INFECTION AND IMMUNITY, Aug. 2006, p. 4766–4777 Vol. 74, No. 8
0019-9567/06/$08.00+0 doi:10.1128/IAI.00316-06

Identification of a Candidate *Streptococcus pneumoniae* Core Genome and Regions of Diversity Correlated with Invasive Pneumococcal Disease†

Caroline Obert,[1]‡ Jack Sublett,[1]‡ Deepak Kaushal,[2] Ernesto Hinojosa,[3] Theresa Barton,[4] Elaine I. Tuomanen,[1] and Carlos J. Orihuela[1]*

Department of Infectious Diseases, St. Jude Children's Research Hospital, 332 North Lauderdale, Memphis, Tennessee 38105[1]; Hartwell Center for Bioinformatics and Biotechnology, St. Jude Children's Research Hospital, 332 North Lauderdale, Memphis, Tennessee 38105[2]; Department of Microbiology and Immunology, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, Texas 78229[3]; and Department of Pediatrics, The University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, Texas 75390[4]

Received 24 February 2006/Returned for modification 13 April 2006/Accepted 10 May 2006

***Streptococcus pneumoniae* is a leading cause of community-acquired pneumonia and gram-positive sepsis. While multiple virulence determinants have been identified, the combination of features that determines the propensity of an isolate to cause invasive pneumococcal disease (IPD) remains unknown. In this study, we determined the genetic composition of 42 invasive and 30 noninvasive clinical isolates of serotypes 6A, 6B, and 14 by comparative genomic hybridization. Comparison of the present/absent gene matrix (i.e., comparative genomic analysis [CGA]) identified a candidate core genome consisting of 1,553 genes (73% of the TIGR4 genome), 154 genes whose presence correlated with the ability to cause IPD, and 176 genes whose presence correlated with the noninvasive phenotype. Genes identified by CGA were cross-referenced with the published signature-tagged mutagenesis studies, which served to identify core and IPD-correlated genes required for in vivo passage. Among these, two pathogenicity islands, region of diversity 8a (RD8a), which encodes a neuraminidase and V-type sodium synthase, and RD10, which encodes PsrP, a protein homologous to the platelet adhesin GspB in *Streptococcus gordonii*, were identified. Mice infected with a PsrP mutant were delayed in the development of bacteremia and demonstrated reduced mortality versus wild-type-infected controls. Finally, the presence of seven RDs was determined to correlate with the noninvasive phenotype, a finding that suggests some RDs may contribute to asymptomatic colonization. In conclusion, RDs are unequally distributed between invasive and noninvasive isolates, RD8a and RD10 are correlated with the propensity of an isolate to cause IPD, and PsrP is required for full virulence in mice.**

Streptococcus pneumoniae is the leading cause of community-acquired pneumonia, sepsis, and meningitis (3). Like other respiratory pathogens, the pneumococcus is primarily a commensal, colonizing the nasopharynx in 5 to 10% of healthy adults and 20 to 40% of healthy children (4, 40). In most instances, colonization is asymptomatic (17). Pneumococcal disease primarily occurs at the extremes of age, in young infants and the elderly; however, certain populations (e.g., Alaskan natives) and the immunocompromised are also highly susceptible to invasive pneumococcal disease (IPD) (18). IPD is marked by progression of the bacteria from the nasopharynx to sterile sites, such as the lungs, blood, and brain (2). Worldwide, it is estimated that *S. pneumoniae* is responsible for 15 cases of IPD per 100,000 persons per year and over a million deaths annually (2, 3).

For over a century, *S. pneumoniae* strains have been categorized by serology, with distinct serotypes identified on the basis of the 92 immunologically and chemically distinct polysaccharide capsules that surround and protect the bacterium from phagocytosis (26). Studies examining the contribution of the capsular type to virulence have demonstrated that only a small subset of serotypes cause the majority of IPD; serotypes 4, 6A, 6B, 14, 23F, 19F, 9V, and 18C account for 80% of the invasive isolates acquired from children 2 to 5 years of age in the United States (5, 13). More recently, molecular typing, such as pulsed-field gel electrophoresis of restriction fragments and multilocus sequence typing, has refined this observation, determining that within invasive serotypes, invasive and noninvasive clones exist (35, 42). Thus, the propensity of an isolate to cause invasive disease is dependent on its serotype and its genomic content.

Since release of the annotated genomes in 2001 (21, 27, 47), the challenge has been to define the genomic content responsible for IPD. Signature-tagged mutagenesis (STM) studies have identified genes required for in vivo passage of *S. pneumoniae* (24, 32, 39). Microarray analysis of in vivo RNA samples has revealed the pneumococcal transcriptome during bacteremia and meningitis (38). Nonetheless, greater clarity is needed as to why certain clonotypes are predisposed to invade while others colonize asymptomatically. As a first assumption,

* Corresponding author. Mailing address: Department of Microbiology and Immunology, Mail Code 7758, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX 78229-3900. Phone: (210) 567-3973. Fax: (210) 567-6612. E-mail: orihuela@uthscsa.edu.

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

‡ C.O. and J.S. contributed equally to the manuscript.

it is reasonable to suggest that invasive isolates carry and express genes that enable disease progression and evasion of the host defense. In contrast, commensal isolates are attenuated as a result of absence of these genes. Most recently, comparative genomic analyses by Hakenbeck et al. and Tettelin et al. have determined that individual isolates of *S. pneumoniae* vary by as much as 10% of their genomic content (23, 47). Moreover, 13 large loci, termed regions of diversity (RDs) by Tettelin et al. (45, 47), account for greater than half the genomic diversity observed between isolates. These RDs are particularly interesting, as they are often composed of atypical GC content, are often flanked by insertion sequences or remnants of mobile genetic elements, encode genes homologous to known virulence determinants, and encode unknown or hypothetical genes demonstrated by STM to be required for passage in mice (24, 32, 39).

In this report we describe the use of comparative genomics to comprehensively examine the genomic content of 72 invasive and noninvasive clinical isolates for features responsible for IPD. Serotypes 6A, 6B, and 14 were chosen for this analysis, as they can be obtained from both IPD and healthy carriers. We identified the candidate core genome and determined that RD and IPD-correlated genes are unequally distributed among serotypes and strains that cause human disease. We identify two RDs, RD8a and RD10, whose presence is highly correlated with the ability to cause human disease in a serotype-independent manner. Finally, we demonstrate that PsrP, a putative adhesin encoded within RD10, is required for efficient entry into the bloodstream of infected mice. Collectively, these studies suggest that RD8a and RD10 are pathogenicity islands, and their acquisition by *S. pneumoniae* increases their propensity to cause IPD.

MATERIALS AND METHODS

Bacterial strains. *S. pneumoniae* was grown on tryptic soy agar (Difco, Detroit, MI) plates supplemented with 3% defibrinated sheep blood or in defined semisynthetic casein liquid medium supplemented with 0.5% yeast extract (31). Clinical isolates were collected at The University of Texas Southwestern Medical Center in Dallas County, Tex., from February 1999 to January 2003. A total of 72 clinical isolates were examined: 23 serotype 6A isolates, 29 serotype 6B isolates, and 20 serotype 14 isolates. Invasive isolates were obtained from blood, cerebrospinal fluid, or aspirates of normally sterile sites from individuals with invasive disease (13 serotype 6A, 17 serotype 6B, and 12 serotype 14); noninvasive isolates were obtained from nasopharyngeal swabs of healthy carriers. Table S1 in the supplemental material lists the clinical isolates used in this study.

Microarray genome content analysis. Microarray experiments were performed by using whole-genome *S. pneumoniae* cDNA microarrays obtained from the Pathogen Functional Genomic Resource Center at The Institute for Genomic Research (TIGR) (http://pfgrc.tigr.org). These arrays have been described in detail previously (38) and consist of PCR products representing segments of 2,131 unique open reading frames (ORFs) from TIGR4, 164 ORFs from strain R6, and 399 ORFs from G54 (21, 27, 47). Microarray experiments, including DNA quality control, Cy3 and Cy5 dye labeling, hybridization, washing, scanning, and data analysis, were performed at the Functional Genomics lab, Hartwell Center for Bioinformatics and Biotechnology, St. Jude Children's Research Hospital. Cy dye labeling was performed using the BioPrime DNA labeling system (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. Microarray hybridizations and washing were performed by using protocols from the Pathogen Functional Genomic Resource Center (http://pfgrc.tigr.org/protocols.shtml). The hybridization probe was constituted by mixture of differentially labeled cDNA derived from (i) sonicated genomic DNA isolated from *S. pneumoniae* strain TIGR4 labeled with Cy3 and (ii) sonicated genomic DNA from the clinical isolates labeled with Cy5. In preliminary investigations, dye bias was not seen to have any impact on results (data not shown), and as a result no dye flips were performed. Microarray slides were scanned using an Axon 4000B dual channel scanner to generate a multi-TIFF image of each slide (Axon Corp., Union City, CA). Images were analyzed by using Axon GenePix 4.1 image analysis software, and the resulting text data files were imported into Spotfire DecisionSite for Functional Genomics (version 8.0; Spotfire, Somerville, MA) (28). Additional analysis was performed in the R language-based Bioconductor (www.bioconductor.org) release 1.9 using the "Array" packages.

