randomly chosen from each of eight genomic DNA libraries (6,144 total clones), each of which was composed of 9,216 clones that had been constructed from a single clinical isolate of *S. pneumoniae.* Approximately 78% (4,793/6,144) of the clones contained inserts of >0.5 kb and were subjected to further analyses, and this represented ~6.5% (4,793/73,728) of the total sequence in the aggregate libraries (Table 1). This method identified 4,043/4,793 (84.4%) clones as being TIGR4-like. Many of these TIGR4-like clones were demonstrated to contain small insertions-deletions and numerous point mutations, but in general they represented allelic variations of known genes.

BLASTn analyses indicated that 750/4,793 (15.6%) clones contained inserts that were unique with respect to the TIGR4 genome. The percentage of sequences not homologous to TIGR4 varied from 10.8 to 21.5% among the individual pneumococcal strains (Table 1). These novel sequences were also compared to the R6 pneumococcal genome and other streptococcal genomes as well as to streptococcal phages listed in GenBank. We identified 263/4,793 (5.5%) clones as having a unique insert of at least 105 nucleotides with respect to all streptococcus-related (referred to as "strep-related" from here on) sequences, with the majority of these clones having no nucleotide matches against any sequence in GenBank. All of the unique DNA sequences used in the distribution study (see below) from these pneumococcal clinical isolates have been deposited with GenBank (accession no. CZ693485 and CZ693542). These unique sequences were examined for overlaps among themselves, from which we assembled nine contigs comprised of 21 cloned sequences.

BLASTx searches were carried out to identify proteins with similarity to the conceptual protein translations of the ORFs corresponding to the non-strep-related clones and contigs. The amino acid identity and similarity values obtained by comparing these hypothetical translations with their closest homologs are listed in Table 2. Many of these clones had only very limited homology to any known protein, suggesting that some of these DNA sequences may encode proteins of novel function, whereas others likely encode virulence factors and proteins associated with metabolic functions and not previously

TABLE 2. Results of protein similarity searches of selected non-TIGR4 and non-R6 sequences

Clone name[a]	Similar protein	Organism	% Identity, % similarity	Nucleotide position of the ORF
SP11_0001_B02	PblB (platelet binding protein)	*Streptococcus mitis* phage SM1	72, 83	59–949
SP11_0001_N20_ORF2	Response regulator aspartate phosphatase	*Bacillus cereus* ATCC 14579	27, 54	1050–1331
SP11_0001_N20_ORF1*	Hypothetical protein	*Plasmodium falciparum* 3D7	22, 45	60–587
SP11_0001_O05	Immunoglobulin A1 protease	*Streptococcus sanguinis*,	53, 70	1–699
		S. pneumoniae R6 and TIGR4	51, 71 (TIGR4)	1–687 (TIGR4)
SP11_0002_A02	Transcriptional regulator	*Clostridium tetani*	40, 64	265–855
SP11_0002_B05_ORF1*	PTS system, IIC component	*Enterococcus faecalis* V583	48, 67	3–530
SP11_0002_B05_ORF2	Phosphonate monoester hydrolase	*Clostridium perfringens*	57, 75	878–1045
SP14_0001_D21	Predicted ATPase	*Streptococcus suis* 89/1591	47, 60	8–739
SP14_0001_F07	Phage-related protein	*Streptococcus suis* 89/1591	26, 43	8–418
SP14_0001_G03	Methyl-accepting chemotaxis protein	*Streptococcus suis*	71, 88	2–289
SP14_0001_J05_ORF1*	12-Oxophytodienoate-10,11 reductase	*Arabidopsis thaliana*	31, 44	54–344
SP14_0001_J05_ORF2	Hypothetical protein gbs0239	*Streptococcus agalactiae* Nem316	99, 99	555–1310
SP14_0001_N13_ORF1*	Hypothetical protein	*Leuconostoc mesenteroides*	34, 51	54–263
SP14_0001_N13_ORF2	Hypothetical protein SPY3_1258	*Streptococcus pyogenes* phage 315.4	38, 52	397–474
SP14_0001_N13_ORF3	Hypothetical protein	*Streptococcus pyogenes* SSI-1	41, 52	467–604
SP14_0001_O12	Hypothetical protein GLP_186_71509_70955	*Giardia lamblia* ATCC 50803	34, 51	41–214
SP14_0002_K23	Acetyl-coA carboxylase precursor	*Plasmodium yoelli* subsp. *yoelli*	22, 42	98–955
SP14_0002_O02	Zinc metalloprotease	*S. pneumoniae* R6	57, 72	4–453
SP18_0001_K21	Glycosyl hydrolase, propable alpha-fucosidase	*Clostridium perfringens* 13	71, 81	3–584
SP18_0001_N06_ORF2	Hypothetical protein Lgas02000227	*Lactobacillus gasseri*	33, 56	842–1261
SP18_0001_N06_ORF1*	ABC-type multidrug transport permease	*Rubrobacter xylanophilis* DSM 9941	43, 58	57–332
SP18_0002_L08/0003_F11_ORF1*	Subtilisin-like serine protease	*Leuconostoc mesenteroides* subsp. *mesenteroides* ATCC 8293	48, 70	234–1364
SP18_0002_L08/0003_F11_ORF2	ATPase of the AAA^+ class	*Leuconostoc mesenteroides* subsp. *mesenteroides* ATCC 8293	71, 86	11–119
SP18_0002_O09_ORF2	Hypothetical protein	*Clostridium perfringens* 13	47, 69	390–977
SP18_0002_O09_ORF1*	Ribulose-5-phosphate 4-epimerase	*Moorella thermoacetica* ATCC 39073	48, 68	37–345
SP18_0002_P11	Transcriptional regulator RggD	*S. pneumoniae* R6	41, 69	120–410
SP19_0001_C10_ORF2*	Hypothetical protein	*Streptococcus* bacteriophage MM1	89, 97	347–661
SP19_0001_C10_ORF1	Permease of the major facilitator superfamily	*Lactobacillus gasseri*	38, 61	25–132
SP19_0001_E12	gp49 homolog (replisome organizer)	*Neisseria gonorrhoeae*	36, 64	8–457
SP23_0001_O05	Putative membrane protein	*Treponema denticola* ATCC 35405	26, 51	51–449
SP23_0002_C12	Lantibiotic mersacidin-modifying enzyme	*Bacillus licheniformis*	25, 47	21–908
SP23_0002_F06	Lantibiotic mersacidin-modifying enzyme	*Bacillus licheniformis*	47, 58	37–237
SP23_0002_F23	Hypothetical protein Ssui801001491	*Streptococcus suis* 89/1591	39, 56	27–476
SP23_0002_G12_ORF2	ATP-binding ABC transporter (possible lantibiotic exporter)	*S. pneumoniae* R6	32, 52	538–1302
SP23_0002_G12_ORF1*	Lantibiotic mersacidin-modifying enzyme	*Bacillus licheniformis*	28, 51	29–517
SP23_0002_I15	Bacteriocin formation protein	*S. pneumoniae* TIGR4	23, 52	51–335
SP23_0002_K13_ORF1	Hypothetical protein	*Helicobacter pylori*	32, 60	315–473
SP23_0002_K13_ORF2	Helicase	*Corynebacterium glutamicum* ATCC 13032	37, 60	90–233
SP23_0002_K23_ORF1*	Hypothetical protein SAG1266	*Streptococcus agalactiae* 2603v/r	24, 50	6–371
SP23_0002_K23_ORF2	Hypothetical protein SMU.1685C	*Streptococcus mutans* UA159	26, 45	207–707
SP23_0002_O09	Arylsulfatase A	*Enterococcus faecium*	81, 91	2–814
SP23_0002_O13	Hypothetical protein TP02-0776	*Theileria parva*	27, 46	18–395

