JOURNAL OF BACTERIOLOGY, Jan. 2007, p. 38–51 Vol. 189, No. 1
0021-9193/07/$08.00+0 doi:10.1128/JB.01148-06

Genome Sequence of Avery's Virulent Serotype 2 Strain D39 of *Streptococcus pneumoniae* and Comparison with That of Unencapsulated Laboratory Strain R6[▽]‡

Joel A. Lanie,[1]† Wai-Leung Ng,[1]† Krystyna M. Kazmierczak,[1]† Tiffany M. Andrzejewski,[1]
Tanja M. Davidsen,[2] Kyle J. Wayne,[1] Hervé Tettelin,[2] John I. Glass,[3]
and Malcolm E. Winkler[1]*

Department of Biology, Indiana University, Bloomington, Indiana 47405[1]; The Institute for Genomic Research, Rockville, Maryland 20850[2]; and J. Craig Venter Institute, Rockville, Maryland 20850[3]

Received 28 July 2006/Accepted 6 October 2006

Streptococcus pneumoniae **(pneumococcus) is a leading human respiratory pathogen that causes a variety of serious mucosal and invasive diseases. D39 is an historically important serotype 2 strain that was used in experiments by Avery and coworkers to demonstrate that DNA is the genetic material. Although isolated nearly a century ago, D39 remains extremely virulent in murine infection models and is perhaps the strain used most frequently in current studies of pneumococcal pathogenesis. To date, the complete genome sequences have been reported for only two *S. pneumoniae* strains: TIGR4, a recent serotype 4 clinical isolate, and laboratory strain R6, an avirulent, unencapsulated derivative of strain D39. We report here the genome sequences and new annotation of two different isolates of strain D39 and the corrected sequence of strain R6. Comparisons of these three related sequences allowed deduction of the likely sequence of the D39 progenitor and mutations that arose in each isolate. Despite its numerous repeated sequences and IS elements, the serotype 2 genome has remained remarkably stable during cultivation, and one of the D39 isolates contains only five relatively minor mutations compared to the deduced D39 progenitor. In contrast, laboratory strain R6 contains 71 single-base-pair changes, six deletions, and four insertions and has lost the cryptic pDP1 plasmid compared to the D39 progenitor strain. Many of these mutations are in or affect the expression of genes that play important roles in regulation, metabolism, and virulence. The nature of the mutations that arose spontaneously in these three strains, the relative global transcription patterns determined by microarray analyses, and the implications of the D39 genome sequences to studies of pneumococcal physiology and pathogenesis are presented and discussed.**

Streptococcus pneumoniae (pneumococcus) is a major human respiratory pathogen that causes several serious diseases, including pneumonia, otitis media (ear infection), sinusitis, meningitis, and septicemia (reviewed in references 81, 112, and 114). Invasive pneumococcal diseases result in high rates of mortality and morbidity, especially among young, elderly, debilitated, and immunosuppressed individuals (54, 55). It is estimated that more than 1 million people die each year from pneumococcal infections worldwide, especially in developing countries (59, 80). In the United States and elsewhere, resistance to a range of antibiotics is increasing at an alarming rate among clinical isolates of *S. pneumoniae* (6, 60). As part of its life cycle, pneumococcus exists as a commensal bacterium that inhabits and colonizes the nasopharynx of up to 20 and 50% of healthy adults and children, respectively, at any time (81, 113). The transition from commensal bacterium to opportunistic pathogen often occurs after another respiratory tract infection. For example, pneumococcal pneumonia has been a leading secondary infection and cause of death during influenza pandemics (9).

Strains of *S. pneumoniae* are categorized into serotypes based on the structures of their exopolysaccharide capsules, of which there are more than 85 kinds (35, 117). To date, the complete genome sequences have been determined for only two strains of *S. pneumoniae* (48, 110), and partial sequences are being determined for other serotype strains that are prevalent as current clinical isolates (see references 20, 28, 42, 104, and 109 and http://genome.microbio.uab.edu/strep/info/and http://www.sanger.ac.uk/Projects/S_pneumoniae/). The genome sequences of the virulent serotype 4 strain, TIGR4, which is a recent clinical isolate (110), and the avirulent, unencapsulated laboratory strain R6 revealed numerous aspects of the metabolism and genome organization of *S. pneumoniae* (48, 110). A startling finding from these initial comparisons is the diversity among the genomes of the different serotypes of *S. pneumoniae* (109). In fact, as many as 10% of the genes may be substantially diverged or absent in comparisons between the genomes of different serotype strains (20, 42, 44, 109). In addition, pneumococcal genomes contain a relatively large number of insertion elements, transposon remnants, and repeat sequences, which suggests that pneumococcal genomes may exhibit considerable plasticity (109). Because of this extreme diversity, it

* Corresponding author. Mailing address: Department of Biology, Indiana University Bloomington, Jordan Hall 142, Bloomington, IN 47405. Phone: (812) 856-1318. Fax: (812) 855-6705. E-mail: mwinkler@bio.indiana.edu.

† J.A.L., W.-L.N., and K.M.K. contributed equally to this study.

‡ Supplemental material for this article may be found at http://jb.asm.org/.

[▽] Published ahead of print on 13 October 2006.

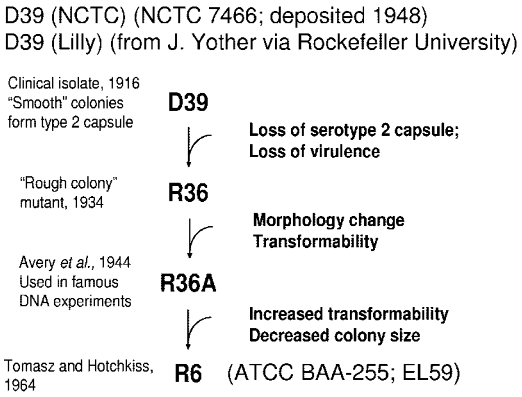


FIG. 1. Pedigree of strains D39 (NCTC), D39 (Lilly), and R6 whose genome sequences are compared in the present study. (The figure was modified from reference 12.) See the introduction and reference 12 for details.

has become imperative to study colonization and virulence in a number of different serotypes of *S. pneumoniae*, each of which causes a distinct pattern of infection in animal models (16, 44, 88).

The genome sequence of strain R6 was determined because it has become a standard laboratory strain for the study of fundamental cellular processes, such as transformation, cell division, and peptidoglycan biosynthesis (e.g., see references 1, 25, 58, 78, and 84). Strain R6 was derived about 40 years ago from strain R36A (12, 89), which was derived about 60 years ago by Avery and coworkers for the landmark demonstration that DNA is the genetic material (Fig. 1) (8, 12, 76). R36A was, in turn, derived from strain R36, which was derived originally from the clinical serotype 2 isolate D39 (Fig. 1) (8, 12, 76). R36, R36A, and R6 were isolated because they were nonreverting mutants lacking type 2 capsule and were increasingly competent for natural transformation, before the discovery of the competence stimulatory peptide (45). Providing an isogenic avirulent strain for comparison with serotype 2 progenitor D39 was not a goal of these derivations.

Besides its historical significance, strain D39 has been adopted as a leading model of pneumococcal pathogenesis (e.g., see references 14, 17, 65, 87, 88, and 96). Although strain D39 was isolated from a patient about 90 years ago (Fig. 1), it remains extremely virulent in animal models of infection (e.g., see below and references 14, 16, 17, 87, 88, and 96) and has proven to be highly tractable in genetic mutant constructions (e.g., see references 14, 16, 17, and 96). Aside from a large deletion of part of the capsule (*cps*) biosynthesis locus, it has often been assumed that the genetic composition of strain D39 and R6 differ by only a limited number of mutations (e.g., see reference 12). However, there is no direct evidence for this assumption, which is partly based on the fact that mutagenesis is not mentioned in the derivation of strains R36, R36A, or R6 (8, 12, 76). It has also been unclear whether pathogenic islands present in strain TIGR4 (110) but absent from strain R6 (48) were lost during derivation and passage. Moreover, not having the genome sequence of the most commonly used strain of *S. pneumoniae* has hampered the interpretation of physiology and pathogenesis studies.

We report here the sequence of two different isolates of serotype 2 strain D39. One isolate designated as D39 (NCTC) is from the National Collection of Type Cultures (NCTC; United Kingdom) and is currently widely used in studies of pneumococcal physiology and pathogenesis (e.g., see references 16 and 17). According to their records, D39 (NCTC) was deposited in the NCTC collection about 60 years ago (Fig. 1). The second isolate designated D39 (Lilly), which was used previously by this laboratory (96), was obtained from Lilly Research Laboratories, who obtained it from Janet Yother (University of Alabama at Birmingham [UAB]). We do not know the exact propagation history of D39 (Lilly), but the two D39 isolates have been separated in captivity by at least 21 years and possibly longer (Janet Yother, unpublished data). Comparison of the genome sequences of these two D39 isolates with the corrected genome sequence of strain R6 indicates mutations that likely arose separately in each of the three strains compared to the D39 progenitor strain. This analysis shows that the genome of serotype 2 strain D39 has been unexpectedly stable during propagation in captivity and that the genome sequence of strain D39 (NCTC) remains extremely close to that of the deduced D39 progenitor strain. In contrast, there are over 80 mutational differences between strain R6 and the D39 progenitor strain (Fig. 1). One important implication of the present study is that some metabolic relationships observed in laboratory strain R6 may not directly apply to pathogenic strain D39. We also report the occurrence of unusual deletion or insertion mutations of immediately adjacent direct repeat sequences in *S. pneumoniae* and the sometimes significant effects of seemingly minor mutations on global transcription patterns of these three strains.