A series of filtration algorithms were applied to remove spots that consistently generated bad data (based on the frequency with which a particular spot failed to reach a minimum required signal-to-noise ratio and the frequency with which a particular spot was flagged bad by the image analysis software, GenePix Pro 4.1). Genes that were flagged or that failed to meet the signal-to-noise ratio criterion 75% of the time were not considered. Lowess global normalization on background-corrected, log-transformed signal values was then performed to remove intensity-specific bias (10). Logarithmic Cy5/Cy3 ratios (log fold changes) were then calculated for every spot. Since each gene was spotted four times per glass microarray, data emerging from each of the valid spots were averaged for a particular gene.

Determinations on the presence or absence of genes were made by comparing hybridization signal strength between the clinical isolate and TIGR4. Values derived from the TIGR4 genomic DNA hybridization were used as baseline. Genes that exhibited normalized log ratios of less than −1.5 were designated as absent, whereas genes with log ratios greater than or equal to −1.5 were considered present. Briefly, the −1.5 cutoff was determined empirically. Primers were used to amplify *hdl*, *pepS*, *ext*, *dinP*, *valS*, *pyrDa*, *tdk*, *degV*, *abcT*, *str*, *spoJ*, and *rok* from the 20 serotype 14 clinical isolates. These genes were selected on the basis that no paralogs were known to be present, genes were >50 kb apart, and no one physiological process was to be oversampled. The ability to amplify these genes (indicating their presence) was subsequently correlated to their normalized log ratios. Strains containing genes with a log ratio of less than −1.5 failed to amplify their corresponding PCR product, whereas those with ratios greater than or equal to −1.5 succeeded. Table S2 in the supplemental material describes the loci used for this analysis, lists the primer pairs used, and summarizes the results.

Comparative genomic analysis (CGA). Phylogenetic relationships among the clinical isolates were extrapolated from the present/absent matrix determined by comparative genomic hybridization. This matrix was used to construct a topology based on the neighbor-joining algorithm and served to sort the clinical isolates into clades (i.e., groups of strains having similar genomic content). The tree was constructed using PAUP version 4.0b10 (44). Statistical support of the branch points for the tree was estimated by performing 5,000 bootstrap replicates in PAUP. Genes whose presence was correlated with the invasive phenotype (i.e., present in strains belonging to invasive clades and absent in strains belonging to noninvasive clades) and the noninvasive phenotype (i.e., absent in strains belonging to invasive clades and present in strains belonging to noninvasive clades) were determined for each serotype by sorting the present/absent matrix using a one-tailed Fisher's exact test ($P < 0.05$) (36). Serotype-independent IPD-correlated genes were also identified; all strains within invasive clades were sorted against all strains in noninvasive clades. Isolates belonging to semi-invasive clades were not included in the analysis.

Cross-reference with STM studies. In order to identify core genes and IPD-correlated genes of interest, we cross-referenced our findings with the three published *S. pneumoniae* STM studies (24, 32, 39). To do so, it was necessary to identity each STM gene in the context of its TIGR annotation (www.tigr.org). Findings by Lau et al. and Polissi et al. (but not Hava et al.) required BLASTN analysis of the raw sequence data obtained from the original STM. Raw sequence data were kindly provided by Alessandra Polissi, Sauli Haataja, and Jeremy Brown.

Insertion duplication mutagenesis. A PsrP-deficient mutant was constructed by insertion duplication mutagenesis. PCR was used to amplify a 451-bp fragment corresponding to bp 61 to 511 of *psrP* (SP1772). EcoRI and BamHI sites were integrated at the 5′ ends of the primers and were used to clone the DNA fragment into pJDC9, a suicide vector (15). The primers used were 5′-NNNNNGAATTCGGGATAGTTGCTGCGGGAGC and 5′-NNNNNGGATCCCCACTGAACGCTTGCGTCGC. PCR fragments and pJDC9 were digested with EcoRI and BamHI, ligated together, and transformed into *Escherichia coli*. Single transformants containing the insert were confirmed by sequencing, and plasmid DNA from these clones was used to transform TIGR4 (11). Chromosomal integration of the vector at the right locus was verified by PCR using primers homologous to plasmid sequences (M13 forward –21 and reverse primers) and to sequences upstream of the point of insertion of the plasmid.

Intranasal challenge model. Female BALB/cJ mice (Jackson Laboratory, Bar Harbor, ME) 4 to 5 weeks old were maintained in biosafety level 2 facilities at The University of Texas Health Science Center in San Antonio. All experiments

were done with mice under general anesthesia with inhaled isoflurane (2.5%; Baxter Healthcare Corp., Deerfield, IL). *S. pneumonia*, at either 10^4 or 10^7 CFU in 20 μl phosphate-buffered saline, was introduced by intranasal administration. Following challenge, bacterial titers in the blood were determined by tail snip and collection of 2 μl of blood, serial dilution, and plating. Bacterial titers in the nasopharynx were determined by nasopharyngeal lavage with 20 μl phosphate-buffered saline, serial dilution, and plating. Statistical analysis of bacterial titers was performed using a nonparametric independent group analysis (Mann-Whitney rank sum). Statistical analysis of survival over time was performed using a Fischer's exact test at day 7.

RESULTS AND DISCUSSION

Core genome. Hybridization of labeled genomic DNA to the microarrays identified 1,553 genes (73% of TIGR4 genes present on the microarray) present in >98% of all the clinical isolates and in TIGR4 (72 of 73 isolates). These findings are consistent with the 80% core genome recently described for *Streptococcus agalactiae* and the earlier *S. pneumoniae* comparative genome hybridization (CGH) studies performed by Hakenbeck et al. and Tettelin et al. (23, 46, 47). Genes that comprise the candidate core genome presumably meet the minimal functions required by the bacterium for colonization of the human nasopharynx. Major virulence factors in the core genome included the following: pneumolysin (SP1923), the hemolytic cytotoxin; autolysin (SP1937), the major murein hydrolase; SpxB (SP0730), the pyruvate oxidase responsible for hydrogen peroxide production; and HtrA (SP2239), a heat shock serine protease. Multiple studies clearly demonstrate a requirement for these genes in nasopharyngeal models of colonization and in animal models of invasive disease (27–29). Other virulence determinants present in the core genome included the following: hyaluronidase (SP0314); LytB, a second murein hydrolase (SP0965); an adhesion lipoprotein (SP1002); enolase (SP1128); various hemolysins (SP1204 and SP1466); and NADH oxidase (SP1469). Essential transporters included two iron transporters (SP0241-2 and SP1869-72) (30) and the *psa* operon, the genes encoding the manganese permease complex (SP1648-50) (31). Table S3 in the supplemental material indicates genes spotted on the microarrays that have been determined to be part of the core genome.

CGH analyses of other bacteria have also shown the presence of virulence determinants in the core genome (1, 33); nonetheless, assuming that the presence of a gene confers gain of function, the presence of the major virulence determinants (i.e., pneumolysin, pyruvate oxidase, and autolysin) in the noninvasive isolates suggests that these genes are necessary but not sufficient to determine the propensity of an isolate to cause disease. Presumably, only genes present in the invasive cohort and absent in the noninvasive cohort confer this property (see "Considerations," below). Nonetheless, since core genes were detected in all the clinical isolates and their DNA sequence is conserved (i.e., detectable by DNA hybridization of nucleotide sequence), it is likely that their gene products play critical roles for the bacterium. As such, core genes represent ideal targets for pharmacological intervention and/or vaccine development.