Continued on facing page

TABLE 2—*Continued*

Clone name[a]	Similar protein	Organism	% Identity, % similarity	Nucleotide position of the ORF
SP3_0001_D02	Zinc metalloprotease	*S. pneumoniae* TIGR4	53, 71	2–451
SP3_0001_E06	Alpha-galactosidase	*Geobacillus stearothermophilis*	56, 74	4–1056
SP3_0002_B15/SP9_0001_G17/ SP3_0001_103	Site-specific recombinase	*Streptococcus suis*	95, 96	723–1316
SP3_0001_O13	Zinc metalloprotease	*S. pneumoniae* TIGR4	43, 59	22–426
SP3_0002_B15_ORF1*	Peptidase E	*Exiguobacterium* sp. strain 255-15	28, 43	91–588
SP3_0002_B15_ORF2	Site-specific recombinase, DNA invertase Pin homologs	*Streptococcus suis* 89/1591	95, 97	723–1244
SP3_0002_H04 ORF1	Hypothetical protein lin2581	*Listeria innocua* Clip11262	43, 67	1–138
SP3_0002_H04 ORF2*	Prophage pi2 protein 29	*Lactococcus lactis* subsp. *lactis* I11403	40, 60	299–577
SP6_0001_A22_ORF2	HLA class II histocompatibility antigen	*Macaca mulatta*	27, 45	456–686
SP6_0001_A22_ORF1*	Hypothetical protein AN8322.2	*Aspergillus nidulans*	23, 45	249–560
SP6_0001_C18_ORF2	Nucleotidyltransferase	*Bacillus licheniformis*	42, 62	490–750
SP6_0001_C18_ORF1*	MSF transporter	*Schizosaccharomyces pombe*	28, 51	113–391
SP6_0001_D01	Hypothetical protein	*Streptococcus suis* bacteriophage phiko2	70, 84	30–287
SP6_0001_D08	gp21		31, 47	59–391
SP6_0001_K19	Leucyl-tRNA synthetase	*Streptococcus suis* 89/1591	77, 85	15–200
SP6_0001_P13	Phage-related protein	*Streptococcus suis* 89/1591	68, 85	113–607
SP6_0002_F09_ORF1	ORF continues for much longer	ORF continues for much longer		–852
SP6_0002_F09_ORF2	ABC-type antimicrobial peptide transport protein	*Enterococcus faecium*	48, 61	1–372
SP9_0001_D22_ORF1*	Hypothetical protein SAG1266	*Streptococcus agalactiae* 2603V/R	24, 50	4–369
SP9_0001_D22_ORF2	Hypothetical cytosolic protein	*Fusobacterium nucleatum* subsp. *nucleatum* ATCC 25586	27, 46	346–732
SP9_0001_G20/0001_H19/ 0001_O09*	Hypothetical protein plu0598	*Photorhabdus luminescens* subsp. *laumondii*	34, 57	25–690
SP9_0001_K08	Choline binding protein G	*S. pneumoniae* R6	58, 74	47–376
SP9_0001_P13_ORF2	Yeast Mcm2 protein	*Guillardia theta*	53, 73	158–235
SP9_0001_P13_ORF1*	Oligopeptide transport, ATP binding	*Pyrococcus horikoshii* OT3	39, 51	45–143
SP9_0002_C16_ORF2*	Phosphotransferase system (PTS) cellobiose-specific protein	*Enterococcus faecium*	63, 81	170–415
SP9_0002_C16_ORF1	PTS system II	*Enterococcus faecium*	58, 75	1–144
SP9_0002_D05_ORF2*	CiaB	*Wolinella succinogens* DSM 1740	45, 62	442–759
SP9_0002_D05_ORF1	Hypothetical protein	*Plasmodium falciparum* 3D7	34, 58	11–248
SP9_0002_I18	NTPase	*Cytophaga hutchinsonii*	27, 54	56–592
SP9_0002_I19	Tail protein	*Clostridium thermocellum* ATCC 27405	30, 47	99–1376
SP9_0002_N02/0001_B10/ 0002_N01	Choline binding protein J	*S. pneumoniae*	67, 80	52–582
SP9_0002_O20	Type I restriction/modification system, S subunit	*Treponema denticola* ATCC 35405	59, 73	1–267
SP9_0003_A11_ORF3	gp3	*Streptococcus mitis* phage SM1	83, 92	562–840
SP9_0003_A11_ORF2*	Inner membrane protein	*Streptococcus mutans* UA159	38, 55	320–475
SP9_0003_A11_ORF1	Membrane permease	*Rhodopseudomonas palustris*	28, 47	73–300
SP9_0003_F08_ORF2	Histidinol phosphate amino-transferase	*Picrophilis turridus* DSM 9290	39, 56	807–1052
SP9_0003_F08_ORF1*	Hypothetical protein	*Gibberella zeae* PH-1	27, 51	46–609
SP9_0003_H13_ORF2*	Related to chloramphenicol acetyltransferase	*Desulfotalea psychrophila*	28, 50	251–511
SP9_0003_H13_ORF1	BabR protein	*Babesia bovis*	45, 70	105–209
SP9_0003_N08	ABC-type multidrug transporter	*Moorella thermoacetum* ATCC 39073	35, 61	1–285
SP9_0003_O13_ORF2	Hypothetical protein plu0597	*Photorhabdus luminescens* subsp. *laumondii*	35, 56	316–1014
SP9_0003_O13_ORF1*	Hypothetical protein plu0598	*Photorhabdus luminescens* subsp. *laumondii*	38, 56	2–316

[a] *, ORF for which the primers were designed.

recognized in the pneumococci. Fifty-eight of these novel sequences were randomly chosen for the distribution study.