MATERIALS AND METHODS

Bacterial strains. A single-colony isolate of strain R6 was assigned the unique strain designation EL59 (48). This isolate originated from the same stock as the sequenced R6 strain deposited in the American Type Culture Collection (ATCC BAA-255) (48). A single-colony isolate of strain D39 obtained from Lilly Research Laboratories (Eli Lilly and Co., Indianapolis, IN) was assigned the unique strain designation IU1680 and is referred to as D39 (Lilly) herein. A second D39 isolate, NCTC 7466, was purchased from the NCTC (London, United Kingdom). The lyophilized sample of NCTC 7466 was resuspended in brain heart infusion (Bacto BHI; Becton Dickinson) broth upon receipt, spread onto Trypticase soy agar II (modified) (Becton Dickinson) containing 5% (vol/vol) defibrinated sheep blood (TSAII BA) for single-colony isolation, and incubated at 37°C in an atmosphere of 5% CO_2. All colonies were alpha-hemolytic and white, but we observed two colony types: larger, shiny, smooth colonies and smaller, flatter, rough colonies. A representative colony of each type was single colony isolated three times on blood agar plates. The representative smooth colony was assigned the unique strain designation IU1690 and is referred to as D39 (NCTC). D39 (NCTC) and D39 (Lilly), whose genome sequences are reported here, are sensitive to optochin (36) and are positive in Quellung reactions to type 2 capsule antiserum (Statens Serum Institut, Denmark) (36). The representative smaller rough colony (strain IU1691) was optochin sensitive but was negative in the Quellung reaction to antisera against type 2, 3, and 4 capsules. IU1691 had spontaneously lost its ability to synthesize capsule and was not characterized further.

General DNA sequencing strategy. Previous determinations of the R6 and TIGR4 genome sequences (48, 109, 110) were hampered by the inability to recover random shotgun clones, presumably due to instability or toxicity, and by difficulties in assembling repeat-rich pneumococcal sequences (see reference 109). To avoid these problems, we used the R6 sequence as a blueprint to design PCR primers and performed a PCR walk around the chromosomes of D39 (NCTC) and D39 (Lilly) (see Fig. 2 and below). We spaced PCR primers to synthesize amplicons of about 900 bp with 300-bp overlaps between amplicons. The coverage of the D39 sequence was one to two separate sequence runs for each amplicon from D39 (NCTC) and D39 (Lilly) in regions that exactly

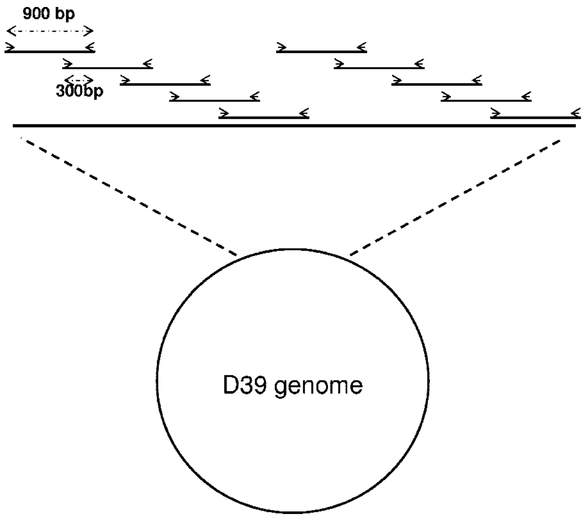


FIG. 2. Primer-walk strategy used to sequence the genomes of strains D39 (NCTC) and D39 (Lilly). Overlapping PCR amplicons were synthesized and sequenced by using primers based on the published sequence of strain R6 (48). See the text for details. The figure is not drawn to scale.

matched the previously published R6 sequence (48). In regions that did not match the R6 sequence or that were different between the two D39 strains, we confirmed sequence differences with two to four separate sequence runs from two or more independent amplicons from each strain, and we rechecked the published sequences in these regions of R6 (48) or the *cps* cluster of D39 (50). Approximately 7,000 PCR primers were generated in this project and are listed in Appendix S1. Automated and manual annotation were performed as described below. The sequence of strain D39 (NCTC) was deposited in GenBank (accession information: s_pneumoniae_d39_1 CP000410), with genes assigned the "SPD" prefix.

Purification of genomic DNA for sequencing. D39 and R6 strains were grown exponentially in static BHI broth to an optical density at 620 nm (OD_{620}) of ~0.2 at 37°C in an atmosphere of 5% CO_2. Bacteria were harvested by centrifugation (3,200 × *g*, 4°C, 10 min). Cells were washed with 1 volume of ice-cold 50 mM Tris–50 mM EDTA (pH 8.0) and then suspended in 1/10 volume of the same buffer. Triton X-100 and sodium dodecyl sulfate were added to the cell suspension at final concentrations of 0.1% (vol/vol) and 0.01% (wt/vol), respectively. Mixtures were incubated at 37°C for 30 min, followed by sequential extraction with buffered phenol, phenol-chloroform-isoamyl alcohol (25:24:1), and chloroform-isoamyl alcohol (24:1). Genomic DNA was precipitated by addition of 1/10 volume of sodium acetate (pH 5.2) and 3 volumes of 100% ethanol. DNA was collected by centrifugation (16,000 × *g*, 15 min, 4°C). DNA pellets were allowed to air dry and dissolved in TE buffer (10 mM Tris-HCl, 1 mM EDTA [pH 8.0]). Concentrations of DNA were determined by absorption at 260 nm (A_{260}).

Generation of PCR amplicons for sequencing. Primers used for PCR and sequencing were designed by using the Vector NTI program (Invitrogen). Primers were chosen based on a predicted T_m of about 58°C and lack of predicted secondary structure. Primers were synthesized by MWG Biotech, Inc., in a 96-well format. PCRs (25 μl) containing genomic DNA and high-fidelity *Pfu* polymerase (Stratagene) were performed according to the manufacturer's instructions. The purity of PCR products was evaluated by agarose gel electrophoresis in a 96-comb/well format apparatus (Owl Scientific). PCRs that yielded expected amplicons were purified by using the Wizard MagneSil purification kit (Promega) according to the manufacturer's instructions. Purified PCR amplicons were sequenced directly. A new pair of PCR primers was designed when PCRs failed to generate sufficient amplicon or yielded nonspecific products. For regions that were recalcitrant to amplification by *Pfu* polymerase, r*Tth* polymerase (Applied Biosystems) was used to generate PCR amplicons.

Sequencing PCR amplicons. Each sequencing reaction (10 μl) contained 5 μl of purified PCR amplicon, 0.4 μM primer, 1 mM $MgCl_2$, and 0.75 to 1.0 μl of BigDye terminator mixture (Applied Biosystems). For regions of predicted strong secondary structure, 5% dimethyl sulfoxide (vol/vol [final concentration]) was added to reactions to facilitate polymerase readthrough. Products of sequencing reactions were resolved by capillary electrophoresis on an ABI 3730 DNA Analyzer (Applied Biosystems) located in the Indiana Molecular Biology Institute. Sequence data were analyzed by using the CodonCode Aligner program to determine Phred scores. Regions with low Phred scores ($Q < 15$; <95% base call accuracy) were discarded and not used in alignments. The published R6 sequence (48) and the type 2 capsule biosynthetic operon (*cps*) (50) were used in the CodonCode Aligner Program to guide sequence alignments. D39 sequences containing differences from the published R6 and *cps* sequences or between the two D39 strains were sequenced multiple times to rule out possible PCR-generated errors. The corresponding regions were resequenced from R6 genomic DNA to identify errors in the published R6 sequence (Table 1).

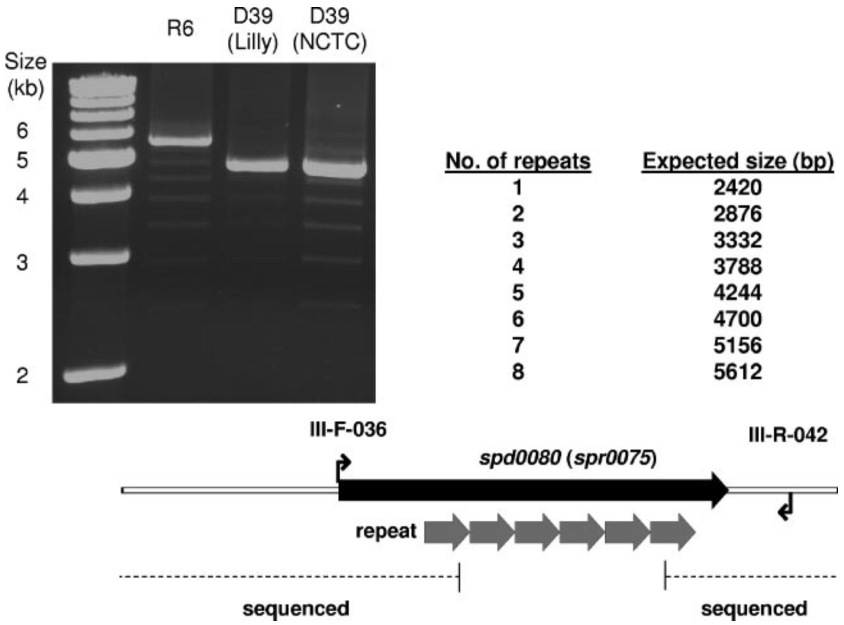
Annotation. (i) Finding genes. All candidate genes were identified by using version 3.01 of the Glimmer gene finding system (26, 99; http://www.cbcb.umd.edu/software/glimmer/). The tRNAscan-SE tool (68) was used to identify tRNAs; rRNA genes and other structural RNAs were identified directly from BLAST (4) search results.

(ii) Homology searches. The translation of each gene prediction was searched against a variety of public and private databases. Blast-Extend-Repraze (BER; http://ber.sourceforge.net) was used to search an internal nonidentical amino acid database constructed from all proteins available from GenBank (http://www.ncbi.nlm.nih.gov), UniProt (http://www.pir2.uniprot.org/), and the Comprehensive Microbial Resource database (http://www.tigr.org/CMR). A multiple alignment of each predicted protein and its BER hits was calculated by using MUSCLE (30). The Pfam (10) and TIGRFAM (41) libraries of hidden Markov

TABLE 1. Interpretation and summary of sequence changes in D39 (NCTC), D39 (Lilly), and R6 (ATCC)

Sequence change presence in:			Interpretation	Changes	
D39 (NCTC)	D39 (Lilly)	R6 (ATCC)		No.	Type
Yes[a]	Yes	No	Change arose in R6 (ATCC)	81	Seventy-one single-base-pair changes, three small insertions, one large insertion, five small deletions, and one large deletion of part of the *cps* locus reported previously (50)
Yes[b]	No	Yes	Change arose in D39 (Lilly)	8	Six single-base-pair changes, one small deletion, and one large deletion
No[b]	Yes	Yes	Change arose in D39 (NCTC)	5	Four single-base-pair changes and one small deletion
Yes[c]	Yes	Yes (resequenced)	Error in published R6 sequence	28	Fourteen in one IS*1167* element
Yes[d]	Yes	NA	Error in published D39 *cps* sequence	12	One frameshift; eleven single-base-pair changes (six in *cps2B*)

[a] Data compiled in Table 2 and Table S1 in the supplemental material.
[b] Data compiled in Table 4 and Table S2 in the supplemental material.
[c] Data compiled in Table S4 in the supplemental material.
[d] Data compiled in Table S5 in the supplemental material.