Cross-reference of the core genome with STM studies served to identify possible virulence determinants of unknown function. One such locus was SP2141-SP2146 (see Table S3 in the supplemental material). Multiple STM hits within this locus by all three STM studies suggest that the region is important and demonstrate that the gene products are required in

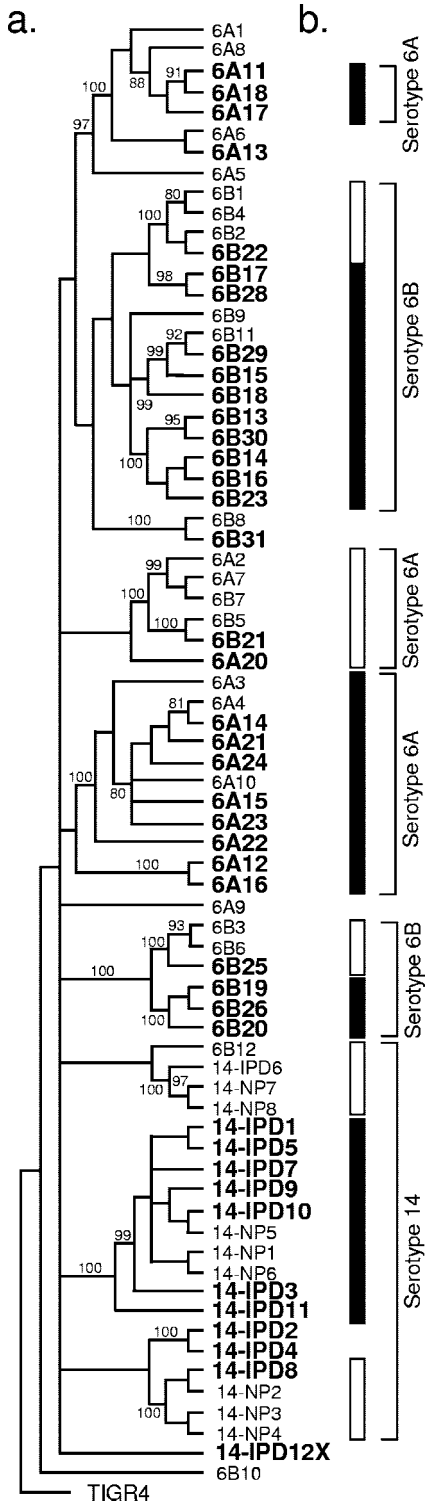

FIG. 1. Phylogenetic clustering of clinical isolates into invasive and noninvasive clades. a) A comparison of the present/absent gene matrix of clinical isolates and TIGR4 was used to construct a phylogenetic tree based on the neighbor-joining algorithm, with invasive isolates shown in boldface type and noninvasive isolates in lightface type. b) Within serotypes, clinical isolates clustered into clades based on their invasive potential. Closed bars, invasive clades; open bars, noninvasive clades.

TABLE 1. Regions of diversity correlated with the invasive or noninvasive phenotype

RD and TIGR ID	STM[b]	CGH correlation[a]		TIGR annotation or gene name
		IPD	Noninvasive	
RD2				
SP0163			6A	Transcriptional regulator PlcR, putative
SP0164			U, 6A	Hypothetical protein
SP0165			U, 6A	Flavoprotein
SP0166			U, 6A	Pyridoxal-dependent decarboxylase, Orn/Lys/Arg family
SP0167			U, 6A	Hypothetical protein
SP0168			U, 6A	Macrolide efflux protein, putative
SP0169			U, 6A	Lactose phosphotransferase system repressor, degenerate
SP0171			U, 6A	ROK family protein
RD5				
SP0691		14	6A	Hypothetical protein
SP0692		14	6A	Hypothetical protein
SP0694		14	6A	Conserved domain protein
SP0695		14	6A	HesA/MoeB/ThiF family protein
SP0696		14	6A	Hypothetical protein
SP0697		14	6A	ABC transporter, ATP-binding protein
SP0698		14	6A	Hypothetical protein
SP0700		14	6A	Transposase, IS*30* family, degenerate
RD6				
SP1046			U, 6A	α-Amylase family protein, authentic point mutation
SP1047			6A	Hypothetical protein
SP1048				Hypothetical protein
SP1049				Hypothetical protein
SP1050				Transcriptional regulator, putative
SP1051				Conserved hypothetical protein
SP1052			6A	Phosphoesterase, putative
SP1053			6A	Conserved domain protein
SP1054				Tn*5252*, Orf 10 protein
SP1055				Tn*5252*, Orf 9 protein
SP1056				Tn*5252*, relaxase
SP1057				Transcriptional regulator PlcR, putative
SP1058				Hypothetical protein
SP1059				Hypothetical protein
SP1060			U	Hypothetical protein
SP1061			U, 14	Protein kinase, putative
SP1062			U, 14	ABC transporter, ATP-binding protein
SP1063				ABC-2 transporter, permease protein, putative
SP1064			U, 6A, 14	Transposase, IS*200* family
SP1065		6A	14	Hypothetical protein
RD7				
SP1129			U, 6B	Integrase/recombinase, phage integrase family
SP1130				Transcriptional regulator
SP1131				Transcriptional regulator, putative
SP1132			U, 6B	Hypothetical protein
SP1133			U, 6B	Hypothetical protein
SP1134			U, 6B	Hypothetical protein
SP1135			6B	Hypothetical protein
SP1136			U, 6B	Conserved domain protein
SP1137			U	GTP-binding protein, putative
SP1138				Hypothetical protein
SP1139				Hypothetical protein
SP1140				Hypothetical protein
SP1141				Hypothetical protein
SP1142				Hypothetical protein
SP1143	H			Conserved hypothetical protein
SP1144				Conserved hypothetical protein
SP1145				Hypothetical protein
SP1146				Hypothetical protein
SP1147			U, 14	Integrase/recombinase, phage integrase family, truncation
RD8				
SP1315		U,6B,14		V-type sodium ATP synthase, subunit D
SP1316		U,6B,14		V-type sodium ATP synthase, subunit B

Continued on following page

TABLE 1—*Continued*

RD and TIGR ID	STM[b]	CGH correlation[a]		TIGR annotation or gene name
		IPD	Noninvasive	
SP1317		U, 6B, 14		V-type sodium ATP synthase, subunit A
SP1318		U, 6B		V-type sodium ATP synthase, subunit G
SP1319		U, 6B, 14		V-type sodium ATP synthase, subunit C
SP1320		U, 6B, 14		V-type sodium ATP synthase, subunit E
SP1321	H	U, 6B, 14		V-type sodium ATP synthase, subunit K
SP1322		U, 6B		V-type sodium ATP synthase, subunit I
SP1323		6B		Hypothetical protein
SP1324		U, 6B, 14		ROK family protein
SP1325		U, 6B		Oxidoreductase, Gfo/Idh/MocA family
SP1326		U, 6B, 14		Neuraminidase, putative
SP1327		U, 6B, 14		Conserved hypothetical protein
SP1328	H	U, 6B, 14		Sodium:solute symporter family protein
SP1329		U, 6B, 14		*N*-Acetylneuraminate lyase
SP1330		U, 6B, 14		*N*-Acetylmannosamine-6-P epimerase, putative
SP1331		U, 6B, 14		Phosphosugar-binding transcriptional regulator, putative
SP1332				Conserved domain protein
SP1333				Hypothetical protein
SP1334			14	Conserved hypothetical protein
SP1335			14	Hypothetical protein
SP1336				Type II DNA modification methyltransferase Spn5252IP
SP1337			U, 14	IS*1380*-Spn1, transposase
SP1338			U, 6B, 14	Hypothetical protein
SP1340			U, 6B, 14	Hypothetical protein
SP1341			U, 6B, 14	ABC transporter, ATP-binding protein
SP1342			U, 6B, 14	Toxin secretion ABC transporter, ATP-binding/permease protein
SP1343	H	6A	U, 6B, 14	Prolyl oligopeptidase family protein
SP1344	H		U, 6B, 14	Conserved hypothetical protein
SP1345			U, 6B, 14	Hypothetical protein
SP1346				Conserved hypothetical protein
SP1347			U, 14	Hypothetical protein
SP1348			U, 14	Conserved hypothetical protein
SP1349			U, 14	Hypothetical protein
SP1350			U, 14	Conserved domain protein
SP1351			U, 14	Hypothetical protein
RD9				
SP1612		U, 14		Conserved domain protein
SP1613		U, 14		IS*3*-Spn1, transposase, authentic point mutation
SP1614				IS*3*-Spn1, hypothetical protein, degenerate
SP1615			6A	Transketolase, authentic frameshift
SP1616			6A	Ribulose-phosphate 3-epimerase family protein
SP1617			6A	PTS system, IIC component
SP1618			6A	PTS system, IIB component
SP1619			6A	PTS system, IIA component
SP1620			6A	PTS system, nitrogen regulatory component IIA, putative
SP1621			6A	Transcription antiterminator BglG family protein, authentic frameshift
SP1622			U, 6A, 14	Transposase, IS*200* family
RD10				
SP1755		U, 6A, 14		Hypothetical protein
SP1756		U, 6A, 14		Conserved domain protein
SP1757		U, 6A, 14		Conserved hypothetical protein
SP1758		U, 6A, 14		Glycosyl transferase, group 1
SP1759		U, 6A, 14		Preprotein translocase, SecA subunit
SP1760	H	U, 6A, 6B		Conserved domain protein
SP1761		U, 6A, 6B		Hypothetical protein
SP1762		U, 6A, 6B		Hypothetical protein
SP1765		U, 6A, 6B		Glycosyl transferase, family 8
SP1766		U, 6A, 6B		Glycosyl transferase, family 8
SP1767		U, 6A		Glycosyl transferase, family 8
SP1768		U, 6A, 6B		Conserved hypothetical protein
SP1769		U, 6B		Glycosyl transferase, authentic frameshift
SP1770	H			Glycosyl transferase, family 8
SP1771	H			Glycosyl transferase, family 2/glycosyl transferase family 8
SP1772	H	U, 6A, 6B		Cell wall surface anchor family protein
SP1773		U, 6A, 14		IS*630*-Spn1, transposase Orf1/Orf2 degenerate