Distribution of novel DNA sequences. The distribution and expression patterns for 58 of the novel DNA sequences (GenBank accession no. CZ693485 and CZ693542) among the eight pneumococcal clinical strains were evaluated using PCR- and RT-PCR-based assays (Table 3). In all cases, genomic DNAs and RNAs from the reference strains, TIGR4 and R6,

TABLE 3. Distribution of non-TIGR4 and non-R6 DNA sequences and RNA transcripts from PCR and RT-PCR analyses

Clone/contig name	Amplimer size	PCR/RT-PCR: results for individual *S. pneumoniae* strains										No. of strains with novel sequences (DNA/RNA)
		BS 68 (Type 9)	*BS 69* (Type 14)	*BS 70* (Type 11)	*BS71* (Type 3)	*BS 72* (Type 23)	*BS 73* (Type 6)	*BS 74* (Type 18)	*BS 75* (Type 19)	TIGR 4	R6	
SP11_0001_B02	330	−/−	−/−	+/+	−/−	−/−	−/−	−/−	−/−	−/−	−/−	1/1
SP11_0001_N20	325	−/−	−/−	+/+	+/+	+/+	+/+	−/−	−/−	−/−	−/−	4/4
SP11_0001_O05	424	−/−	−/−	+/+	+/+	+/+	+/+	−/−	−/−	−/−	−/−	4/4
SP11_0002_A02	126	−/−	−/−	+/+	+/+	+/+	+/−	+/+	+/+	−/−	−/−	6/5
SP11_0002_B05	382	+/+	+/+	+/+	+/+	+/+	+/−	+/−	+/+	−/−	−/−	8/6
SP14_0001_D21	156	+/+	+/+	−/−	−/−	−/−	+/+	−/−	−/−	−/−	−/−	3/3
SP14_0001_F07	162	+/+	+/−	−/−	+/+	+/−	+/−	−/−	+/−	−/−	−/−	6/2
SP14_0001_G03	185	+/+	+/+	−/−	+/+	+/+	+/+	−/−	+/+	−/−	−/−	6/6
SP14_0001_J05	147	−/−	+/+	−/−	−/−	−/−	−/−	−/−	−/−	−/−	−/−	1/1
SP14_0001_N13	107	−/−	+/+	−/−	−/−	−/−	−/−	−/−	−/−	−/−	−/−	1/1
SP14_0001_O12	156	−/−	+/−	+/−	−/−	−/−	−/−	+/−	+/+	−/−	−/−	4/1
SP14_0002_K23	156	−/−	+/+	−/−	−/−	−/−	−/−	+/+	−/−	−/−	−/−	2/2
SP14_0002_O02	144	+/+	+/+	+/+	−/−	−/−	−/−	+/+	−/−	−/−	−/−	4/4
SP18_0001_K21	271	−/−	−/−	−/−	+/−	−/−	−/−	+/+	−/−	−/−	−/−	2/1
SP18_0001_N06	108	+/+	+/+	+/+	+/+	+/+	−/−	+/−	−/−	−/−	−/−	6/5
SP18_0002_L08/0003_F11	246	−/−	−/−	−/−	+/−	−/−	+/+	+/+	+/+	−/−	−/−	4/3
SP18_0002_O09	112	−/−	−/−	−/−	+/−	−/−	−/−	+/+	−/−	−/−	−/−	2/1
SP18_0002_P11	107	−/−	−/−	+/+	+/+	+/+	+/+	+/+	+/+	−/−	−/−	6/6
SP19_0001_C10	129	+/−	+/−	+/+	+/+	+/−	+/−	+/−	+/+	−/−	−/−	8/3
SP19_0001_E12	228	−/−	+/+	+/+	+/+	−/−	+/+	−/−	+/−	−/−	−/−	5/4
SP23_0001_O05	110	−/−	−/−	−/−	−/−	+/+	−/−	−/−	−/−	−/−	−/−	1/1
SP23_0002_C12	366	−/−	−/−	−/−	+/+	+/+	+/−	−/−	−/−	−/−	−/−	3/2
SP23_0002_F06	113	−/−	−/−	−/−	+/+	+/+	+/−	−/−	−/−	−/−	−/−	3/2
SP23_0002_F23	120	−/−	−/−	−/−	−/−	+/+	−/−	−/−	−/−	−/−	−/−	1/1
SP23_0002_G12	130	−/−	−/−	−/−	+/+	+/+	+/−	−/−	−/−	−/−	−/−	3/2
SP23_0002_I15	123	−/−	−/−	−/−	−/−	+/+	−/−	−/−	−/−	−/−	−/−	1/1
SP23_0002_K13	125	+/+	−/−	−/−	+/+	+/+	+/+	−/−	+/+	−/−	−/−	5/5
SP23_0002_K23	121	+/+	+/+	−/−	−/−	+/−	−/−	−/−	−/−	−/−	−/−	3/2
SP23_0002_O09	98	+/+	−/−	+/+	+/+	+/+	−/−	−/−	+/+	−/−	−/−	5/5
SP23_0002_O13	294	−/−	−/−	−/−	+/+	+/+	−/−	−/−	−/−	−/−	−/−	2/2
SP3_0001_D02	167	−/−	−/−	−/−	+/+	+/+	−/−	+/+	−/−	−/−	−/−	3/3
SP3_0001_E06	116	−/−	−/−	−/−	+/−	−/−	−/−	+/+	−/−	−/−	−/−	2/1
SP3_0001_O13	168	−/−	−/−	−/−	+/+	−/−	−/−	−/−	−/−	−/−	−/−	1/1
SP3_0002_B15/0001_103	132	+/+	+/+	−/−	+/−	−/−	−/−	+/+	+/+	−/−	−/−	5/4
SP3_0002_H04	177	−/−	−/−	+/−	+/+	−/−	+/+	−/−	−/−	−/−	−/−	3/2
SP6_0001_A22	113	−/−	−/−	−/−	−/−	−/−	+/+	−/−	−/−	−/−	−/−	1/1
SP6_0001_C18	109	−/−	−/−	−/−	−/−	−/−	+/+	−/−	−/−	−/−	−/−	1/1
SP6_0001_D01	150	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	−/−	−/−	8/8
SP6_0001_D08	127	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	−/−	−/−	8/8
SP6_0001_K19	113	+/+	+/−	−/−	+/+	+/+	+/+	−/−	+/+	−/−	−/−	6/5
SP6_0001_P13	112	+/+	+/+	−/−	+/+	+/+	+/+	−/−	+/+	−/−	−/−	6/6
SP6_0002_F09	118	+/+	−/−	+/−	−/−	−/−	+/−	−/−	−/−	−/−	−/−	3/1
SP9_0001_D22	115	+/+	−/−	−/−	−/−	−/−	−/−	−/−	−/−	−/−	−/−	1/1
SP9_0001_G17	229	+/+	+/+	+/+	+/+	−/−	−/−	−/−	+/−	−/−	−/−	5/4
SP9_0001_G20/0001_H19/0001_O09	103	+/+	−/−	−/−	−/−	−/−	−/−	−/−	−/−	−/−	−/−	1/1
SP9_0001_K08	143	+/+	−/−	−/−	−/−	−/−	−/−	−/−	−/−	−/−	−/−	1/1
SP9_0001_P13	95	+/+	−/−	−/−	−/−	+/+	+/+	−/−	+/+	−/−	−/−	4/4
SP9_0002_C16	110	+/+	−/−	+/+	+/+	+/+	−/−	−/−	+/+	−/−	−/−	5/5
SP9_0002_D05	110	+/+	−/−	−/−	−/−	−/−	−/−	−/−	−/−	−/−	−/−	1/1
SP9_0002_I18	121	+/+	+/+	−/−	+/+	−/−	−/−	−/−	−/−	−/−	−/−	3/3
SP9_0002_I19	112	+/−	+/−	−/−	+/−	+/−	+/+	+/+	+/+	−/−	−/−	7/3
SP9_0002_N02/0001_B10	139	+/+	+/+	+/+	+/+	+/+	−/−	+/+	−/−	−/−	−/−	6/6
SP9_0002_O20	118	+/+	−/−	+/+	+/+	+/+	+/+	+/+	+/+	−/−	−/−	7/7
SP9_0003_A11	108	+/+	+/+	−/−	−/−	−/−	−/−	−/−	−/−	−/−	−/−	2/2
SP9_0003_F08	153	+/+	+/+	−/−	−/−	−/−	−/−	−/−	−/−	−/−	−/−	2/2
SP9_0003_H13	137	+/+	+/+	−/−	+/−	−/−	−/−	−/−	+/+	−/−	−/−	4/3
SP9_0003_N08	80	+/−	+/−	+/−	−/−	−/−	−/−	+/+	−/−	−/−	−/−	4/1
SP9_0003_O13	107	+/+	−/−	+/−	−/−	−/−	−/−	−/−	−/−	−/−	−/−	2/1
No. of novel sequences		32	26	22	35	29	26	20	22	0	0	212
No. of novel transcripts		29	20	17	28	25	18	16	19	0	0	172
No. of non-TIGR4/R6 DNA sequences found exclusively in this strain		4	2	1	1	3	2	0	0	0	0	13
No. of non-TIGR4/R6 RNA transcripts found exclusively in this strain		6	2	1	1	3	2	4	1	0	0	20