No. of repeats	Expected size (bp)
1	2420
2	2876
3	3332
4	3788
5	4244
6	4700
7	5156
8	5612

FIG. 3. PCR analysis of the repeated region in the spd0080 gene of D39 (spr0075 of R6). Primers III-F-036 and III-R-042 flanking the multiple 456-bp repeats of the SSURE domain (21) were used in PCRs to amplify the region as described in Materials and Methods. Amplification products from strains R6, D39 (Lilly), and D39 (NCTC) are shown next to a 1-kb molecular mass marker ladder (left lane) and the expected sizes of amplicons containing one to eight repeat units. The predominant species indicated repeats of eight or six SSURE units in this gene in R6 or D39, respectively. The faint ladder of bands differing by one repeat unit are likely a PCR amplification artifact and were not observed in Southern blots (see text). The extent to which this region was sequenced is indicated at the bottom of the figure. As was done previously (48), we inserted a placeholder sequence containing the five repeats reported previously for R6 into this region of the D39 genome sequence (see the text).

models (HMMs) were searched by using programs from the HMMer package (29). Other amino acid sequence signatures, domains, or functional sites were predicted by searching all proteins against the PROSITE database (31). The SignalP (13) and TMHMM (62) algorithms were used to predict putative signal sequences and membrane-spanning domains, respectively. Family prediction was also done by searching (using BLAST) all predicted proteins against version 2 of the NCBI clusters of orthologous genes (COG) database (108). Domain-based paralogous families were built on the basis of HMM hits and homologous regions (detected by using BLAST) not covered by the HMM models.

(iii) Automated annotation. A computer program developed at The Institute for Genomic Research (TIGR), AutoAnnotate, analyzed the BER and HMM search results and putatively assigned a common name, gene symbol, enzyme commission (EC) number (http://www.expasy.ch/enzyme), and TIGR and gene ontology (GO) (7) role categories in an automated fashion. AutoAnnotate uses a hierarchical approach first by evaluating the isology (sequence similarity) and score of each HMM match, followed by an evaluation of the length and the percent similarity of each BER alignment in order to predict a functional assignment to each gene. If there was a hit to an equivalog-level HMM (41) with a score above the trusted cutoff, the identifying information attached to that HMM (common name, role category, gene symbol, and EC number if applicable) was assigned to the predicted coding region. If no high scoring equivalog-level hit existed, the BER search results were evaluated. The program looked for a full-length match of at least 80% of the length of the subject, with at least 35% identity. If more than one match was found, the program attempted to choose a match with a name that followed TIGR's naming conventions and assigned a TIGR role category. If the chosen BER match was a hypothetical protein from another species or if no pairwise matches met the match criteria, AutoAnnotate went back to the HMM results and looked for nonequivalog hits. If any hits existed, the protein was assigned a family name based on the HMM name. Proteins with a pairwise match to a hypothetical protein from another species, but no HMM hit, were named a "conserved hypothetical protein." Proteins with no HMM or BER matches remained named "hypothetical protein."

(iv) Curating gene models. The results of the homology searches were manually analyzed to curate predicted initiator codons and to identify potential frameshifted genes or genes with introduced stop codons. The assembly sequence was checked against the traces to determine whether frameshift or point mutations were introduced during the closure or assembly process or whether they were authentic. Overlapping genes were manually resolved either by evaluating and editing the initiation codon for each overlap or by retaining the one overlapping gene with sequence similarity to some other protein or domain.

Genes missed by Glimmer were identified by a BLAST search against an internal nonidentical amino acid database of the six-frame translation of regions of the genome where no homology evidence for the existing gene set was present. The output was manually reviewed, and new genes were added to the data set as required.

(v) Manual annotation. All available evidence for each protein was evaluated by a human curator by using the manual annotation tool Manatee (http://manatee.sourceforge.net). Based on the evidence associated with each protein, the following descriptive information was assigned where appropriate: common name, gene symbol, EC number, TIGR role category, designation of homologous genes in the R6 (48) and TIGR4 (110) genome sequences, and gene ontology (7) terms. Specific functional annotations were assigned based on high-quality matches to experimentally characterized proteins found in the BER results or to equivalog-level HMM matches. Less specific and family level annotations were based on domain, superfamily, or subfamily HMMs or on the presence of specific motifs such as TMHMM or SignalP. In addition, the presence of genes in a putative gene cluster or pathway was important information included in each evaluation.

Determination of the number of repeats in spd0080 (spr0075). The number of direct repeats in spd0080 (spr0075) was determined by two approaches. First, different pairs of flanking primers (e.g., III-F-036 and III-R-042 in Fig. 3) were used in PCRs containing r*Tth* polymerase to amplify the repeat region. Amplicon sizes were determined by agarose gel electrophoresis relative to a 1-kb molecular weight marker ladder run in parallel (Fig. 3). Second, Southern blots were performed on genomic DNA from strains digested with NheI, which cuts outside of the region containing the direct repeats. Digested DNA was resolved by agarose gel electrophoresis and transferred to Hybond N+ membrane (GE Healthcare) by standard protocols (100). A 294-bp amplicon probe corresponding to an internal region of each repeat was synthesized by PCR using primers III-F-038 and III-R-040. Nonradioactive labeling of the amplicon probe, hybridization to the blots, and signal detection were performed according to instructions provided in the gene image AlkaPhos Direct kit (GE Healthcare).

Microarray analyses. Bacteria were grown statically in BHI broth at 37°C in an atmosphere of 5% CO_2. Overnight cultures were adjusted to an OD_{620} of ~0.1

(16-mm tubes) and diluted a further 50-fold into fresh BHI. Growth was monitored by change in OD_{620} as measured in a Spectronic 20 spectrophotometer. Cultures were harvested at an OD_{620} of ~0.1, and RNA was extracted by a hot lysis-acid phenol protocol, followed by purification using the RNAeasy minikit (QIAGEN) as described previously (96). Thirty micrograms of total RNA was used to synthesize cDNAs for each sample, followed by direct labeling using Cy3- or Cy5-dCTP (GE Healthcare). Synthesis, labeling, hybridization (16 to 18 h at 42°C), and washing protocols were performed as recommended by the microarray manufacturer (see http://www.ocimumbio.com/web/arrays/assets/downloads/manuals/manual_bacteria.pdf). *S. pneumoniae* R6 microarrays were purchased from MWG Biotech, Inc. (now available from Ocimum Biosolutions). Details of the *S. pneumoniae* oligonucleotide array are described in GEO platform entry GPL536 (http://www.ncbi.nlm.nih.gov/projects/geo/). Microarray slides were scanned on an Axon GenePix 4200A microarray scanner, and images were analyzed by using GenePix Pro 5.0 software (Molecular Devices). Microarray data were obtained from three independent biological replicates, including one dye swap. The data normalization was performed with GeneTraffic 3.2 software (Iobion Informatics) using the Lowess (subgrid) method. The data were normalized both with or without background subtraction and used to calculate expression ratios. Ratios did not differ significantly with or without background subtraction.

Expression ratios from the three replicates were averaged to obtain the average fold changes expressed in Tables S6 and S7 in the supplemental material. Bayesian *P* values were calculated by using the Cyber-T Web interface (http://visitor.ics.uci.edu/genex/cybert/) (66). The cutoff for significant changes in relative transcript amounts was set at positive or negative 1.8-fold with a Bayesian *P* value of <0.001. Intensity data and expression ratio data are deposited in the GEO database (accession no. GSE5375).

Animal models of infection. Experiments involving murine models of infection were conducted with prior approval by the Bloomington Institutional Animal Care and Use Committee and were performed in strict compliance to the *Guide for the Care and Use of Laboratory Animals*, prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Animal Resources, National Research Council (82). ICR outbred male mice (21 to 30 g; Harlan Sprague) were inoculated by intraperitoneal injection (~100 CFU in 100 μl), intratracheally (~2.5×10^6 CFU in 50 μl), or intranasally (10^5 or 10^6 CFU in 50 μl) by standard methods published previously (17, 96). Time to moribundity was monitored and recorded, after which animals were sacrificed by CO_2 asphyxiation. Death was not used as an endpoint. Survival curves were analyzed by using GraphPad Prism software and were compared by using the log-rank test. For biophotonic imaging, ICR mice were infected intratracheally with D39 (NCTC) *luxABCDE* (designated IU1912) or D39 (Lilly) *luxABCDE* (designated IU1933), which were constructed by transformation with a PCR amplicon containing the Tn*4001 luxABCDE* cassette synthesized from strain Xen 35 (38, 88). Images of bioluminescent animals were obtained essentially as described previously (87, 88) at approximately 10-h intervals.

RESULTS AND DISCUSSION

Summary of DNA sequence results. We determined and annotated the genome sequence of the two D39 isolates D39 (NCTC) and D39 (Lilly) (Fig. 1) as described in Materials and Methods. Differences between the sequences of D39 (NCTC) and D39 (Lilly) are described here. We compared the two D39 sequences with the published sequence of strain R6 (Table 1 and see Tables S1 and S2 in the supplemental material). At positions where there were differences among the three sequences, we repeated and confirmed the D39 sequence determinations, and we resequenced these regions in the R6 genome (see Materials and Methods). This analysis revealed 28 sequence errors in the previously published R6 sequence (48) (Tables 1 and S4 in the supplemental material). Fourteen of these errors were in a single IS*1167* element and likely arose as an artifact of sequence assembly. We also identified 12 differences that may be errors in the previously published sequence of the capsule (*cps*) biosynthesis region of strain D39 (Tables 1 and S5). Six of these sequence differences were clustered in the *cps2B* gene.

By comparing these three closely related sequences, we deduced the likely sequence of the D39 progenitor strain and mutations that arose separately in the D39 (NCTC) and D39 (Lilly) isolates and during the multistep isolation of laboratory strain R6 (Fig. 1). This interpretation indicated that the widely used D39 (NCTC) strain is most closely related to the D39 progenitor strain, with only five mutational differences (Table 1). D39 (Lilly) is also closely related to the progenitor strain with only eight mutational differences, including one large deletion that arose spontaneously during propagation in this laboratory. Therefore, D39 (NCTC) and D39 (Lilly) differ from each other by 13 mutations. In contrast, laboratory strain R6 contains 81 mutational differences compared to the D39 progenitor strain (Table 1). This large number of mutational differences in laboratory strain R6 was unexpected and does not support the commonly held notion that strain R6 was derived from strain D39 through a limited number of mutational events (e.g., see reference 12). These mutations and their possible impacts on physiology and pathogenesis are considered below.