Continued on following page

TABLE 1—*Continued*

RD and TIGR ID	STM[b]	CGH correlation[a]		TIGR annotation or gene name
		IPD	Noninvasive	
RD13				
SP2158			14	L-Fucose isomerase
SP2159	H	6A	14	Lucolectin-related protein
SP2160		6A	14	Conserved hypothetical protein
SP2161		6A	14	PTS system, IID component
SP2162	H	6A	14	PTS system, IIC component
SP2163		6A	14	PTS system, IIB component
SP2164	H	6A	14	PTS system, IIA component
SP2165		6A	14	Fucose operon FucU protein
SP2166		6A	14	L-Fuculose phosphate aldotase

[a] 6A, 6B, and 14 indicate genes whose presence is correlated with the invasive cohort or the noninvasive cohort. U indicates genes whose presence is correlated with either cohort in a serotype-independent manner.
[b] H indicates genes determined by STM to be required for in vivo passage (Hava and Camilli [24]).

vivo. Furthermore, microarrays have shown this locus to be up-regulated during pneumococcal contact with epithelial cells (38). SP2141-SP2146 encodes a cell wall anchor protein, a glycosyl hydrolase, and four conserved hypothetical proteins. To date, their function is unknown. Other STM core loci of interest include a phophoribosylamide synthase operon(s) (SP0043-SP0056) and ZmpB, a zinc metalloprotease and its surrounding genes (SP0663-SP0667) (14). Table S3 in the supplemental material provides a comprehensive list of the TIGR4 genome, indicates core genes, and lists the genes previously identified by STM.

Comparative genomic analysis. Phylogenetic relationships extrapolated from the present/absent matrix determined that the majority of the 72 clinical isolates clustered into clades (i.e., groups of genetically similar strains) not only as expected, within their own serotype, but also by their ability to cause IPD (Fig. 1a). These findings were consistent with phylogenetic analysis of pulsed-field gel electrophoresis profiles of the 72 isolates (data not shown) and published studies that demonstrated clonal properties contributing to IPD (35, 41, 42). To identify genes whose presence was correlated with IPD, we compared the genomic content of isolates within invasive clades to that of isolates present in the noninvasive clades (Fig. 1b). This analysis was done at the individual serotype level (serotype 6A, 6 noninvasive and 14 invasive; serotype 6B, 7 noninvasive and 15 invasive; serotype 14, 8 noninvasive and 10 invasive) and without regards to serotype (noninvasive, 21 isolates; invasive, 39 isolates). Comparative genomic analysis of clades instead of a direct comparison of invasive and noninvasive isolates was necessary, as this approach disregarded host and environmental factors which may have contributed to the designation of isolates as invasive or noninvasive. For example, it is possible that invasive isolates may have been collected from the nasopharynx during a colonization period, and noninvasive isolates may have been collected from an immunocompromised host. Comparison of clades allowed the comparison of genetic material more frequently associated with IPD to that more frequently associated with asymptomatic colonization. Mixed clades containing equal numbers of invasive and noninvasive isolates were excluded from the analysis due to their ambiguous phenotype. Figure 1b outlines the details of the genetic comparison.

CGA identified 47, 54, and 61 genes whose presence correlated with strains in the invasive clades of serotype 6A, serotype 6B, and serotype 14, respectively; 99 genes were identified whose presence correlated with the invasive phenotype irrespective of serotype (see Table S3 in the supplemental material). In contrast, CGA also identified 65, 24, and 92 genes whose presence correlated with strains in the noninvasive clades of serotype 6A, serotype 6B, and serotype 14; 93 genes were identified whose presence correlated with the noninvasive phenotype irrespective of serotype (see Table S3). Surprisingly, analysis of the 13 RDs determined that the majority of RDs correlated with the noninvasive phenotype (Table 1). This would imply that these RDs contribute to long-term colonization (see below). Interestingly, the presence of the first half of RD8, SP1315-SP1332 (RD8a), correlated with the invasive phenotype, whereas the second half, SP1333-SP1351 (RD8b), correlated with the noninvasive phenotype, a finding that indicates that RD8 is composed of two distinct regions that are rarely present together in clinical isolates. Of particular interest were RD8a and RD10. Genes within these regions were associated with IPD in a serotype-independent manner; moreover, they were associated with atypical (RD8) and highly atypical (RD10) GC contents (45), the latter implying that RD8 and RD10 were acquired horizontally. Furthermore, several genes within RD8a and RD10 have been previously identified by STM (24). Thus, genes within RD8a and RD10 are required for in vivo passage.

RD8a and RD10. RD8 is composed of 40,358 nucleotides encoding 31 predicted coding regions. Based on spatial organization, commonalities in predicted function, and analysis of noncoding regions, it appears that that the genes are organized into five operons; two within RD8a (RD8a1 and -2) and three within RD8b (RD8b1 to -3) (Fig. 2a). As indicated, distribution of RD8 in clinical isolates divides the locus in half, suggesting that RD8 is composed of two RD located adjacent to each other in TIGR4 (Table 1). Interestingly, the presence of RD8a correlates with the invasive phenotype, whereas the presence of RD8b correlates with the noninvasive phenotype. RD8a1 encodes a V-type sodium ATP synthase and is potentially required for homeostasis of metal ions or catalyzing the transmembrane movement of substances. The presence of an oxidoreductase (SP1325) in RD8a2 is suggestive of such a role.