were used as negative controls. The *S. pneumoniae* gene *pbp2x*, encoding a cell division protein, was selected as the positive control for all PCR- and RT-PCR-based assays. DNA and RNA preparations from all 10 *S. pneumoniae* strains supported PCR and RT-PCR, respectively, with the *pbp2x* gene primers (data not shown). All distribution and expression assays were performed in quadruplicate; a positive call required that at least three of four results were positive. Southern hybridization was used to validate the PCR-based distribution analyses for six clones and provided an overall concordance rate of 87% (52/60 identical calls). Each technique found four positive results that were missed by the other (data not shown).

TABLE 4. Frequency of *S. pneumoniae* strains showing unique sequences

No. of strains with unique sequences	Frequency of unique DNA sequences	% of unique DNA sequences	Frequency of unique RNA transcripts	% of unique RNA sequences
8	4	6.9	2	3.4
7	2	3.4	1	1.7
6	8	13.8	5	8.6
5	6	10.3	6	10.3
4	8	13.8	7	12.1
3	9	15.5	7	12.1
2	8	13.8	10	17.2
1	13	22.4	20	34.5
Total	58	100	58	100

TABLE 5. Pairwise tabulations of unique gene content and expression among eight pneumococcal clinical isolates

RNA sequence (serotype)	DNA sequence (serotype)[a]							
	BS 68 (9)	*BS 69* (14)	*BS 70* (11)	*BS 71* (3)	*BS 72* (23)	*BS 73* (6)	*BS 74* (18)	*BS 75* (19)
BS 68 (9)		**16**	**26**	**29**	**27**	**30**	**30**	**20**
BS 69 (14)	17		**26**	**29**	**31**	**30**	**22**	**20**
BS 70 (11)	26	21		**25**	**25**	**24**	**18**	**20**
BS 71 (3)	27	28	15		**16**	**19**	**23**	**17**
BS 72 (23)	28	31	18	11		**17**	**27**	**19**
BS 73 (6)	29	26	21	22	21		**28**	**16**
BS 74 (18)	33	24	19	30	27	22		**20**
BS 75 (19)	22	25	18	21	18	15	19	

[a] Numbers in bold refer to the number of unique genes between each pair of strains, and numbers that are not bold refer to the number of unique genes that are expressed differentially between members of a strain pair. In all cases, the number of differences is over a denominator of 58.

A global examination of the genic distribution of the 58 novel sequences among the eight clinical pneumococcal strains revealed that the novel genes, on average, were found in 45% (212/464) of the eight genomes. All of the novel sequences were amplified from one or more of the clinical isolates' genomes, indicating that none of these sequences represented contaminating fragments that entered our genomic libraries. All 58 primer pairs for the unique sequences failed, as expected, to support amplification from either the TIGR4 or R6 genome. Only four clones (7%) were identified in all eight of the clinical strains (e.g., SP11_0002_B05), while 13 of the unique ORFs were each found in only one of the clinical isolates (e.g., SP14_0001_J05) (Table 4). Strain BS71 (serotype 3) had the most novel sequences (35/58), whereas strain BS74 (serotype 18) had the least (20/58). Each of the eight strains had, on average, 26 of the 58 novel genes. The modal distribution of the unique clones among the eight strains demonstrated that the majority (38/58) of these sequences were found in 50% or fewer (four or fewer) of the strains.

One possible source of error with regard to our analysis of the distribution of the novel sequences is that different strains may possess different alleles of the same gene, thus preventing amplification with the primers designed from the sequenced clone. If this is an issue, it would mean that we have underestimated the distribution frequencies of some of the novel genes. However, we do not think any underestimation is significant, as Southern blot hybridization experiments using the entire cloned inserts as probes largely corroborated the PCR-based results. Nevertheless, it remains possible that genes encoding surface-exposed antigens or other proteins that come in contact with the host immune system could display heterogeneity beyond what could be detected by Southern blotting (SB). One possible example of this is clone SP11_0001_O05 (detected in 4 of the 10 strains by PCR and SB), which shows ~50% amino acid (aa) identity to the immunoglobulin A1 (IgA1) proteases of TIGR4 and R6 but no observable nucleotide homology to either of these genes. Whether this gene encodes an IgA1 protease will be the subject of further investigation.