How the unexpectedly large number of mutations arose in strain R6 is unknown. The predecessors of R6 (Fig. 1) were propagated repeatedly in the presence of antibody against serotype 2 capsule to select for the stable loss of capsule biosynthesis (8, 76). It is unknown whether this kind of serial propagation under stress could have led to increased mutagenesis. Alternatively, it is possible that some type of agent-induced mutagenesis was performed during the derivation of R6 from progenitor strain D39. Mutagenesis was not reported in the derivation of strain R6 (Fig. 1) (see references 8, 76, and 89). On the other hand, mutagenesis would have been entirely appropriate given the goals of those earlier experiments, which were to provide a laboratory background for transformation and genetic studies (see the introduction) (8, 76, 89). R6 remains a suitable laboratory strain for basic experiments on fundamental genetic and physiological mechanisms in pneumococcus. However, because laboratory strain R6 contains so many mutations compared to the D39 strain, it cannot be assumed that physiological and metabolic properties studied in strain R6 will extrapolate to pathogenic strain D39. Future experiments aimed at comparing the physiological, metabolic, and pathogenic properties of mutants may best be performed directly in D39 (NCTC) or in an isogenic avirulent derivative of D39 (NCTC). To this end, we have constructed stable unencapsulated mutants of strain D39 (NCTC) [strain IU1824 = D39 (NCTC) Δ*cps2A′*-*cps2H′ rpsL41* and strain IU1945 = IU1824 *rpsL*$^+$], which contain the same 7,505-bp deletion within the capsule (*cps*) biosynthetic region present in R6-related laboratory strains. IU1824 was constructed by using the Janus cassette method for allele replacement (107) and contains the *rpsL41* allele, which imparts resistance to streptomycin but does not seem to impair virulence (data not shown; see Materials and Methods). Strain IU1945 is an *rpsL*$^+$ transformant of strain IU1824.

There is another important difference between the D39 strains and laboratory strain R6. PCR analyses showed that both D39 (NCTC) and D39 (Lilly) carry the cryptic pDP1 plasmid (15, 86, 103, 105), whereas laboratory strain R6 does not (data not shown). Thus, the R6 strain has been cured of this plasmid. We sequenced the 3,161-bp pDP1 plasmids from both D39 stains and found the sequence to be identical to that reported previously (86). The cryptic pDP1 plasmid contains seven open reading frames but, aside from a replication protein, the other six putative open reading frames do not show significant homology to other genes, and their functions remain unknown (86). Although the function(s) and selection that maintain the cryptic pDP1 plasmid remain unknown, the replication apparatus of pDP1 has been used to construct shuttle vectors (94).

Besides the large 7,505-bp deletion of the *cps2A-cps2H* region, there is one other large mutational difference between the genome sequences of the D39 and R6 strains. The spd0080 gene in D39 (spr0075 in R6) encodes a surface protein containing tandem direct repeats of a 456-bp sequence (Fig. 3). The spd0080 (spr0075) gene is conserved in other pneumococcus serotypes, and four repeat units have been reported in this gene for serotype 4 strain TIGR4 (110). The repeated amino acids of this unusual protein constitute SSURE domains that bind to the extracellular matrix protein fibrinogen and may play a role in adhesion to eukaryotic host cells (21). PCR analysis using pairs of different primers flanking the repeat regions showed that both D39 strains contain six copies of the repeat (Fig. 3). In contrast, laboratory strain R6 contains eight copies of the repeat and thus has acquired two additional copies compared to the D39 strains. The eight copies in strain R6 reported here differ from the seven copies reported previously (48). Besides prominent single bands, these PCR analyses revealed a faint ladder of amplicons containing one to more than six repeats (Fig. 3). This ladder, which was used to assign the number of repeats in the prominent amplicons, is likely an amplification artifact that resulted from internal pairing of the repeats during the PCR. Consistent with this interpretation, faint amplicons containing more than six repeats were often observed for the D39 strains (Fig. 3). In addition, Southern blots of genomic DNA were consistent with the presence of six or eight copies of the repeats in D39 or R6, respectively, and did not show ladders of additional bands (data not shown). The exact sequence of the repeat region in spd0080 (spr0075) could not be determined by the primer walk approach or standard sequencing methods. As was done before (48), we inserted a placeholder sequence containing the five repeats used previously for R6 into this region of the D39 genome sequence. The rest of the genome sequence was determined completely for both D39 strains.

Major sequence differences between strains D39 and R6. The genome sequence of strain R6 has many more mutations compared to the deduced progenitor D39 strain than either of the current D39 isolates (Tables 1 and 2 and see Tables S1 and S2 in the supplemental material). Nevertheless, the synteny between the R6 and D39 progenitor strain is nearly 100% (data not shown), and the vast majority of mutational differences are single-base-pair changes, only three of which result in stop codons (Table 2 and Tables S1 and S3 in the supplemental material). This finding is significant for several reasons. Pairwise comparisons between the R6 and TIGR4 genome sequences revealed ca. 10% differences, including six regions containing insertions or deletions and five regions of different content between conserved flanking sequences (20, 104, 109). It was not entirely clear whether laboratory strain R6 had lost some of these regions during its derivation and cultivation. In fact, aside from the *cps* deletion, the loss of large pathogenic islands has not occurred in R6 or the two current D39 isolates (Tables 1 and 2 and Tables S1 and S2 in the supplemental material). This observation supports the view that the serotype 2 and 4 strains of pneumococcus have fundamental differences in their genomes that did not arise as a result of cultivation.

TABLE 2. Sequence changes that arose in strain R6 compared to progenitor strain D39[a]

Change category (no. of changes)	Description
Single-base-pair changes[b] (71)	Fifty-one missense (two following frameshift mutations); three nonsense; eight intercistronic (some correlated with changes in transcript levels); nine silent
Deletions (6)	ΔC (frameshift); ΔA (intercistronic); ΔA (frameshift); Δ8 bp of adjacent direct repeat (intercistronic); Δ41 bp (intercistronic); Δ7,505 bp that removes part of the capsule biosynthesis (*cps*) locus (50)
Insertions (4)	Insertion of two 456-bp direct repeat units (912 bp) in spr0075; inversion and insertion of 25-bp internal to *cps2A* at the junction of *cps* deletion; insertion of 30-bp adjacent direct repeat (in frame) in *pcpA*, encoding choline-binding protein; insertion of 2 bp (frameshift)

[a] Specific mutations are listed in Tables S1 and S3 in the supplemental material. R6 has also lost cryptic plasmid pDP1 (see the text).

[b] There were 45 transition mutations: 19 GC to AT (27%) and 26 AT to GC (37%). There were 26 transversion mutations: 7 GC to TA (10%); 10 AT to CG (14%); 7 GC to CG (10%); and 2 AT to TA (3%).

An interesting feature of pneumococcal genomes is the presence of numerous active and inactive IS elements and inverted sequences termed BOX or RUP elements (48, 79, 110). It has been speculated that these repeated sequences may promote genetic rearrangement and genomic plasticity. Comparison between the R6 and D39 progenitor sequence shows that rearrangements and transposition have not occurred in the approximately 70 years since strain R36 was first separated from the progenitor D39 strain (Fig. 1).

Several properties of the mutations that arose in laboratory strain R6 are noteworthy. The mutations result in changes in the amino acid sequences of numerous important proteins that mediate physiological processes, such as murein biosynthesis and transcription regulation, or virulence in the encapsulated D39 strains (Table 3). Very few of these mutated genes are linked directly to competence development (e.g., *cinA*), and there is no apparent correlation between mutations in competence genes and the increased transformability reported for R6 compared to the D39 progenitor strain (Fig. 1). The transcription patterns discussed below suggest a more complicated correlation at the level of gene expression.

If mutagenesis was used to generate the R6 strain, interpretation of the distribution of single-base-pair mutations in R6 compared to the D39 progenitor strain becomes problematic (Tables 1 and 2 and Tables S1 and S3 in the supplemental material). Nonetheless, there are two interesting mutational differences. First, there are three separate instances of deletion or insertion of adjacent repeat sequences. These range from an

TABLE 3. Virulence and physiologically important genes with altered sequences in strain D39 compared to strain R6[a]

D39 locus tag	Gene	Function	Source or reference[b]
SPD_0063	*strH*	β-*N*-Acetylhexosaminidase	STM (43)
SPD_0065	*bgaC*	β-Galactosidase	52
SPD_0080		SSURE fibronectin binding	21
SPD_0081	*rr08*	Response regulator	63, 111
SPD_0168	*ribE*	Riboflavin synthesis	71
SPD_0315-SPD_0323	Δ*cps*	Capsule biosynthesis	3, 69
SPD_0336	*pbp1A*	Penicillin-binding protein transglycosylase/transpeptidase	49, 64, 90
SPD_0479	*nusA*	Transcription termination/antitermination	18
SPD_0636	*spxB*	Pyruvate oxidase	STM/experimental (12, 65, 87, 91, 92, 106)
SPD_0660	*ftsX*	ABC transporter/cell division	102
SPD_0665	*pyrDA*	Dihydro-orotate dehydrogenase (pyrimidine biosynthesis)	5
SPD_0773	*fruA*	Fructose PTS	STM (43)
SPD_0854	*flpA*(*pavA*)	Fibronectin binding	STM/experimental (47, 64, 95)
SPD_0967	*murA1*	Homolog of MurA (first step in murein biosynthesis)	19, 70
SPD_1131	*carB*	Carbamoylphosphate synthase, heavy subunit (pyrimidine biosynthesis)	73
SPD_1337	*atpA*	Proton-translocating ATPase, F1 α subunit	32
SPD_1346		Hypothetical	STM (43)
SPD_1461	*psaB*	Manganese transport	Experimental (24, 27, 72, 75, 85)
SPD_1462	*psaC*	Manganese transport	Experimental (24, 27, 72, 75, 85)
SPD_1512	*secA*	Preprotein translocase subunit	93
SPD_1671	*amiA*	Oligopeptide transport	2, 57
SPD_1740	*cinA*	Competence induced	STM (43)
SPD_1758	*rpoC*	RNA polymerase β′ subunit	116
SPD_1797	*ccpA*	Catabolite control	STM/experimental (37, 53)
SPD_1961		Hypothetical (putative transcription regulator)	STM (43)
SPD_1965	*pcpA*	Choline-binding protein	STM/experimental (43, 101)
SPD_1974		Conserved hypothetical	STM (43)
SPD_1987		Hypothetical (fucolectin-related protein)	STM (43)
SPD_2005	*dltA*	D-Ala ligase	STM (43)
SPD_2012	*glpO*	α-Glycerophosphate oxidase	98
SPD_2022	*clpC*	Stress-related ATPase	Experimental (22, 23, 51, 96, 97)
SPD_2028	*cbpD*	Murein hydrolase	STM/experimental (, 40, 43, 46, 56)

[a] Specific mutations are listed in Tables S1 and S3 in the supplemental material.
[b] Evidence for a role in virulence exists where indicated. STM, signature-tagged mutagenesis.