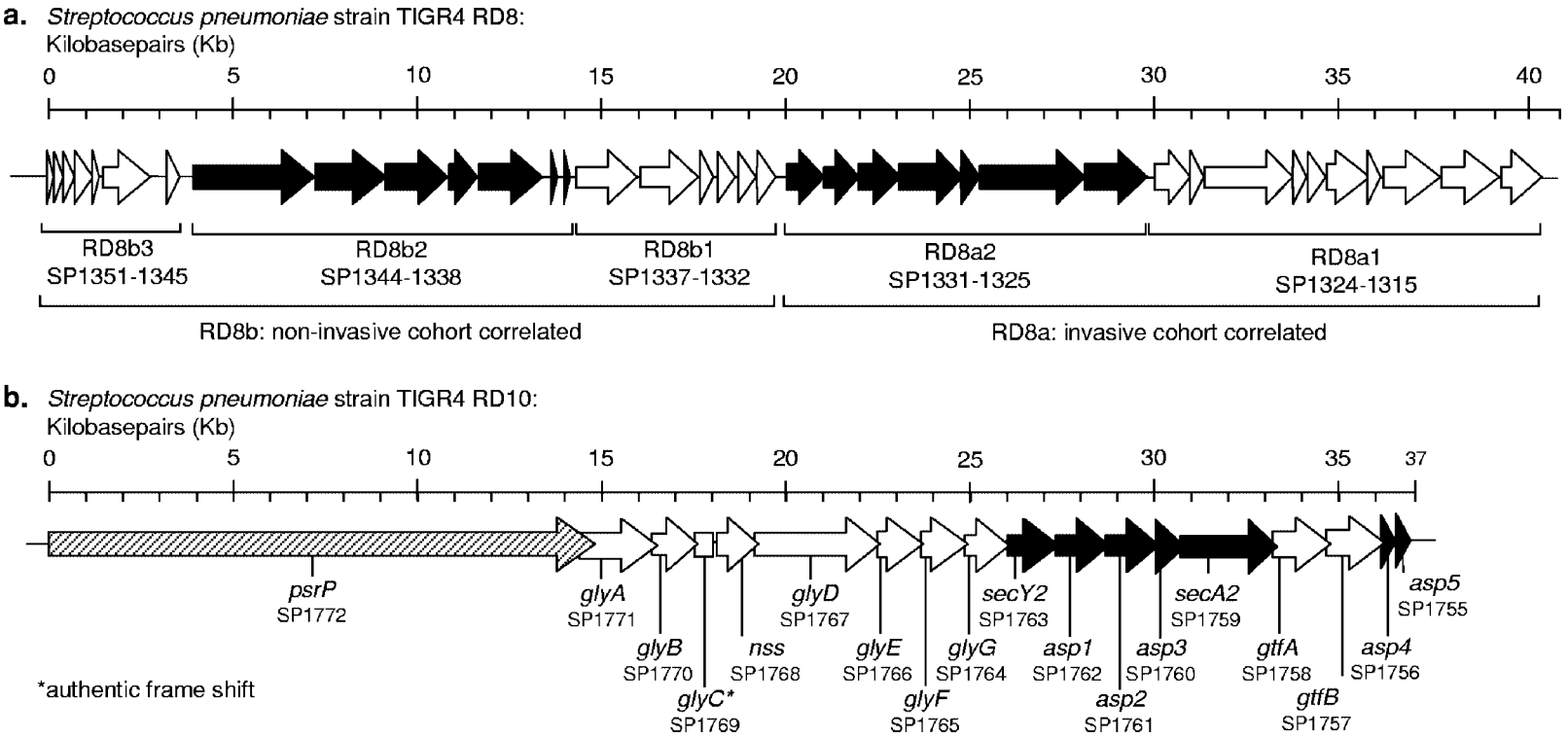

FIG. 2. Schematic representation of gene structures of RD8 and RD10. a) Five predicted operons within RD8 and the division between RD8a and RD8b are shown. b) Illustration of RD10. Genes homologous to glycosyl transferases are indicated in white, whereas those homologous to genes required for secretion are in black.

TABLE 2. Comparison of RD10 to the *S. gordonii gspB-secY2A2* operon

S. pneumoniae RD10			*S. gordonii* homolog			
Gene	Proposed gene symbol	Size (amino acids)	Gene	Size (amino acids)	Function	BLASTP E value
SP1772	*psrP*	4,777	*gspB*	3,072	Platelet-binding protein GspB	4.00E-03
SP1771	*glyA*	697	*gly*	682	Glycosyl transferase family 8	1.00E-103
SP1770	*glyB*	405	*gly*	682	Glycosyl transferase family 8	1.00E-63
SP1769	*glyC*[a]	NA[c]				
SP1768	*nss*	337	*nss*	334	Nucleotide sugar synthetase-like protein Nss	1.00E-103
SP1767	*glyD*	815	*gly*	682	Glycosyl transferase family 8	8.00E-108
SP1766	*glyE*	407	*gly*	682	Glycosyl transferase family 8	6.00E-55
SP1765	*glyF*	399	*gly*	682	Glycosyl transferase family 8	2.00E-29
SP1764	*glyG*[b]	302				
SP1763	*secY2*	406	*secY*	407	Eubacterial SecY protein	4.00E-103
SP1762	*asp1*	527	*asp1*	526	Accessory secretory protein Asp1	7.00E-136
SP1761	*asp2*	512	*asp2*	510	Accessory secretory protein Asp2	2.00E-152
SP1760	*asp3*	147	*asp3*	159	Accessory secretory protein Asp3	3.00E-26
SP1759	*secA2*	791	*secA2*	793	SecA-like protein	0.00
SP1758	*gtfA*	504	*gtfA*	506	Glycosyl transferase group 1	0.00
SP1757	*gtfB*	446	*gtfB*	450	Glycosyl transferase family 8	4.00E-142
SP1756	*asp4*	60	*asp4*	60	Accessory secretory protein Asp4	2.00E-04
SP1755	*asp5*	75	*asp5*	73	Accessory secretory protein Asp5	5.2

[a] Authentic frameshift.
[b] Homologous to a glycosyl transferase family 2 protein not present in *S. gordonii.*
[c] NA, not applicable.

RD8a2 encodes the oxidoreductase, a neuraminidase, and associated sugar-modifying enzymes. Previously, neuraminidase A was demonstrated to contribute to pneumococcal pathogenesis by cleaving sialic acid residues on the surface of the cell and exposing eukaryotic receptors that enhance adhesion (49). The neuraminidase/sugar-modifying enzymes may also alter glycosylated host components, such as secretory immunoglobulin A2, lactoferrin, and C-reactive protein, that attach to the bacteria and facilitate clearance (30).

RD10 in TIGR4 is composed of 36,179 bp encoding 17 predicted coding regions, one of which is truncated by a frameshift mutation (SP1769). Close examination of the nucleotide sequence indicates that the genes overlap, suggesting they may be transcribed as a single unit. BLAST analysis of the predicted amino acid sequences subsequently determined that RD10 is most similar to the *gspB-secY2A2* operon in *Streptococcus gordonii* (6). Figure 2b illustrates RD10, whereas Table 2 indicates the proposed names for the genes in RD10 and their homology to the locus in *S. gordonii*. In *S. gordonii*, the *gspB-secY2A2* operon encodes GspB (also known as Hsa), a 204- to 286-kDa (size is strain specific) protein that has been characterized as a sialic acid binding hemagglutinin. GspB mediates binding to the platelet membrane glycoprotein Ibα and is thought to play a central role in the development of infective endocarditis (7). Other genes in the operon encode glycosyl transferases that glycosylate GspB within the bacterial cytoplasm and an alternate SecA/SecY protein transport system that is responsible for transport of GspB bearing 70 to 100 monosaccharide residues of *N*-acetylglucosamine and glucose (6, 7). GspB in *S. gordonii* is characterized by a 90-amino-acid signal peptide that is three times longer than signals for export meditated by SecA (9), and mutations in SecA2 or SecY2 result in accumulation of GspB in the cytoplasm (7, 8).

Like *S. gordonii*, RD10 encodes an extremely long serine-rich protein (Fig. 2b). SP1772 (hereafter termed PsrP, for *p*neumococcal *s*erine-rich *r*epeat *p*rotein) is composed of 14,331 bp encoding a 4,776-amino-acid protein with a predicted molecular mass of 412 kDa. Like GspB, PsrP consists of a large signal peptide (72 amino acids), a short serine-rich repeat region (SRR1; 49 amino acids), a basic region (272 amino acids), a second extremely large serine-rich repeat area (SRR2; 4,319 amino acids), and a cell wall anchor domain at the carboxy terminus (62 amino acids) (Fig. 3). The serine-rich repeat region is composed of approximately 539 SASASAST repeats. RD10 also contains near-identical homologs to each gene in the entire *gspB-secY2A2* operon; however, unlike *S. gordonii*, RD10 contains an additional seven glycosyl transferases that occupy the region between *gly* and *nss* and *nss* and *secY2* (Fig. 2b; Table 2).

PsrP is required for virulence but not nasopharyngeal colonization. To confirm our hypothesis that IPD-correlated genes contribute to virulence, PsrP (SP1772), the adhesin encoded in RD10, was deleted by insertion duplication mutagenesis in TIGR4 (T4 ΔPsrP), a virulent serotype 4 isolate (15). Intranasal challenge of 5-week-old BALB/cJ mice with the mutant and wild type (WT) showed that deletion of *psrP* slowed bacterial entry into the bloodstream (Fig. 4B); moreover, the mutant was unable to kill mice as effectively as the WT (Fig. 4A) (WT, 6% survival; T4 ΔPsrP, 75% survival; $P = 0.002$). Examination of nasal lavage in these mice 2 days postchallenge showed that deletion of *psrP* did not affect nasopharyngeal colonization (WT, 2.10×10^6; T4 ΔPsrP, 1.43×10^6; $P = 0.757$); this finding was independently confirmed in a low-dose (10^4 CFU) nasopharyngeal colonization model (Fig. 4C). Moreover, T4 ΔPsrP grew normally in blood and was comparable to wild type following intravenous injection of mice (data not shown). Thus, disruption of *psrP* (RD10) affects only the ability to progress into the bloodstream, presumably from the lungs, and not colonization or survival in blood.