Overall, the 58 novel genes were expressed as RNA transcripts in ~81% (172/212) of their genic occurrences within the eight clinical strains. Every ORF under study was expressed as an RNA transcript in at least one of the strains, indicating that all of the novel sequences under study correspond to transcriptionally active genes. A majority of the genes (44/58) were expressed in 50% or fewer of the clinical pneumococcal strains (four or fewer) under in vitro planktonic conditions. Thirty-five (60%) of the ORFs were transcriptionally active in all strains harboring the ORF. Strain BS71 (serotype 3), which contained the most novel DNA sequences, had an 80% RNA transcription rate (28/35) for these novel sequences. Similarly, strain BS74 (serotype 18), which had the fewest of the novel DNA sequences, also expressed 80% (16/20) of its novel ORFs (Table 3).

Pairwise genomic comparisons of the eight clinical strains were carried out using the 58 clone/contig sequences to estimate the level of genomic diversity among strains (Table 5). The greatest difference was found between strains BS68 (serotype 9) and BS74 (serotype 18), with 32 differences in DNA content. The most closely related strains were BS71 (serotype 3) and BS72 (serotype 23), which differed by 15 genes. Interestingly, in a blinded experiment using the chinchilla OM model to assess the relative pathogenicities of these eight clinical isolates, the last two strains produced similar infection profiles, whereas BS68 and BS74 produced the most divergent clinical outcomes in terms of mortality (Forbes et al., unpublished data). Overall, BS68 was found to have the greatest number of DNA differences ($n = 178$) compared to all other strains, while BS75 (type 19) had the fewest total genic differences.

Identification of novel virulence genes. (i) Subtilisin-like serine protease. Two contiguous ORFs from serotype 18 encode hypothetical proteins most similar to subtilisin-like serine proteases and ATPases of the AAA^+ class. The greatest homology for both proteins was found to *Leuconostoc mesenteroides* (48% aa identity with 70% aa similarity and 71% aa identity with 86% similarity, respectively). Somewhat lesser degrees of homology were observed to genes found in the *Mycoplasma* spp. and *Listeria* spp.

The subtilisin-like serine protease superfamily contains >200 subtilases that are widely distributed across all kingdoms and phyla (46). Microbial pathogens frequently use extracellular subtilases to degrade host proteins or to modify their own pathogenic proteins (33). Harris et al. (24) found in the group B streptococci that a mutation of a CspA subtilase resulted in a 10-fold reduction in virulence; the wild-type enzyme was able to cleave the human fibrinogen α chain, whereas the mutant could not. Similarly, in a mouse model of infection, a significant increase in survival was noted for animals infected with a $PrtA^-$ (subtilisin) *S. pneumoniae* strain compared with congenic wild-type bacteria (7). The

nucleotide sequence encoding the putative subtilase described herein and present in four of our clinical strains revealed no observable nucleotide or amino acid homology to either of the published streptococcal subtilisin genes.

(ii) Alpha-galactosidase clone. A clone from the serotype 3 strain contains an ORF whose hypothetical translation produces a protein similar to the α-galactosidases. The highest degree of relatedness was observed for a *Geobacillus stearothermophilus* enzyme (56% identity and 74% similarity), with many other gram-positive α-galactosidases showing similar levels of homology; a much lesser degree of relatedness was also observed with a TIGR4 α-galactosidase (Table 2). The finding of additional saccharide catabolic capability in pneumococci is in keeping with their documented ability to metabolize a larger number of sugars than other nasopharyngeal pathogens (26). This expanded substrate utilization capacity probably allows the pneumococci to occupy a unique physiological niche in an otherwise microbially complex environment. Moreover, α-galactosidase function and galactose transport have been shown to be essential for *S. pneumoniae* virulence (44, 47), as knockout mutants grow normally in vitro but are avirulent, likely owing to a reduction in their capsular polysaccharide of 50%. The finding of a second α-galactosidase in two of our more virulent clinical strains (Forbes et al., submitted for publication) suggests that genic redundancy in this case might be associated with increased virulence.

(iii) Immunoglobulin A1 protease. Our serotype 11 strain contained an ORF whose hypothetical translation displayed 53% identity to immunoglobulin A proteases from *Streptococcus sanguinis* and *S. pneumoniae* (Table 2); however, the greatest homology (56% identity, 73% similarity) was to an IgA protease from *Gemella hemolysans*, a member of the staphylococci. The streptococci, like many other pathogens, produce an IgA1 protease (39) that cleaves the J peptide of the divalent IgA, rendering it monomeric and greatly reducing its mucosal protective capacity (29, 42). All streptococcal IgA proteases are phylogenetically related and cleave a conserved Pro-Thr peptide bond in the human IgA heavy chain (41); however, there is significant variation and mosaicism among strains, likely as a mechanism to escape immune surveillance. It is probable that this newly identified gene provides genetic redundancy for a key virulence trait.

(iv) Lantibiotic mersacidin-like sequences. A three-clone contig from our serotype 23 strain contained two ORFs that appear to encode proteins related to mersacidin lantibiotic synthesis and transport. The first ORF is most similar to the *mrsM* gene of *Bacillus licheniformis*, which encodes a mersacidin-modifying enzyme (Table 2) (1). Three regions of similarity were identified between these two large proteins (1,026 amino acids), located at aa positions 500 to 565, 667 to 908, and 865 to 1022, with similarities of 52%, 56%, and 51%, respectively (8). Interestingly, the nucleotide sequences for the two clones which collectively comprise the first ORF did not match precisely in the overlapping region, indicating that these two clones originated from two different copies of a duplicated gene within the strain BS72 genome. A second ORF in this contig also showed similarity (48%) with the *B. licheniformis* lantibiotic transport gene *mrsT*, suggesting that these two genes were likely acquired together. These genes were also identified in two of our other clinical strains.

Lantibiotics are natural antibacterial peptides (10) encoded by genes from a wide array of gram-positive bacteria, and although present in a number of the streptococci, they have not previously been identified in any pneumococcal strains. It has recently been speculated that coordinated bacteriocin production and competence development may be two aspects of a conjoint mechanism for taking up DNA from neighboring species, as this would provide bacteria capable of inducing autocompetence with a supply of exogenous DNA by inducing lysis in nearby related species (31).

(v) Platelet binding protein. The serotype 11 strain library also yielded an ORF whose hypothetical translation produces a protein with similarity to the platelet binding protein PblB of *Streptococcus mitis*, which is encoded within the lysogenic SM1 prophage (6). Platelet binding by *S. mitis* was shown to be affected by this protein and a related PblA protein (5). These proteins appear to be multifunctional, as they also play a role in bacteriophage tail assembly. Since they are encoded by a lysogenic phage, their transmission among a broad range of bacterial species is likely. Indeed, both *Streptococcus pyogenes* and *Enterococcus faecalis* harbor PblA and/or PblB homologs. PblB represents the first reported case of a bacteriophage encoding a human tissue adhesin. However, virulence genes are commonly encoded by phages, as in the case of *Haemophilus influenzae* and *Vibrio cholerae*.