8-bp deletion in an intercistronic region between *manL* and *adh* to two larger insertions in potential virulence factor genes: a 30-bp in-frame insertion in *pcpA* and insertions of two tandem 456-bp in-frame SSURE domain repeat units in spr0075 (Fig. 3). Deletion and insertion of adjacent repeat sequences such as these likely arise by RecA-independent misalignment mechanisms (reviewed in reference 67). This kind of mutation has not been reported previously for *S. pneumoniae* and may contribute to the regulation and genetic diversity in this and other bacteria with relatively small genomes (67, 91). Second, of four cases of small deletion or insertion mutations, only one occurred in a homopolymeric run of bases (see Tables S1 and S2 in the supplemental material). Spontaneous deletion and insertion in runs of bases seem to play roles in phase variation phenomena of bacterial pathogens (reviewed in reference 11) and have been reported previously in other serotypes of *S. pneumoniae* (91). However, they were not prevalent during the derivation of strain R6 from the D39 progenitor strain.

Microarray analyses of transcript amounts in strain D39 (NCTC) compared to R6 were performed for bacteria growing statically in BHI broth in an atmosphere of 5% CO_2 (see Materials and Methods). The D39 and R6 sequences differ mainly by single base substitutions (Tables 1 and 2 and Table S1 in the supplemental material). Therefore, we expected that commercially available microarrays containing 50-mer oligomers based on the R6 sequence would work well for studies of D39 transcription patterns, including transcription of the capsule biosynthesis (*cps*) genes, which comprise an operon (50) and are partly included on the R6 microarrays. This prediction is confirmed by the comparison in Fig. 4. Most transcript levels were comparable in the two strains grown under these conditions (Fig. 4 and Table S6 in the supplemental material). However, there are some notable differences. As expected, deletion of part of the capsule (*cps*) region in R6 was reflected in the microarray comparison (Fig. 4). Unexpectedly, mutations in several intercistronic regions in R6 seemed to affect the transcript amounts of adjacent gene clusters, suggesting the cotranscription of genes from possible promoters or regulatory sites in these intercistronic regions (Fig. 4). In addition, the relative transcript amounts of eight competence regulon genes increased in R6 compared to D39 (Fig. 4; see also Table S6 in the supplemental material). However, none of these genes contained mutations in R6 compared to D39, and the mechanism of their induction remains unknown. The increased expression of this set of genes could, in part, contribute to the reported increase in transformability of R6 compared to the D39 progenitor (Fig. 1).

Major sequence differences between two isolates of strain D39. In contrast to strain R6, there are relatively few mutational changes between D39 (NCTC) and D39 (Lilly) and the deduced D39 progenitor strain (Tables 1 and 4). D39 (NCTC) and D39 (Lilly) contain only five and eight mutational differ-

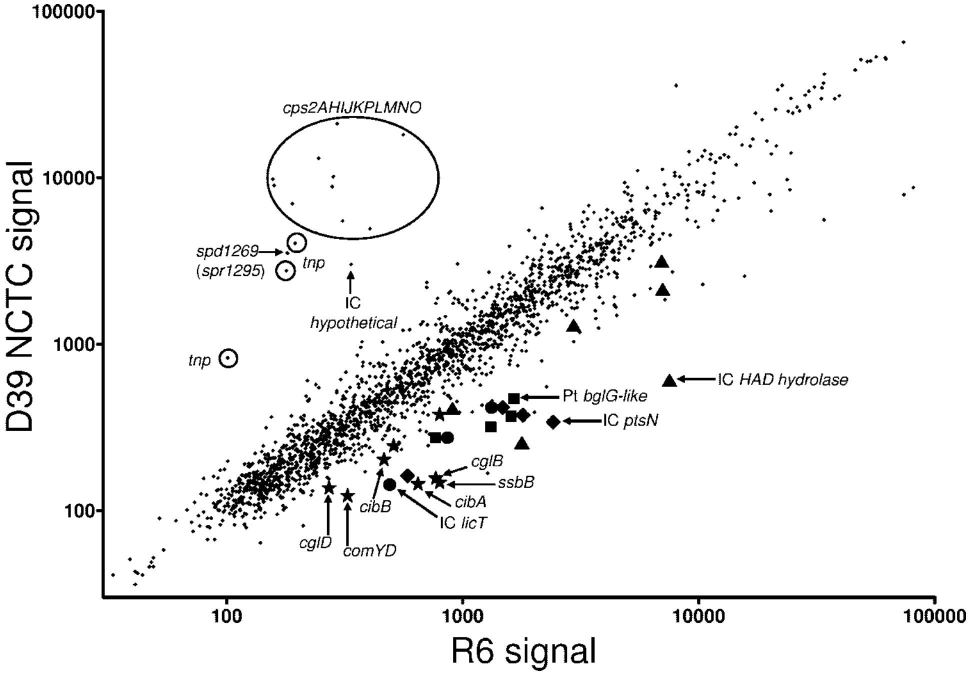


FIG. 4. Microarray analysis of relative transcript amounts in strains D39 (NCTC) and R6 grown exponentially in BHI. Microarray analyses were performed as described in Materials and Methods. A representative log-scale scatter plot of relative transcript amounts is shown, and fold changes and Bayesian *P* values for transcripts changing at least 1.9 are listed in Table S6 in the supplemental material. Genes comprising the capsule locus are circled. Symbols indicate coregulated gene clusters as follows: stars, competence-induced genes (*cibBC* [allolysis], single-strand binding protein [*ssb*], and *cglABCD comYD* [peptidoglycan-spanning structural proteins]); circles (spd0501 to spd0503; spr0504 to spr0506), genes coregulated with *licT* (spd0501; spr0504); diamonds (spd0559 to spd0562; spr0562 to spr0565), genes coregulated with *ptsN* (spd0559; spr0562); squares (spd1958 to spd1961; spr1937 to spr1940), genes coregulated with a BglG-family transcriptional regulator (spd1961; spr1940); triangles (spd1029 to spd1034; spr1052 to spr1057), genes coregulated with an HAD superfamily hydrolase (spd1034; spr1057). IC, mutation in upstream intercistronic region; Pt, point mutation within gene coding sequence. See the text for additional details.

ences, respectively, compared to the D39 progenitor and thirteen mutational differences between themselves. These mutations almost certainly arose spontaneously and reveal the pattern of mutations that arose during repeated culturing of these serotype 2 strains. D39 (NCTC) shows relatively minor changes compared to the D39 progenitor (Tables 1 and 4 and Table S2 in the supplemental material). The single-base-pair change in the IC (spd1248/spd1249 intercistronic region [corresponds to IC spr1274/spr1275 in R6] does not influence the transcript amounts of adjacent genes (see Table S7 in the supplemental material). The lone frameshift mutation is in a hypothetical gene (spd0800; corresponds to R6 spr0806) and results in a carboxyl-terminal truncation of only three amino acids (see Table S2 in the supplemental material). Of the three missense mutations, two are in hypothetical genes (spd1137 [corresponds to R6 spr1161] and spd1751 [corresponds to R6 spr1769]), and one is in a putative regulator of capsule biosynthesis (*cpsY*) (spd0818 [corresponds to R6 spr0828]) (61). spd1751 (corresponds to R6 spr1769) and *cpsY* (spd0818) were identified as putative virulence factors in previous STM screens (Table 5) (43). The microarray analyses described below did not detect a relative change in *cps* transcript amounts between D39 (NCTC) and D39 (Lilly), which contains the *cpsY* progenitor sequence. Moreover, D39 (NCTC) remains extremely virulent in murine models of infection using intranasal or intraperitoneal inoculation (see below and Materials and Methods), implying that the missense mutations that arose in spd1751 and *cpsY* in D39 (NCTC) do not strongly affect virulence. Thus, overall, the effects of the mutational changes in D39 (NCTC) seem minimal, and D39 (NCTC) is the serotype 2 isolate of choice for future physiological, genetic, and virulence studies.

TABLE 4. Sequence changes that arose in strain D39 (NCTC) or D39 (Lilly) compared to progenitor strain D39[a]

Strain	Changes
D39 (NCTC)	ΔG (frameshift); three missense mutations[b]; one single base change in IC region[b]
D39 (Lilly)	Five missense mutations[c]; one silent change[c]; one ΔT (frameshift); one Δ*treP-amiC* deletion (7,958 bp) between 14-bp direct repeats (AGTCGCTGTATAAG)

[a] Specific mutations are listed in Table S2 in the supplemental material.

[b] There were three transition mutations: one GC to AT (25%) and two AT to GC (50%). There was one transversion mutation: 1 GC to CG (25%).

[c] There were four transition mutations: two GC to AT (33%) and two AT to GC (33%). There were two transversion mutations: one AT to CG (17%) and one GC to CG (17%).