Given (i) the highly atypical DNA content of RD8a and

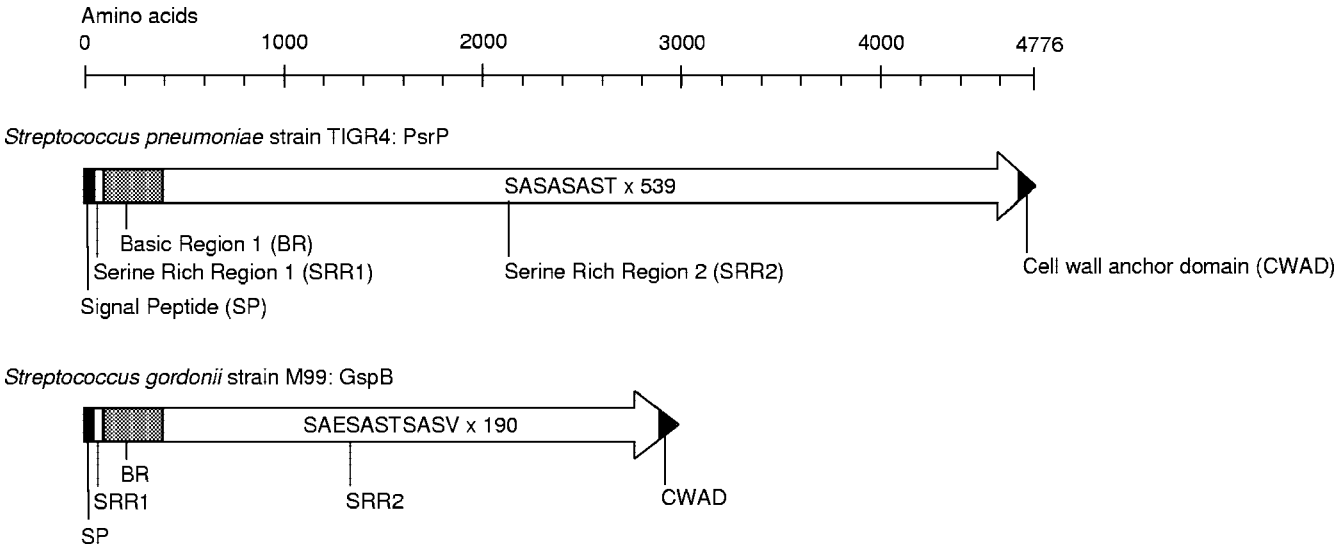

FIG. 3. Domain structures of PsrP and GspB.

RD10 (45), (ii) the correlation of these loci with human IPD in a serotype-independent manner, (iii) the finding that transposon mutagenesis (STM) of genes within these loci reduced the ability of TIGR4 to passage through mice (24), and (iv) the attenuated phenotype of T4 ΔPsrP in mice, it is reasonable to suggest that RD8a and RD10 are pathogenicity islands that facilitate the development of IPD in humans. Ongoing studies are focused on characterizing the function of these RDs and confirming their role in virulence.

Table S3 in the supplemental material lists the complete annotated genome of TIGR4 and indicates other IPD-correlated genes identified by the three STM studies. Several small operons were also identified that are not discussed.

RD8b and genes correlated with the noninvasive phenotype. Molecular epidemiology has demonstrated that within serotypes, invasive and noninvasive clones exist (42). Certain clones are associated with a high attack rate and are rarely isolated from healthy carriers; alternatively, certain clones are routinely isolated from healthy carriers and are rarely responsible for invasive disease. One interpretation of these observations is that pneumococci within invasive serotypes are adapted to spread as the result of symptoms brought on by invasive disease (e.g., coughing). Alternatively, noninvasive clones must be adapted for spread during asymptomatic colonization. Presumably, the inability to cause symptoms that facilitate infectivity must be offset by a prolonged period of colonization during which the pneumococcus has an equal opportunity to spread. This view is supported by the observation that noninvasive serotypes colonize the nasopharynx for extended periods; moreover, they do so for longer periods than invasive serotypes (17).

CGA indicated that the majority of RDs correlated with the noninvasive cohorts. In addition to RD8b, genes within RD1, RD2, RD6, and RD7 correlated with noninvasive colonization in a serotype-independent manner. Likewise, genes within RD5, RD9, and RD13 did so in a serotype-dependent manner. It is interesting that the preponderance of known virulence determinants that correlated with the noninvasive cohort has been demonstrated to contribute to nasopharyngeal colonization. For example, RD1 encodes an immunoglobulin A1 protease (50). RD4 encodes the *rlr* pathogenicity islet demonstrated by Hava et al. to be required for colonization of the nasopharynx and lung infection but is dispensable for systemic infection (24, 25). RD13 encodes a fucose transferase system; Coyne et al. demonstrated that surface fucosylation of bacteria facilitates colonization (16). Noninvasive correlated virulence genes outside of RD include choline binding protein F (serotype 6A; SP0391) (22), neuraminidase A (serotype independent; SP1639) (48, 49), and choline binding protein A (serotype independent; SP2190) (37). Multiple studies have clearly demonstrated that these genes contribute to nasopharyngeal colonization. Thus, the sorting of these genes with the noninvasive cohort may reflect the necessity of noninvasive isolates to colonize the nasopharynx more efficiently. RDs may also represent "antivirulence genes" (20). Using *Shigella flexneri* and enteroinvasive *Escherichia coli* as a model system, Maurelli et al. demonstrated that the presence of certain pathogenicity islands attenuates virulence and loss of these regions enhances the virulence of the organism (19, 34). Thus, in addition to providing a mechanism for prolonged adhesion, it is possible that some RDs may facilitate asymptomatic colonization by reducing virulence. We are currently performing experiments to determine if this is the case.

Serotype-dependent distribution of RDs. While capsular polysaccharide is absolutely required for virulence, the relationship between serotype and genome is complex and not fully understood. For example, conversion of an isolate with one capsule type to another has variable results. Kelly et al. demonstrated that serotype conversion of an avirulent serotype 6B isolate to serotype 3 increased virulence, whereas conversion of a highly virulent serotype 5 isolate to type 3 attenuated the bacterium (29). Thus, for each serotype a distinct set of genetic requirements may be required to cause IPD. Examination of RD distribution determined that several RDs were distributed in a serotype-dependent manner. For example, the presence of RD5 correlated with IPD for serotype 6A isolates but corre-

lated with the noninvasive phenotype among serotype 14 isolates. Likewise, the presence of RD13 correlated with noninvasive isolates of 6A and invasive isolates of 14. One possible explanation for these serotype-dependent distributions may be that RDs are required to complement the unique physiological properties of each capsule type. For example, a pneumococcus containing a capsule type that is susceptible to opsonophagocytosis would require genetic determinants that counteract this host defense. Alternatively, a capsule type that is highly resistant to opsonophagocytosis may require additional adhesins to overcome the hindrance to adhesion imposed by the capsule.

Considerations. Genomic profiling and subsequent analysis of genes are powerful new tools to bring to bear on the dissection of bacterial pathogenesis. Nonetheless, limitations exist that need to be considered. Foremost, microarrays can only detect genes that are included on the array, and targeted genes must have considerable homology to the nucleic acid sequence of the probes. For example, *pspA*, a variable virulence determinant (12), was detected in only 14 of the 72 clinical isolates (data not shown). To determine if these were false-negative results, primers designed from the TIGR4 sequence were used to amplify full-length *pspA* from the 20 serotype 14 clinical isolates (microarrays detected *pspA* in 8 of 20). Successful PCR amplification of *pspA* from each of the 20 clinical isolates indicated that the microarrays failed to detect the presence of *pspA* in the majority of the isolates. Thus, in addition to the absence of the gene, significant gene variation or overlap of the microarray probe with a hypervariable region of the gene(s) may also result in false-negative results. Ultimately, empirical analysis of a negative signal is required for final determination of the absence of a gene. Thus, the designation of a "core gene" in this study pertains only to nonvariable genes; variable genes, such as *pspA*, may belong to the genetic core but remained undetected. Interestingly, four different sizes of amplified *pspA* were observed, and these differences corresponded to the distinct clades identified by the phylogenetic analysis (data not shown).

Lastly, we hypothesize that genes whose presence is correlated with an invasive cohort contribute to invasive disease. This view does not consider the impact of differential gene expression between isolates and assumes that the presence of a gene confers a gain of function. It is highly likely that core virulence determinants are expressed differentially between invasive and noninvasive isolates. Comparative transcriptome analysis would be required to determine if this were the case. Furthermore, the comparative analysis described in this study was dependent on sorting of the strains without regard to the immune status of the infected individual. Despite these caveats, we have demonstrated that significantly valuable information can be gained by CGA of cohorts of clinical strains, justifying study of other serotypes/clonotypes.