Multilocus sequence typing. MLST has been used as a DNA sequence-based approach to characterizing pathogenic isolates of numerous naturally transformable bacterial species, including *Neisseria meningitidis*, *Haemophilus influenzae*, and *S. pneumoniae* (19, 21, 35). MLST is useful for tracking epidemics, identifying prevalent clonal lineages, and studying the evolutionary relationships among strains of a given bacterial pathogen. Previously, we compared this method for evolutionarily grouping clinical isolates of *H. influenzae* with a novel gene distribution study (45) and found that each gene used in MLST produced a different evolutionary tree. In the current study, we again utilized MLST to characterize eight clinical isolates of *S. pneumoniae* and compared it with our novel gene distribution studies as a means to discriminate relatedness among the strains. The seven housekeeping genes used to determine the pneumococcal sequence types (ST) were sequenced from all eight clinical strains and analyzed as described by Enright and Spratt (19). All sequences were compared with the MLST database, and existing locus numbers were assigned to strains with sequence-matched loci. Each of the eight strains had a different ST. Only one strain, BS74 (serotype 18), received a new ST based upon a novel *ddl* allele (Table 6). It is noteworthy that both the ST and the serotype of each of the eight clinical strains matched both the ST and serotypes of previously catalogued strains listed in the MLST database, e.g., strain BS69 (serotype 14) had ST 124, and all 23 other strains in the database which shared ST 124 were also serotype 14; similarly, strain BS71 (serotype 3) shared sequence type 180 with 20 other serotype 3 strains.

The experiments reported herein were designed as a partial test of the distributed genome hypothesis, which states that (i) among chronic bacterial pathogens such as *S. pneumoniae*, no two clinical strains have the same genic content; (ii) chronic pathogens utilize a polyclonal infection strategy wherein the collective population-based supragenome is substantially larger than the ge-

TABLE 6. Characteristics of the eight clinical strains of *S. pneumoniae* used for this study[a]

Strain (serotype)	Serotype	Allelic profile for each gene							ST	No. of other strains in MLST database
		aroE	*gdh*	*gki*	*recP*	*spi*	*xpt*	*ddl*		
BS 68 (9)	9V	7	11	10	1	6	76	14	1269	1
BS 69 (14)	14	7	5	1	8	14	11	14	124	23
BS 70 (11)	11	2	5	29	12	16	3	14	62	3
BS 71 (3)	3	7	15	2	10	6	1	22	180	20
BS 72 (23)	23F	1	8	6	2	6	4	6	37	5
BS 73 (6)	6A	5	7	4	10	10	1	27	460	1
BS 74 (18)	18C	7	6	1	2	6	15	New	New	0
BS 75 (19)	19F	1	5	1	1	1	1	8	485	1

[a] All strains were isolated in 2003 from patients in Pennsylvania.

nome of any individual strain (13, 45); and (iii) recombination among the strains during persistent infections provides a mechanism to counteract the host's adaptive defense responses. Therefore, the use of reference genomes, such as TIGR4 and R6 in the case of *S. pneumoniae*, would identify only a subset of the genes extant within the population supragenome. To begin to address these hypotheses, we constructed individual genomic libraries from each of eight low-passage-number clinical isolates of *S. pneumoniae*, each of a different serotype and MLST, obtained from pediatric participants in a polyvalent pneumococcal vaccine trial. We surveyed approximately 6.5% of the clones in these libraries and identified ~15.6% of these clones as being novel with respect to the TIGR4 genome; moreover, 5.5% of the clones were novel with respect to all streptococcus-related sequences available in public databases.

The natural competence and transformation mechanisms of *S. pneumoniae* provide an obvious means for horizontal gene transfer, and not surprisingly, slightly more than 50% of the 58 novel genes we evaluated in detail were predicted to encode proteins that evidenced their highest degree of similarity to proteins encoded by other streptococcal species. However, the remainder of the unique sequences, based on protein homology searches, appear to have origins in more distantly related genera. Our genic distribution analyses demonstrated that these novel genes were nonuniformly distributed among the eight clinical strains and were not universally expressed under in vitro planktonic growth conditions. Our ability to amplify each of these novel sequences from one or more of the genomic DNAs isolated from the various clinical strains minimizes the likelihood that any of these genes represents a contaminating sequence from the laboratory. More than 65% (38/58) of the novel genes were present in 50% or fewer of the strains. All of the unique genes were found to be expressed in at least one strain, strongly suggesting that all of these unique sequences represented functional genes and not junk DNA. There were 13/58 gene sequences, however, identified exclusively in a single strain. Thus, these become strain-specific markers that can be used to analyze issues of gene flow during future polyclonal animal model experiments.

The degree of genomic plasticity and the pattern of genic distribution observed in the present study mirror the results obtained in a parallel study of the gram-negative nasopharyngeal pathogen *Haemophilus influenzae* (45). Thus, it appears that the presence of a population-based supragenome and the nonuniform distribution of virulence genes among the numerous strains of a species are common features of bacterial pathogens and cut across all phylogenetic classes. There are, however, differences in the degrees of sharing between the pneumococcal strains and the nontypeable *H. influenzae* strains. Among nontypeable *H. influenzae* strains, 32% of the novel genes were found in 90% or more of all clinical strains examined, whereas among the pneumococci, only 10% of the novel genes were found in 90% of the clinical strains examined. What is more interesting is that the overall degree of sharing of novel genes was less in the pneumococci. This greater degree of genomic plasticity among the pneumococcal strains we examined may result from the tentative observation that there is more limited gene sharing among strains of different serotypes. It will be important in future comparative genomic studies of the pneumococci to compare intraserotype diversity with interserotype diversity.

It is clear that each housekeeping gene has limited discriminating power compared to the concatenated unique sequences to identify the genetic clade to which a particular pneumococcal clinical strain belongs. This is analogous to the relationship between the individual novel clones and the sum of all novel clones, with the latter providing greater power for characterizing the relationships among different *S. pneumoniae* strains.

It has been demonstrated that bacteria are found predominantly in biofilms, the formation of which has alternate gene expression patterns and new gene generation capability. This may allow the bacteria to adapt more rapidly to changes in the environment. Our data revealed a wide variety of genic components in different clinical strains. The virulence difference for some strains disclosed by pathogenicity scores given to chinchilla OM animal models correlated well with our gene distribution patterns (e.g., BS71 and BS72 compared to BS68 and BS74) (Forbes et al., unpublished). There exists a high possibility that the novel nucleotide sequences in our study include some unrecognized virulence genes which are components of a set of contingency genes that are available to *S. pneumoniae* at a population level.