In contrast, the eight unique mutations in strain D39 (Lilly) compared to the D39 progenitor strain are not as benign as those in strain D39 (NCTC) (Tables 4 and 5 and Table S2 in the supplemental material). Several of these mutational differences are in genes important for virulence identified in previous STM screens in murine models of infection, including genes encoding a putative copper transporter CtpA (Spd0635),

TABLE 5. Virulence and physiologically important genes with altered sequences in strain D39 (NCTC) compared to strain D39 (Lilly)[a]

D39 locus tag	Gene	Function	Source or reference[e]
SPD_0133[b]	*cibB*	Competence-induced bacteriocin (allolysis)	spr0128 (48; see reference 40)
SPD_0157[b]	*hk07*	Histidine kinase	63, 111
SPD_0383[b]	*fabD*	Malonyl coenzyme A-acyl carrier protein transacylase	115
SPD_0635[b]	*ctpA*	Cation transport	STM (43)
SPD_0636[b]	*spxB*	Pyruvate oxidase	STM/experimental (12, 65, 87, 91, 92, 106)
SPD_0800[c]		Hypothetical	spr0806 (48)
SPD_0818[c]	*cpsY*	Capsule regulation	43, 61
SPD_1137[c]		Hypothetical (ABC transporter)	spr1161 in (48)
SPD_1428[b]	*cmk*	Cytidylate kinase	118
SPD_1669[b,d] (Δ*treP-amiC*)	Δ*amiD*	Oligopeptide transport	STM (43)
SPD_1670[b,d] (Δ*treP-amiC*)	Δ*amiC*	Oligopeptide transport	STM (43)
SPD_1751[c]		Hypothetical (putative membrane protein)	STM (43)

[a] Specific mutations are listed in Table S2 in the supplemental material.
[b] Arose in strain D39 (Lilly) compared to D39 progenitor strain.
[c] Arose in strain D39 (NCTC) compared to D39 progenitor strain.
[d] Arose spontaneously during propagation of D39 (Lilly) in this laboratory.
[e] Evidence for a role in virulence exists where indicated. STM, signature-tagged mutagenesis.

pyruvate oxidase SpxB (Spd0636), and oligopeptide transporters AmiC (Spd1670) and AmiD (Spd1669) (Table 5 and Table S2 in the supplemental material). Other sequence differences are in key regulatory and metabolic genes, such as the genes encoding competence-induced bacteriocin CibB involved in allolysis (Spd0133), histidine kinase HK07 (Spd0157), malonyl acyl carrier protein transacylase FabD (Spd0383), and cytidylate kinase Cmk (Spd1428) (Table 5).

Because of these mutational differences, we tested whether D39 (Lilly) was attenuated for virulence compared to D39 (NCTC). Both strains are virulent in an ICR mouse model of infection (see Materials and Methods) (Fig. 5). Intraperitoneal infection of either strain resulted in severe moribundity of mice within 1 day (data not shown). Survival curves after intranasal inoculation with either strain were statistically indistinguishable over 6 days (data not shown). However, survival curves after intratracheal inoculation suggested that D39 (Lilly) is partially attenuated for virulence compared to D39 (NCTC) (data not shown). Monitoring of infection by biophotonic imaging (see Materials and Methods) confirmed this conclusion (Fig. 5). In a typical experiment, comparable infections were confined to the lungs for both strains about 19 h after inoculation (Fig. 5A). By 85 h after inoculation, the D39 (NCTC) strain had spread from the lungs into the bloodstream in most mice and was causing fatal, systemic infection (Fig. 5B, left) (see references 87 and 88). In contrast, the D39 (Lilly) strain often remained confined to the lungs and, in some cases, appeared to be cleared from the lungs (Fig. 5B, right).

The compilation of the kinds of mutations that arose spontaneously in the two D39 isolates is informative (Table 4). Most of the single base changes are transition mutations, which are nearly evenly distributed between GC→AT and AT→GC changes (Table 4). Both D39 (NCTC) and D39 (Lilly) produce significant amounts of hydrogen peroxide, as reported previously for other pneumococcal strains (12, 92), although D39 (Lilly) produces ~3-fold less than D39 (NCTC) (S. Ramos-Montanez, unpublished results). Despite this production of hydrogen peroxide, GC→AT mutations, which are indicative of oxidative damage, do not predominate as might have been anticipated from an earlier study of spontaneous mutations that arise in different pneumococcal strains (91). The only large spontaneous deletion found in strain D39 (Lilly) arose between two nontandem 14-bp direct repeats in the *treP-amiC* region of the chromosome (Table 4 and Table S2 in the supplemental material). PCR analyses of different stocks and derivatives of D39 (Lilly) showed that this deletion arose spontaneously in this laboratory during routine propagation (single colony isolation on TSAII BA, followed by liquid culture in BHI) without obvious selection to replenish frozen stock vials. Deletions between such short nontandem direct repeats have been reported before in pneumococcus (91) and likely arise by a RecA-independent mechanism involving DNA misalignment during replication (67). Finally, the single spontaneous frameshift mutation in D39 (Lilly) again did not occur in a homopolymeric run of bases (Table 4 and Table S2 in the supplemental material) and causes a severe truncation of CtpA, which has been implicated in copper uptake and virulence (34, 43).

The 13 mutational differences between D39 (NCTC) and D39 (Lilly) caused marked changes in relative transcript amounts when the strains were grown exponentially in BHI (Fig. 6 and Table S7 in the supplemental material). In fact, there are about as many transcripts with changed relative amounts in D39 (NCTC) compared to D39 (Lilly) as there are in D39 (NCTC) compared to R6 (compare Fig. 4 and 6). Three regulatory patterns can be discerned in the comparison between transcript amounts in D39 (NCTC) and D39 (Lilly). The deletion in D39 (Lilly) of the *amiCDEF* operon (Fig. 6), which encodes an oligopeptide transporter (Table 5), may be linked to the increase in expression of the *ilvBNC* operon in D39 (Lilly) compared to D39 (NCTC) (Fig. 6 and Table S7 in the supplemental material). The *ilvBNC* operon, which mediates the biosynthesis of isoleucine and valine from pyruvate (83), is likely a member of the CodY regulon in *S. pneumoniae* (39) and may also be regulated by attenuation (77; http://cmgm.stanford.edu/~merino/streptococcus_pneumoniae_R6/indice_alpha.html). In D39 (NCTC) growing in BHI, the intact AmiABCDEF transporter likely increases the intracellular concentration of isoleucine compared to D39 (Lilly). Branched -chain amino acids are the corepressors of CodY that presumably

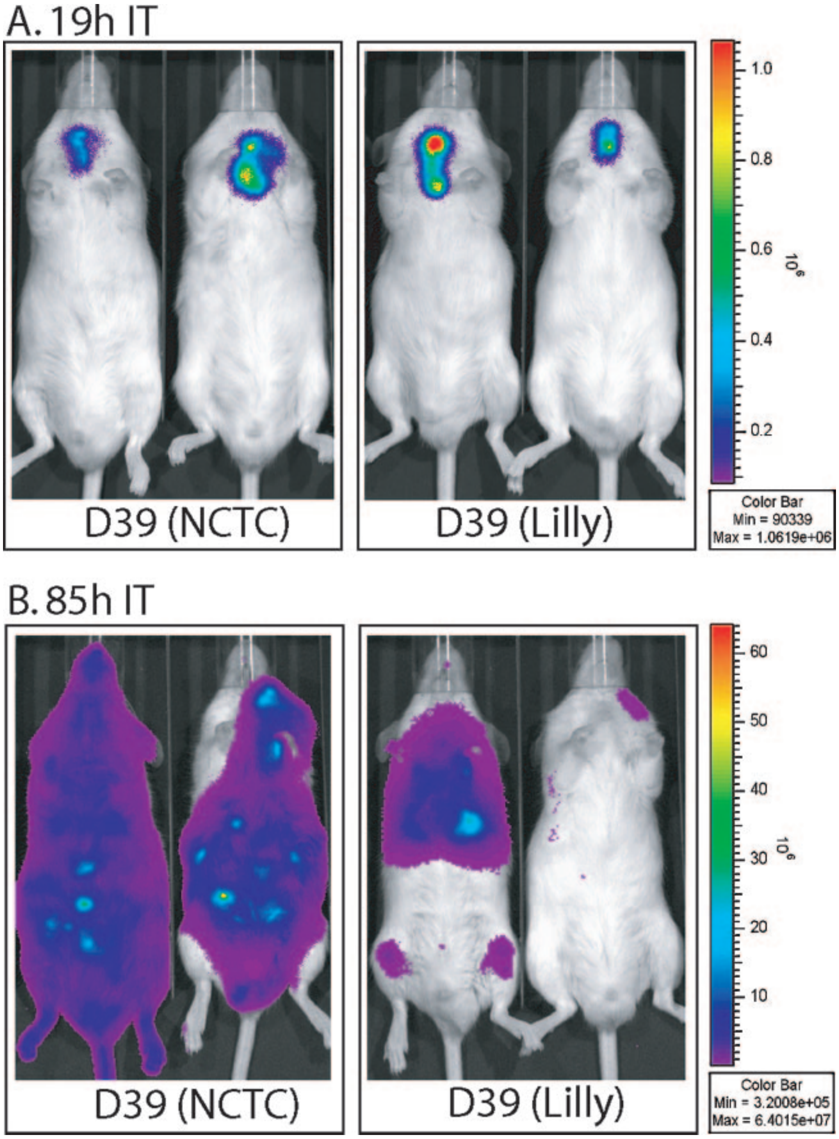


FIG. 5. Biophotonic imaging of ICR male mice (ca. 25 to 30 g) infected with D39 (NCTC) *luxABCDE* or D39 (Lilly) *luxABCDE*. Each mouse was infected intratracheally (IT) with 50 μl of buffered saline containing ~2.5 × 10^6 CFU of each strain (see Materials and Methods) (96). Bioluminescence was monitored and recorded at various times in separate animals (87, 88). Typical images from 1-min exposures are shown 19 h (A) and 85 h (B) after infection. The mouse in the lower right corner has cleared the D39 (Lilly) infection.

lead to repression of *ilvBNC* transcription in D39 (NCTC) relative to D39 (Lilly).

A second transcription pattern is the increased expression of the CiaRH two-component regulatory system in D39 (NCTC) compared to D39 (Lilly) (Fig. 6). We do not know the basis for this increase in CiaRH transcript amounts and whether it is somehow linked to deletion of *amiCDEF* in D39 (Lilly). Nevertheless, the increased transcription of *ciaRH* likely leads to a "Cia-on" state (74), the hallmarks of which include relative increases or decreases in the transcript amounts of *malPM*, *axe1*, spd0913 (corresponds to R6 spr0931), and *htrA* or *comCD*, respectively, in D39 (NCTC) compared to D39 (Lilly) (Fig. 6). Finally, we observed increased relative transcript amounts of the *malXCD* operon in D39 (NCTC) compared to D39 (Lilly) (Fig. 6). We do not know the basis of this increase, but it could be related to the increased expression of *malPM* in the "Cia-on" state in D39 (NCTC).