Conclusion. We have described the use of CGH analysis to identify genes whose presence is correlated with invasive disease. This is distinct from a recent study by Shen et al., who used CGH to identify novel genes present in *S. pneumoniae* isolates collected from symptomatic pediatric patients (43), and is in contrast to a study by Lindsay et al., who using CGH failed to correlate gene diversity with *Staphylococcus aureus* invasion (33). Using CGH, we have identified a candidate core genome of the pneumococcus and determined that the distri-

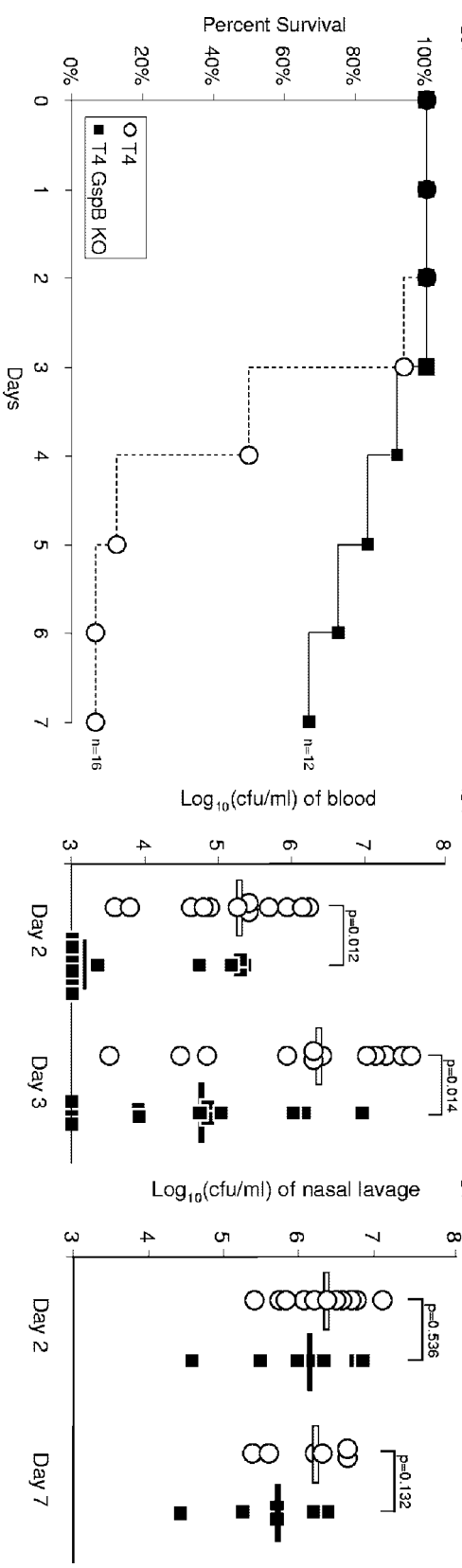

FIG. 4. Mutants deficient in PsrP are attenuated in an intranasal challenge model of invasive disease but are able to colonize the nasopharynx normally. a) Kaplan-Meier plot illustrating the enhanced survival of mice infected intranasally with T4 ΔPsrP (■) versus wild type (○) following intranasal challenge with 10^7 CFU. b) Bacterial titers in the blood 2 days and 3 days post-intranasal challenge. c) Bacterial titers in nasal lavage of mice challenged with 10^4 CFU 2 days and 7 days post-intranasal challenge. Horizontal bars indicate the median values.

butions of genes, in particular RDs, are correlated with the propensity of an isolate to cause invasive disease. We have identified two RDs, RD8a and RD10, which encode genes homologous to known virulence determinants and have been shown by STM to be required in vivo. We confirmed a role for RD10 by deletion of *psrP* and the observation of an attenuated phenotype in mice. Use of comparative genomics in this manner is, in effect, an in silico subtractive hybridization. Cross-reference of these findings with other published reports, such as transcriptional data in vivo, could serve to further elucidate the mechanisms by which the pneumococcus causes invasive disease.

ACKNOWLEDGMENTS

We thank Robert Fleischmann and Scott Peterson at the Pathogen Functional Genomics Resource Center at The Institute for Genomic Research for providing the pneumococcal microarrays necessary for this project. We thank Geli Gao and Nelson Velazquez for invaluable technical assistance.

This work was supported by NIH RO1 AI27913, The American Lebanese Syrian Associated Charities, and project grant 121919 of the Executive Research Committee Research Fund at UTHSCSA.