ACKNOWLEDGMENTS

We thank Mary O'Toole for her help with the preparation of the manuscript.

This work was supported by Allegheny-Singer Research Institute, Allegheny General Hospital, and the National Institute on Deafness and Other Communication Disorders grants DC 05659 (J.C.P.) and DC05659-02S1 (J.C.P.).

REFERENCES

1. **Altena, K., A. Guder, C. Cramer, and G. Bierbaum.** 2000. Biosynthesis of the lantibiotic mersacidin: organization of a type B lantibiotic gene cluster. Appl. Environ. Microbiol. **66:**2565–2571.
2. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
3. **Ausubel, F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl.** 1990. Current protocols in molecular biology. Greene Publishing Associates and Wiley-Interscience, New York, N.Y.
4. **Avery, O. T., C. M. MacLeod, and M. McCarty.** 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J. Exp. Med. **89:**137–158.
5. **Bensing, B. A., C. Rubens, and P. M. Sullam.** 2001. Genetic loci of *Streptococcus mitis* that mediate binding to human platelets. Infect. Immun. **69:** 1373–1380.
6. **Bensing, B. A., I. R. Siboo, and P. M. Sullam.** 2001. Proteins PblA and PblB of *Streptococcus mitis*, which promote binding to human platelets, are encoded within a lysogenic bacteriophage. Infect. Immun. **69:**6186–6192.
7. **Bethe, G., R. Nau, A. Wellmer, R. Hakenbeck, R. R. Reinert, H. P. Heinz, and G. Zysk.** 2001. The cell wall-associated serine protease PrtA: a highly con-