Further insights into pneumococcal physiology, metabolism, and virulence. As noted above, the overall genome structure appears to have been remarkably stable for virulent serotype 2 D39 strains, which have been maintained in captivity for nearly a century. Although D39 and R6 have the capacity for genetic plasticity (48, 110), major deletions and rearrangements have not occurred in these strains during cultivation, and the D39 strains have maintained their extreme virulence in murine models of infection (Fig. 5) (see Materials and Methods) (96). The D39 genome sequences reported here confirm several previous conclusions that were tentatively based on the genome sequence of laboratory strain R6 (48, 110). Notably, major differences in the genetic complements of D39 and

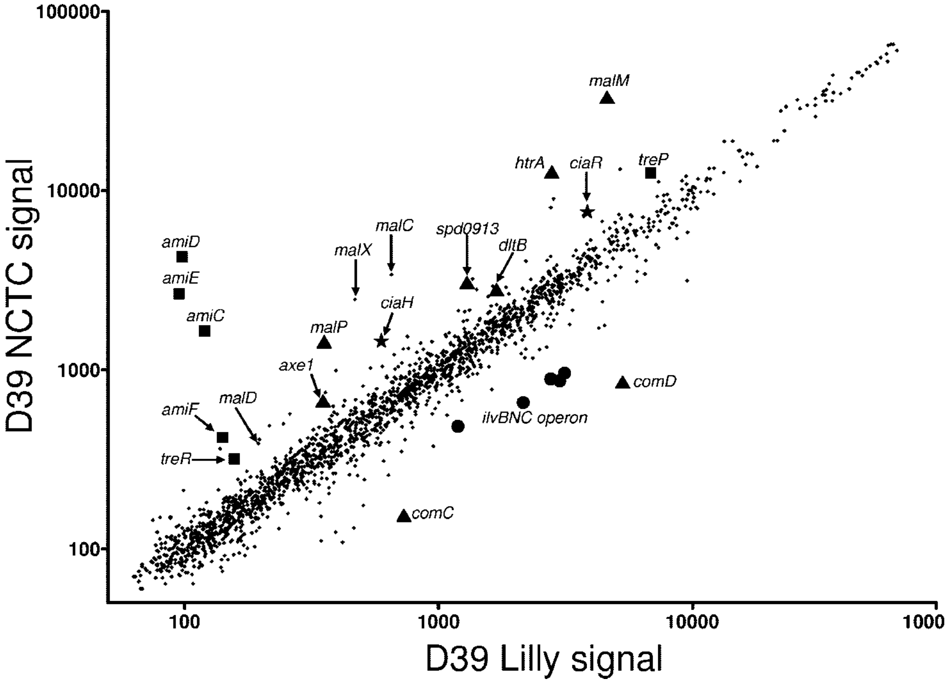


FIG. 6. Microarray analysis of relative transcript amounts in strains D39 (NCTC) and D39 Lilly) grown exponentially in BHI. Microarray analyses were performed as described in Materials and Methods. A representative log-scale scatter plot of relative transcript amounts is shown, and the fold changes and Bayesian *P* values for transcripts changing at least 1.8-fold are listed in Table S7 in the supplemental material. Symbols are as follows: squares (spd1664 to spd1670), genes partially or completely deleted in D39 (Lilly); circles (spd0404 to spd0408), *ilvBNC* gene cluster likely regulated to compensate for deletion of *amiCDEF*; stars, *ciaRH* two-component signal transduction system; triangles, genes regulated in a "*cia*-on" mutant (74).

TIGR4 were confirmed, such as the presence of a single sortase gene in D39 compared to four sortase genes in TIGR4 (48, 109, 110). Indeed, D39 and TIGR4 contain significant genetic differences that contribute to the different diseases that they cause in infection models (see the introduction). An unusual process confirmed by the D39 sequences is the devolution of amino acid biosynthetic and transposase genes in serotype 2 pneumococcus. The D39 and R6 sequences contain four truncated amino acid biosynthetic genes, *serB*, *metY*, *leuD*, and *leuA*, which mediate serine, methionine, and leucine biosynthesis. Other members of these key pathways, such as *serA*, are already missing from the genome, and it appears that the remaining genes in these pathways exist as inactive remnants. Likewise, there are many (>30) inactive transposase gene remnants in the D39 chromosome. In summary, the genome sequence of strain D39 should significantly increase our understanding of pneumococcal physiology, pathogenesis, and evolution; help in the interpretations of previously performed experiments; and allow the design of future studies, including swapping pathogenic islands between different pneumococcal serotype strains.

ACKNOWLEDGMENTS

We thank Lawrence Washington and Rudy Raff for access to sequencing facilities at the Indiana Institute for Molecular Biology and John Colbourne and Peter Cherbas for the use of the microarray facilities in the Center for Genomics and Bioinformatics at Indiana University Bloomington. We thank the following undergraduate students who contributed confirmatory data to this project: Talin Robinson, Michael Foster, and Rebecca Lohmeyer. We thank Lilly Research Laboratories for providing strain D39 (Lilly) that was originally obtained from Janet Yother (UAB), and we thank Tim Mitchell (Glasgow), Janet Yother (UAB), and Susan Hollingshead (UAB) for helpful comments about the project. We also thank Francesco Iannelli (LAMMB, Siena, Italy) for information about the corrected sequence of the *cps* region.

This study was supported primarily by an NSF/USDA-CSREES Microbial Sequencing Award 0412141 to M.E.W. This project was also supported by funds from grant AI060744-01 from the National Institutes of Health and funds from the Indiana METACyt Initiative of Indiana University Bloomington, funded in part through a major grant from the Lilly Endowment, Inc., to M.E.W.

ADDENDUM IN PROOF

The difference in the sequence of the *dltA* gene in strains D39 and R6 (Table 3) was recently independently confirmed by M. Kovacs, A. Halfmann, J. Fedtke, M. Heintz, A. Peschel, W. Vollmer, R. Hakenbeck, and R. Bruckner (J. Bacteriol. **188:**5797–5805, 2006). They show that the mutation in strain R6 inactivates *dltA* (D-alanine ligase), resulting in sensitivity to cationic antimicrobial peptides. The 12 differences in the sequence of the *cps* (capsule biosynthesis) region reported in Tables 1 and S5 have been confirmed as sequencing errors in a previous paper (50; F. Iannelli, personal communication). Corresponding corrections have been made by F. Iannelli to GenBank file AF026471.

REFERENCES

1. **Alloing, G., B. Martin, C. Granadel, and J. P. Claverys.** 1998. Development of competence in *Streptococcus pneumoniae*: pheromone autoinduction and control of quorum sensing by the oligopeptide permease. Mol. Microbiol. **29:**75–83.
2. **Alloing, G., M. C. Trombe, and J. P. Claverys.** 1990. The *ami* locus of the gram-positive bacterium *Streptococcus pneumoniae* is similar to binding protein-dependent transport operons of gram-negative bacteria. Mol. Microbiol. **4:**633–644.
3. **Alonso De Velasco, E., A. F. Verheul, J. Verhoef, and H. Snippe.** 1995.

Streptococcus pneumoniae: virulence factors, pathogenesis, and vaccines. Microbiol. Rev. **59:**591–603.