REFERENCES

1. **Anjum, M. F., C. Marooney, M. Fookes, S. Baker, G. Dougan, A. Ivens, and M. J. Woodward.** 2005. Identification of core and variable components of the *Salmonella enterica* subspecies I genome by microarray. Infect. Immun. **73:** 7894–7905.
2. **Anonymous.** 2000. Preventing pneumococcal disease among infants and young children. Recommendations of the Advisory Committee on Immunization Practices (ACIP). Morb. Mortal. Wkly. Rep. Recomm. Rep. **49:**1–35.
3. **Anonymous.** 1999. Pneumococcal vaccines. WHO position paper. Wkly. Epidemiol. Rec. **74:**177–183.
4. **Austrian, R.** 1986. Some aspects of the pneumococcal carrier state. J. Antimicrob. Chemother. **18**(Suppl. A)**:**35–45.
5. **Babl, F. E., S. I. Pelton, S. Theodore, and J. O. Klein.** 2001. Constancy of distribution of serogroups of invasive pneumococcal isolates among children: experience during 4 decades. Clin. Infect. Dis. **32:**1155–1161.
6. **Bensing, B. A., B. W. Gibson, and P. M. Sullam.** 2004. The *Streptococcus gordonii* platelet binding protein GspB undergoes glycosylation independently of export. J. Bacteriol. **186:**638–645.
7. **Bensing, B. A., J. A. Lopez, and P. M. Sullam.** 2004. The *Streptococcus gordonii* surface proteins GspB and Hsa mediate binding to sialylated carbohydrate epitopes on the platelet membrane glycoprotein Ibα. Infect. Immun. **72:**6528–6537.
8. **Bensing, B. A., and P. M. Sullam.** 2002. An accessory sec locus of Streptococcus gordonii is required for export of the surface protein GspB and for normal levels of binding to human platelets. Mol. Microbiol. **44:**1081–1094.
9. **Bensing, B. A., D. Takamatsu, and P. M. Sullam.** 2005. Determinants of the streptococcal surface glycoprotein GspB that facilitate export by the accessory Sec system. Mol. Microbiol. **58:**1468–1481.
10. **Berger, J. A., S. Hautaniemi, A. K. Jarvinen, H. Edgren, S. K. Mitra, and J. Astola.** 2004. Optimized LOWESS normalizatin parameter selection for DNA microarray data. BMC Bioinformatics **9:**194.
11. **Bricker, A. L., and A. Camilli.** 1999. Transformation of a type 4 encapsulated strain of Streptococcus pneumoniae. FEMS Microbiol. Lett. **172:**131–135.
12. **Briles, D. E., M. J. Crain, B. M. Gray, C. Forman, and J. Yother.** 1992. Strong association between capsular type and virulence for mice among human isolates of *Streptococcus pneumoniae*. Infect. Immun. **60:**111–116.
13. **Butler, J. C., R. F. Breiman, H. B. Lipman, J. Hofmann, and R. R. Facklam.** 1995. Serotype distribution of Streptococcus pneumoniae infections among preschool children in the United States, 1978–1994: implications for development of a conjugate vaccine. J. Infect. Dis. **171:**885–889.
14. **Camilli, R., E. Pettini, M. D. Grosso, G. Pozzi, A. Pantosti, and M. R. Oggioni.** 2006. Zinc metalloproteinase genes in clinical isolates of Streptococcus pneumoniae: association of the full array with a clonal cluster comprising serotypes 8 and 11A. Microbiology **152:**313–321.
15. **Chen, J. D., and D. A. Morrison.** 1988. Construction and properties of a new insertion vector, pJDC9, that is protected by transcriptional terminators and useful for cloning of DNA from Streptococcus pneumoniae. Gene **64:**155–164.
16. **Coyne, M. J., B. Reinap, M. M. Lee, and L. E. Comstock.** 2005. Human symbionts use a host-like pathway for surface fucosylation. Science **307:** 1778–1781.
17. **Crook, D. W., A. D. Brueggemann, K. L. Sleeman, and T. E. Peto.** 2004. Pneumococcal carriage, p. 136–147. *In* E. I. Tuomanen, T. J. Mitchell, D. A. Morrison, and B. G. Spratt (ed.), The pneumococcus. ASM Press, Washington, D.C.
18. **Davidson, M., A. J. Parkinson, L. R. Bulkow, M. A. Fitzgerald, H. V. Peters, and D. J. Parks.** 1994. The epidemiology of invasive pneumococcal disease in Alaska, 1986–1990—ethnic differences and opportunities for prevention. J. Infect. Dis. **170:**368–376.
19. **Day, W. A., Jr., R. E. Fernandez, and A. T. Maurelli.** 2001. Pathoadaptive mutations that enhance virulence: genetic organization of the *cadA* regions of *Shigella* spp. Infect. Immun. **69:**7471–7480.
20. **Day, W. A., and A. T. Maurelli.** 2006. Black holes and antivirulence genes: selection for gene loss as part of the evolution of bacterial pathogens, p. 109–122. *In* H. S. Seifert and V. J. DiRita (ed.), The evolution of bacterial pathogens. ASM Press, Washington, D.C.
21. **Dopazo, J., A. Mendoza, J. Herrero, F. Caldara, Y. Humbert, L. Friedli, M. Guerrier, E. Grand-Schenk, C. Gandin, M. de Francesco, A. Polissi, G. Buell, G. Feger, E. Garcia, M. Peitsch, and J. F. Garcia-Bustos.** 2001. Annotated draft genomic sequence from a Streptococcus pneumoniae type 19F clinical isolate. Microb. Drug Resist. **7:**99–125.
22. **Gosink, K. K., E. R. Mann, C. Guglielmo, E. I. Tuomanen, and H. R. Masure.** 2000. Role of novel choline binding proteins in virulence of *Streptococcus pneumoniae*. Infect. Immun. **68:**5690–5695.
23. **Hakenbeck, R., N. Balmelle, B. Weber, C. Gardes, W. Keck, and A. de Saizieu.** 2001. Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of *Streptococcus pneumoniae*. Infect. Immun. **69:**2477–2486.
24. **Hava, D. L., and A. Camilli.** 2002. Large-scale identification of serotype 4 Streptococcus pneumoniae virulence factors. Mol. Microbiol. **45:**1389–1406.
25. **Hava, D. L., C. J. Hemsley, and A. Camilli.** 2003. Transcriptional regulation in the *Streptococcus pneumoniae rlrA* pathogenicity islet by RlrA. J. Bacteriol. **185:**413–421.
26. **Henrichsen, J.** 1995. Six newly recognized types of *Streptococcus pneumoniae*. J. Clin. Microbiol. **33:**2759–2762.
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. E. 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. **Kaushal, D., and C. Naeve.** 2004. Analyzing and visualizing expression data with Spotfire, p. 7.9.1–7.9.43. *In* A. D. Baxevanis, D. B. Davison, R. D. M. Page, G. A. Petsko, L. D. Stein, and S. G.D. (ed.), Current protocols in bioinformatics. John Wiley and Sons, Inc., New York, N.Y.
29. **Kelly, T., J. P. Dillard, and J. Yother.** 1994. Effect of genetic switching of capsular type on virulence of *Streptococcus pneumoniae*. Infect. Immun. **62:**1813–1819.
30. **King, S. J., K. R. Hippe, J. M. Gould, D. Bae, S. Peterson, R. T. Cline, C. Fasching, E. N. Janoff, and J. N. Weiser.** 2004. Phase variable desialylation of host proteins that bind to Streptococcus pneumoniae in vivo and protect the airway. Mol. Microbiol. **54:**159–171.
31. **Lacks, S., and R. D. Hotchkiss.** 1960. A study of the genetic material determining an enzyme in pneumococcus. Biochim. Biophys. Acta **39:**508–518.
32. **Lau, G. W., S. Haataja, M. Lonetto, S. E. Kensit, A. Marra, A. P. Bryant, D. McDevitt, D. A. Morrison, and D. W. Holden.** 2001. A functional genomic analysis of type 3 Streptococcus pneumoniae virulence. Mol. Microbiol. **40:**555–571.
33. **Lindsay, J. A., C. E. Moore, N. P. Day, S. J. Peacock, A. A. Witney, R. A. Stabler, S. E. Husain, P. D. Butcher, and J. Hinds.** 2006. Microarrays reveal that each of the 10 dominant lineages of *Staphylococcus aureus* has a unique combination of surface-associated and regulatory genes. J. Bacteriol. **188:** 669–676.
34. **Maurelli, A. T., R. E. Fernandez, C. A. Bloch, C. K. Rode, and A. Fasano.** 1998. "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of Shigella spp. and enteroinvasive Escherichia coli. Proc. Natl. Acad. Sci. USA **95:**3943–3948.
35. **Mizrachi Nebenzahl, Y., N. Porat, S. Lifshitz, S. Novick, A. Levi, E. Ling, O. Liron, S. Mordechai, R. K. Sahu, and R. Dagan.** 2004. Virulence of Streptococcus pneumoniae may be determined independently of capsular polysaccharide. FEMS Microbiol. Lett. **233:**147–152.
36. **Obert, R.** 2005. Fischer exact, Excel addin. [Online.] http://www.obertfamily.com/software/fisherexact.html.
37. **Orihuela, C. J., G. Gao, K. P. Francis, J. Yu, and E. I. Tuomanen.** 2004. Tissue-specific contributions of pneumococcal virulence factors to pathogenesis. J. Infect. Dis. **190:**1661–1669.
38. **Orihuela, C. J., J. N. Radin, J. E. Sublett, G. Gao, D. Kaushal, and E. I. Tuomanen.** 2004. Microarray analysis of pneumococcal gene expression during invasive disease. Infect. Immun. **72:**5582–5596.
39. **Polissi, A., A. Pontiggia, G. Feger, M. Altieri, H. Mottl, L. Ferrari, and D.**

Simon. 1998. Large-scale identification of virulence genes from *Streptococcus pneumoniae*. Infect. Immun. **66:**5620–5629.
40. **Regev-Yochay, G., M. Raz, R. Dagan, N. Porat, B. Shainberg, E. Pinco, N. Keller, and E. Rubinstein.** 2004. Nasopharyngeal carriage of Streptococcus pneumoniae by adults and children in community and family settings. Clin. Infect. Dis. **38:**632–639.
41. **Sandgren, A., B. Albiger, C. J. Orihuela, E. Tuomanen, S. Normark, and B. Henriques-Normark.** 2005. Virulence in mice of pneumococcal clonal types with known invasive disease potential in humans. J. Infect. Dis. **192:**791–800.
42. **Sandgren, A., K. Sjostrom, B. Olsson-Liljequist, B. Christensson, A. Samuelsson, G. Kronvall, and B. Henriques Normark.** 2004. Effect of clonal and serotype-specific properties on the invasive capacity of Streptococcus pneumoniae. J. Infect. Dis. **189:**785–796.
43. **Shen, K., J. Gladitz, P. Antalis, B. Dice, B. Janto, R. Keefe, J. Hayes, A. Ahmed, R. Dopico, N. Ehrlich, J. Jocz, L. Kropp, S. Yu, L. Nistico, D. P. Greenberg, K. Barbadora, R. A. Preston, J. C. Post, G. D. Ehrlich, and F. Z. Hu.** 2006. Characterization, distribution, and expression of novel genes among eight clinical isolates of *Streptococcus pneumoniae*. Infect. Immun. **74:**321–330.
44. **Swofford, D.** 1997. PAUP*: phylogenetic analysis using parsimony (* and other methods), 4.0 ed. Sinauer Associates, Sunderland, Mass.
45. **Tettelin, H., and S. K. Hollingshead.** 2004. Comparative genomics of *Streptococcus pneumoniae*: intrastrain diversity and genome plasticity, p. 15–29. *In* E. I. Tuomanen, T. Mitchell, D. A. Morrison, and B. G. Spratt (ed.), The pneumococcus. ASM Press, Washington, D.C.
46. **Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. Deboy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser.** 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome." Proc. Natl. Acad. Sci. USA **102:**13950–13955.
47. **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.
48. **Tong, H. H., L. E. Blue, M. A. James, and T. F. DeMaria.** 2000. Evaluation of the virulence of a *Streptococcus pneumoniae* neuraminidase-deficient mutant in nasopharyngeal colonization and development of otitis media in the chinchilla model. Infect. Immun. **68:**921–924.
49. **Tong, H. H., M. James, I. Grants, X. Liu, G. Shi, and T. F. DeMaria.** 2001. Comparison of structural changes of cell surface carbohydrates in the eustachian tube epithelium of chinchillas infected with a Streptococcus pneumoniae neuraminidase-deficient mutant or its isogenic parent strain. Microb. Pathog. **31:**309–317.
50. **Weiser, J. N., D. Bae, C. Fasching, R. W. Scamurra, A. J. Ratner, and E. N. Janoff.** 2003. Antibody-enhanced pneumococcal adherence requires IgA1 protease. Proc. Natl. Acad. Sci. USA **100:**4215–4220.

Editor: J. N. Weiser