served virulence factor of Streptococcus pneumoniae. FEMS Microbiol. Lett. **205:**99–104.
8. **Bierbaum, G., H. Brotz, K. P. Koller, and H. G. Sahl.** 1995. Cloning, sequencing and production of the lantibiotic mersacidin. FEMS Microbiol. Lett. **127:**121–126.
9. **Borriello, G., E. Werner, F. Roe, A. M. Kim, G. D. Ehrlich, and P. S. Stewart.** 2004. Oxygen limitation contributes to antibiotic tolerance of *Pseudomonas aeruginosa* in biofilms. Antimicrob. Agents Chemother. **48:**2659–2664.
10. **Brotz, H., G. Bierbaum, A. Markus, E. Molitor, and H. G. Sahl.** 1995. Mode of action of the lantibiotic mersacidin: inhibition of peptidoglycan biosynthesis via a novel mechanism. Antimicrob. Agents Chemother. **39:**714–719.
11. **Bruckner, R., M. Nuhn, P. Reichmann, B. Weber, and R. Hakenbeck.** 2004. Mosaic genes and mosaic chromosomes—genomic variation in *Streptococcus pneumoniae*. Int. J. Med. Microbiol. **294:**157–168.
12. **Chole, R. A., and B. T. Faddis.** 2003. Anatomical evidence of microbial biofilms in tonsillar tissues: a possible mechanism to explain chronicity. Arch. Otolaryngol. Head Neck Surg. **129:**634–636.
13. **Claverys, J. P., M. Prudhomme, I. Mortier-Barriere, and B. Martin.** 2000. Adaptation to the environment: Streptococcus pneumoniae, a paradigm for recombination-mediated genetic plasticity? Mol. Microbiol. **35:**251–259.
14. **Costerton, W., R. Veeh, M. Shirtliff, M. Pasmore, C. Post, and G. Ehrlich.** 2003. The application of biofilm science to the study and control of chronic bacterial infections. J. Clin. Investig. **112:**1466–1477.
15. **Ehrlich, G. D., R. Veeh, X. Wang, J. W. Costerton, J. D. Hayes, F. Z. Hu, B. J. Daigle, M. D. Ehrlich, and J. C. Post.** 2002. Mucosal biofilm formation on middle-ear mucosa in the chinchilla model of otitis media. JAMA **287:**1710–1715.
16. **Ehrlich, G. D., F. Z. Hu, and J. C. Post.** 2004. Role for biofilms in infectious disease, p. 332–358. *In* M. Ghannoum and G. A. O'Toole (ed.), Microbial biofilms. ASM Press, Washington, D.C.
17. **Ehrlich, G. D., F. Z. Hu, K. Shen, P. Stoodley, and J. C. Post.** 2005. Bacterial plurality as a general mechanism driving persistence in chronic infections. Clin. Orthop. Relat. Res. **437:**20–24.
18. **Ehrlich, G. D., P. Stoodley, S. Kathju, Y. Zhao, B. R. McLeod, N. Balaban, F. Z. Hu, N. G. Sotereanos, J. W. Costerton, P. S. Stewart, J. C. Post, and Q. Lin.** 2005. Engineering approaches for the detection and control of orthopaedic biofilm infections. Clin. Orthop. Relat. Res. **437:**59–66.
19. **Enright, M. C., and B. G. Spratt.** 1998. A multilocus sequence typing scheme for Streptococcus pneumoniae: identification of clones associated with serious invasive disease. Microbiology **144:**3049–3060.
20. **Erdos, G., S. Sayeed, P. Antalis, F. Z. Hu, J. Hayes, J. Goodwin, R. Dopico, J. C. Post, and G. D. Ehrlich.** 2003. Development and characterization of a pooled Haemophilus influenzae genomic library for the evaluation of gene expression changes associated with mucosal biofilm formation in otitis media. Int. J. Pediatr. Otorhinolaryngol. **67:**749–755.
21. **Feil, E. J., M. C. Enright, and B. G. Spratt.** 2000. Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between Neisseria meningitidis and Streptococcus pneumoniae. Res. Microbiol. **151:**465–469.
22. **Fux, C. A., M. Shirtliff, P. Stoodley, and J. W. Costerton.** 2005. Can laboratory reference strains mirror "real-world" pathogenesis? Trends Microbiol. **13:**58–63.
23. **Griffith, F.** 1928. The significance of pneumococcal types. J. Hyg. **27:**113–159.
24. **Harris, T. O., D. W. Shelver, J. F. Bohnsack, and C. E. Rubens.** 2003. A novel streptococcal surface protease promotes virulence, resistance to opsonophagocytosis, and cleavage of human fibrinogen. J. Clin. Investig. **111:**61–70.
25. **Hollingshead, S. K., and D. E. Briles.** 2001. Streptococcus pneumoniae: new tools for an old pathogen. Curr. Opin. Microbiol. **4:**71–77.
26. **Holt, J., N. R. Krieg, P. H. A. Sneath, J. T. A. Staley, and S. T. Williams.** 1994. Bergey's manual of determinative bacteriology. Williams & Wilkins, Baltimore, Md.
27. **Hoskins, J., W. E. Alborn, Jr., J. Arnold, L. C. Blaszczak, S. Burgett, B. S. DeHoff, S. T. Estrem, L. Fritz, D. J. Fu, W. Fuller, C. Geringer, R. Gilmour, J. S. Glass, H. Khoja, A. R. Kraft, R. E. Lagace, D. J. LeBlanc, L. N. Lee, E. J. Lefkowitz, J. Lu, P. Matsushima, S. M. McAhren, M. McHenney, K. McLeaster, C. W. Mundy, T. I. Nicas, F. H. Norris, M. O'Gara, R. B. Peery, G. T. Robertson, P. Rockey, P. M. Sun, M. W. Winkler, Y. Yang, M. Young-Bellido, G. Zhao, C. A. Zook, R. H. Baltz, S. R. Jaskunas, P. R. Rosteck, Jr., P. L. Skatrud, and J. I. Glass.** 2001. Genome of the bacterium *Streptococcus pneumoniae* strain R6. J. Bacteriol. **183:**5709–5717.
28. **Hotchkiss, R. D.** 1951. Transfer of penicillin resistance in pneumococci by the desoxyribonucleate derived from resistant cultures. Cold Spring Harbor Symp. Quant. Biol. **16:**457–461.
29. **Kilian, M., J. Mestecky, and M. W. Russell.** 1988. Defense mechanisms involving Fc-dependent functions of immunoglobulin A and their subversion by bacterial immunoglobulin A proteases. Microbiol. Rev. **252:**296–303.
30. **Klein, J. O.** 2000. The burden of otitis media. Vaccine **19**(Suppl. 1)**:**S2–S8.
31. **Kreth, J., J. Merritt, W. Shi, and F. Qi.** 2005. Co-ordinated bacteriocin production and competence development: a possible mechanism for taking up DNA from neighbouring species. Mol. Microbiol. **57:**392–404.
32. **Lomholt, H.** 1995. Evidence of recombination and an antigenically diverse immunoglobulin A1 protease among strains of *Streptococcus pneumoniae*. Infect. Immun. **63:**4238–4243.
33. **Maeda, H.** 1996. Role of microbial proteases in pathogenesis. Microbiol. Immunol. **40:**685–699.
34. **Mavroidi, A., D. Godoy, D. M. Aanensen, D. A. Robinson, S. K. Hollingshead, and B. G. Spratt.** 2004. Evolutionary genetics of the capsular locus of serogroup 6 pneumococci. J. Bacteriol. **186:**8181–8192.
35. **Meats, E., E. J. Feil, S. Stringer, A. Cody, R. Goldstein, J. S. Kroll, T. Popovic, and B. G. Spratt.** 2003. Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by multilocus sequence typing. J. Clin. Microbiol. **41:**1623–1636.
36. **Molin, S., and T. Tolker-Nielsen.** 2003. Gene transfer occurs with enhanced efficiency in biofilms and induces enhanced stabilisation of the biofilm structure. Curr. Opin. Biotechnol. **14:**255–261.
37. **Palmer, R. J., Jr., S. M. Gordon, J. O. Cisar, and P. E. Kolenbrander.** 2003. Coaggregation-mediated interactions of streptococci and actinomyces detected in initial human dental plaque. J. Bacteriol. **185:**3400–3409.
38. **Peterson, S. N., C. K. Sung, R. Cline, B. V. Desai, E. C. Snesrud, P. Luo, J. Walling, H. Li, M. Mintz, G. Tsegaye, P. C. Burr, Y. Do, S. Ahn, J. Gilbert, R. D. Fleischmann, and D. A. Morrison.** 2004. Identification of competence pheromone responsive genes in Streptococcus pneumoniae by use of DNA microarrays. Mol. Microbiol. **51:**1051–1070.
39. **Poulsen, K., J. Reinholdt, C. Jespersgaard, K. Boye, T. A. Brown, M. Hauge, and M. Kilian.** 1998. A comprehensive genetic study of streptococcal immunoglobulin A1 proteases: evidence for recombination within and between species. Infect. Immun. **66:**181–190.
40. **Rayner, M. G., Y. Zhang, M. C. Gorry, Y. Chen, J. C. Post, and G. D. Ehrlich.** 1998. Evidence of bacterial metabolic activity in culture-negative otitis media with effusion. JAMA **279:**296–299.
41. **Reinholdt, J., M. Tomana, S. B. Mortensen, and M. Kilian.** 1990. Molecular aspects of immunoglobulin A1 degradation by oral streptococci. Infect. Immun. **58:**1186–1194.
42. **Reinholdt, J., and M. Kilian.** 1997. Comparative analysis of immunoglobulin A1 protease activity among bacteria representing different genera, species, and strains. Infect. Immun. **65:**4452–4459.
43. **Robinson, D. A., D. E. Briles, M. J. Crain, and S. K. Hollingshead.** 2002. Evolution and virulence of serogroup 6 pneumococci on a global scale. J. Bacteriol. **184:**6367–6375.
44. **Rosenow, C., M. Maniar, and J. Trias.** 1999. Regulation of the alpha-galactosidase activity in Streptococcus pneumoniae: characterization of the raffinose utilization system. Genome Res. **9:**1189–1197.
45. **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.
46. **Siezen, R. J., and J. A. Leunissen.** 1997. Subtilases: the superfamily of subtilisin-like serine proteases. Protein Sci. **6:**501–523.
47. **Smith, A. W., H. Roche, M. C. Trombe, D. E. Briles, and A. Hakansson.** 2002. Characterization of the dihydrolipoamide dehydrogenase from Streptococcus pneumoniae and its role in pneumococcal infection. Mol. Microbiol. **44:**431–448.
48. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98:**503–517.
49. **Tettelin, H., K. E. Nelson, I. T. Paulsen, J. A. Eisen, T. D. Read, S. Peterson, J. Heidelberg, R. T. DeBoy, D. H. Haft, R. J. Dodson, A. S. Durkin, M. Gwinn, J. F. Kolonay, W. C. Nelson, J. D. Peterson, L. A. Umayam, O. White, S. L. Salzberg, M. R. Lewis, D. Radune, E. Holtzapple, H. Khouri, A. M. Wolf, T. R. Utterback, C. L. Hansen, L. A. McDonald, T. V. Feldblyum, S. Angiuoli, T. Dickinson, E. K. Hickey, I. E. Holt, B. J. Loftus, F. Yang, H. O. Smith, J. C. Venter, B. A. Dougherty, D. A. Morrison, S. K. Hollingshead, and C. M. Fraser.** 2001. Complete genome sequence of a virulent isolate of Streptococcus pneumoniae. Science **293:**498–506.

Editor: J. N. Weiser