4. **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.
5. **Andersen, P. S., P. J. Jansen, and K. Hammer.** 1994. Two different dihydroorotate dehydrogenases in *Lactococcus lactis*. J. Bacteriol. **176:**3975–3982.
6. **Appelbaum, P. C.** 2002. Resistance among *Streptococcus pneumoniae*: implications for drug selection. Clin. Infect. Dis. **34:**1613–1620.
7. **Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, et al.** 2000. Gene ontology: tool for the unification of biology. Nat. Genet. **25:**25–29.
8. **Avery, O. T., C. M. MacLeod, and M. McCarty.** 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J. Exp. Med. **79:**137–158.
9. **Barry, J. M.** 2005. The great influenza: the epic story of the deadliest plague in history. Penguin Books, New York, NY.
10. **Bateman, A., E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. Sonnhammer.** 2000. The Pfam protein families database. Nucleic Acids Res. **28:**263–266.
11. **Bayliss, C. D., D. Field, and E. R. Moxon.** 2001. The simple sequence contingency loci of *Haemophilus influenzae* and *Neisseria meningitidis*. J. Clin. Investig. **107:**657–662.
12. **Belanger, A. E., M. J. Clague, J. I. Glass, and D. J. Leblanc.** 2004. Pyruvate oxidase is a determinant of Avery's rough morphology. J. Bacteriol. **186:** 8164–8171.
13. **Bendtsen, J. D., H. Nielsen, G. von Heijne, and S. Brunak.** 2004. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. **340:**783–795.
14. **Benton, K. A., M. P. Everson, and D. E. Briles.** 1995. A pneumolysin-negative mutant of *Streptococcus pneumoniae* causes chronic bacteremia rather than acute sepsis in mice. Infect. Immun. **63:**448–455.
15. **Berry, A. M., E. M. Glare, D. Hansman, and J. C. Paton.** 1989. Presence of a small plasmid in clinical isolates of *Streptococcus pneumoniae*. FEMS Microbiol. Lett. **53:**275–278.
16. **Blue, C. E., and T. J. Mitchell.** 2003. Contribution of a response regulator to the virulence of *Streptococcus pneumoniae* is strain dependent. Infect. Immun. **71:**4405–4413.
17. **Blue, C. E., G. K. Paterson, A. R. Kerr, M. Berge, J. P. Claverys, and T. J. Mitchell.** 2003. ZmpB, a novel virulence factor of *Streptococcus pneumoniae* that induces tumor necrosis factor alpha production in the respiratory tract. Infect. Immun. **71:**4925–4935.
18. **Borukhov, S., J. Lee, and O. Laptenko.** 2005. Bacterial transcription elongation factors: new insights into molecular mechanism of action. Mol. Microbiol. **55:**1315–1324.
19. **Brown, E. D., E. I. Vivas, C. T. Walsh, and R. Kolter.** 1995. MurA (MurZ), the enzyme that catalyzes the first committed step in peptidoglycan biosynthesis, is essential in *Escherichia coli*. J. Bacteriol. **177:**4194–4197.
20. **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.
21. **Bumbaca, D., J. E. Littlejohn, H. Nayakanti, D. J. Rigden, M. Y. Galperin, and M. J. Jedrzejas.** 2004. Sequence analysis and characterization of a novel fibronectin-binding repeat domain from the surface of *Streptococcus pneumoniae*. Omics **8:**341–356.
22. **Charpentier, E., R. Novak, and E. Tuomanen.** 2000. Regulation of growth inhibition at high temperature, autolysis, transformation, and adherence in *Streptococcus pneumoniae* by *clpC*. Mol. Microbiol. **37:**717–726.
23. **Chastanet, A., M. Prudhomme, J. P. Claverys, and T. Msadek.** 2001. Regulation of *Streptococcus pneumoniae clp* genes and their role in competence development and stress survival. J. Bacteriol. **183:**7295–7307.
24. **Claverys, J. P., C. Granadel, A. M. Berry, and J. C. Paton.** 1999. Penicillin tolerance in *Streptococcus pneumoniae*, autolysis, and the Psa ATP-binding cassette (ABC) manganese permease. Mol. Microbiol. **32:**881–883.
25. **De Las Rivas, B., J. L. Garcia, R. Lopez, and P. Garcia.** 2002. Purification and polar localization of pneumococcal LytB, a putative endo-β-*N*-acetylglucosaminidase: the chain-dispersing murein hydrolase. J. Bacteriol. **184:** 4988–5000.
26. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27:**4636–4641.
27. **Dintilhac, A., G. Alloing, C. Granadel, and J. P. Claverys.** 1997. Competence and virulence of *Streptococcus pneumoniae*: Adc and PsaA mutants exhibit a requirement for Zn and Mn resulting from inactivation of putative ABC metal permeases. Mol. Microbiol. **25:**727–739.
28. **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.
29. **Eddy, S. R.** 1998. Profile hidden Markov models. Bioinformatics **14:**755–763.
30. **Edgar, R. C.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:**1792–1797.
31. **Falquet, L., M. Pagni, P. Bucher, N. Hulo, C. J. Sigrist, K. Hofmann, and A. Bairoch.** 2002. The PROSITE database, its status in 2002. Nucleic Acids Res. **30:**235–238.
32. **Fenoll, A., R. Munoz, E. Garcia, and A. G. de la Campa.** 1994. Molecular basis of the optochin-sensitive phenotype of pneumococcus: characterization of the genes encoding the F0 complex of the *Streptococcus pneumoniae* and *Streptococcus oralis* H^+-ATPases. Mol. Microbiol. **12:**587–598.
33. **Francis, K. P., J. Yu, C. Bellinger-Kawahara, D. Joh, M. J. Hawkinson, G. Xiao, T. F. Purchio, M. G. Caparon, M. Lipsitch, and P. R. Contag.** 2001. Visualizing pneumococcal infections in the lungs of live mice using bioluminescent *Streptococcus pneumoniae* transformed with a novel gram-positive *lux* transposon. Infect. Immun. **69:**3350–3358.
34. **Francis, M. S., and C. J. Thomas.** 1997. The *Listeria monocytogenes* gene *ctpA* encodes a putative P-type ATPase involved in copper transport. Mol. Gen. Genet. **253:**484–491.
35. **Garcia, E., D. Llull, R. Munoz, M. Mollerach, and R. Lopez.** 2000. Current trends in capsular polysaccharide biosynthesis of *Streptococcus pneumoniae*. Res. Microbiol. **151:**429–435.
36. **Gerhardt, P., R. G. E. Murray, W. A. Wood, and N. R. Krieg.** 1994. Methods for general and molecular bacteriology. ASM Press, Washington, DC.
37. **Giammarinaro, P., and J. C. Paton.** 2002. Role of RegM, a homologue of the catabolite repressor protein CcpA, in the virulence of *Streptococcus pneumoniae*. Infect. Immun. **70:**5454–5461.
38. **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.
39. **Guedon, E., B. Sperandio, N. Pons, S. D. Ehrlich, and P. Renault.** 2005. Overall control of nitrogen metabolism in *Lactococcus lactis* by CodY, and possible models for CodY regulation in firmicutes. Microbiology **151:**3895–3909.
40. **Guiral, S., T. J. Mitchell, B. Martin, and J. P. Claverys.** 2005. Competence-programmed predation of noncompetent cells in the human pathogen *Streptococcus pneumoniae*: genetic requirements. Proc. Natl. Acad. Sci. USA **102:**8710–8715.
41. **Haft, D. H., B. J. Loftus, D. L. Richardson, F. Yang, J. A. Eisen, I. T. Paulsen, and O. White.** 2001. TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res. **29:**41–43.
42. **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.
43. **Hava, D. L., and A. Camilli.** 2002. Large-scale identification of serotype 4 *Streptococcus pneumoniae* virulence factors. Mol. Microbiol. **45:**1389–1406.
44. **Hava, D. L., J. LeMieux, and A. Camilli.** 2003. From nose to lung: the regulation behind *Streptococcus pneumoniae* virulence factors. Mol. Microbiol. **50:**1103–1110.
45. **Havarstein, L. S., G. Coomaraswamy, and D. A. Morrison.** 1995. An unmodified heptadecapeptide pheromone induces competence for genetic transformation in *Streptococcus pneumoniae*. Proc. Natl. Acad. Sci. USA **92:**11140–11144.
46. **Havarstein, L. S., B. Martin, O. Johnsborg, C. Granadel, and J. P. Claverys.** 2006. New insights into the pneumococcal fratricide: relationship to clumping and identification of a novel immunity factor. Mol. Microbiol. **59:**1297–1307.
47. **Holmes, A. R., R. McNab, K. W. Millsap, M. Rohde, S. Hammerschmidt, J. L. Mawdsley, and H. F. Jenkinson.** 2001. The *pavA* gene of *Streptococcus pneumoniae* encodes a fibronectin-binding protein that is essential for virulence. Mol. Microbiol. **41:**1395–1408.
48. **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.
49. **Hoskins, J., P. Matsushima, D. L. Mullen, J. Tang, G. Zhao, T. I. Meier, T. I. Nicas, and S. R. Jaskunas.** 1999. Gene disruption studies of penicillin-binding proteins 1a, 1b, and 2a in *Streptococcus pneumoniae*. J. Bacteriol. **181:**6552–6555.
50. **Iannelli, F., B. J. Pearce, and G. Pozzi.** 1999. The type 2 capsule locus of *Streptococcus pneumoniae*. J. Bacteriol. **181:**2652–2654.
51. **Ibrahim, Y. M., A. R. Kerr, N. A. Silva, and T. J. Mitchell.** 2005. Contribution of the ATP-dependent protease ClpCP to the autolysis and virulence of *Streptococcus pneumoniae*. Infect. Immun. **73:**730–740.
52. **Ito, Y., and T. Sasaki.** 1997. Cloning and characterization of the gene

encoding a novel beta-galactosidase from *Bacillus circulans*. Biosci. Biotechnol. Biochem. **61:**1270–1276.

53. **Iyer, R., N. S. Baliga, and A. Camilli.** 2005. Catabolite control protein A (CcpA) contributes to virulence and regulation of sugar metabolism in *Streptococcus pneumoniae*. J. Bacteriol. **187:**8340–8349.

54. **Janoff, E. N., and J. B. Rubins.** 2004. Immunodeficiency and invasive pneumococcal disease, p. 252–280. *In* E. I. Tuomanen, T. J. Mitchell, D. A. Morrison, and B. G. Spratt (ed.), The pneumococcus. ASM Press, Washington, DC.

55. **Janoff, E. N., and J. B. Rubins.** 1997. Invasive pneumococcal disease in the immunocompromised host. Microb. Drug Resist. **3:**215–232.

56. **Kausmally, L., O. Johnsborg, M. Lunde, E. Knutsen, and L. S. Havarstein.** 2005. Choline-binding protein D (CbpD) in *Streptococcus pneumoniae* is essential for competence-induced cell lysis. J. Bacteriol. **187:**4338–4345.

57. **Kerr, A. R., P. V. Adrian, S. Estevao, R. de Groot, G. Alloing, J. P. Claverys, T. J. Mitchell, and P. W. Hermans.** 2004. The Ami-AliA/AliB permease of *Streptococcus pneumoniae* is involved in nasopharyngeal colonization but not in invasive disease. Infect. Immun. **72:**3902–3906.

58. **Kharat, A. S., and A. Tomasz.** 2003. Inactivation of the *srtA* gene affects localization of surface proteins and decreases adhesion of *Streptococcus pneumoniae* to human pharyngeal cells in vitro. Infect. Immun. **71:**2758–2765.

59. **Klein, D. L.** 1999. Pneumococcal disease and the role of conjugate vaccines. Microb. Drug Resist. **5:**147–157.

60. **Klugman, K. P.** 2004. Clinical relevance of antibiotic resistance in pneumococcal infections, p. 331–338. *In* E. I. Tuomanen, T. J. Mitchell, D. A. Morrison, and B. G. Spratt (ed.), The pneumococcus. ASM Press, Washington, DC.

61. **Koskiniemi, S., M. Sellin, and M. Norgren.** 1998. Identification of two genes, *cpsX* and *cpsY*, with putative regulatory function on capsule expression in group B streptococci. FEMS Immunol. Med. Microbiol. **21:**159–168.

62. **Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer.** 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. **305:**567–580.

63. **Lange, R., C. Wagner, A. de Saizieu, N. Flint, J. Molnos, M. Stieger, P. Caspers, M. Kamber, W. Keck, and K. E. Amrein.** 1999. Domain organization and molecular characterization of 13 two-component systems identified by genome sequencing of *Streptococcus pneumoniae*. Gene **237:**223–234.

64. **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.

65. **LeMessurier, K. S., A. D. Ogunniyi, and J. C. Paton.** 2006. Differential expression of key pneumococcal virulence genes in vivo. Microbiology **152:** 305–311.

66. **Long, A. D., H. J. Mangalam, B. Y. Chan, L. Tolleri, G. W. Hatfield, and P. Baldi.** 2001. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in *Escherichia coli* K-12. J. Biol. Chem. **276:**19937–19944.

67. **Lovett, S. T.** 2004. Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. Mol. Microbiol. **52:** 1243–1253.

68. **Lowe, T. M., and S. R. Eddy.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:**955–964.

69. **Magee, A. D., and J. Yother.** 2001. Requirement for capsule in colonization by *Streptococcus pneumoniae*. Infect. Immun. **69:**3755–3761.

70. **Marquardt, J. L., D. A. Siegele, R. Kolter, and C. T. Walsh.** 1992. Cloning and sequencing of *Escherichia coli murZ* and purification of its product, a UDP-*N*-acetylglucosamine enolpyruvyl transferase. J. Bacteriol. **174:**5748–5752.

71. **Marra, A., J. Asundi, M. Bartilson, S. Lawson, F. Fang, J. Christine, C. Wiesner, D. Brigham, W. P. Schneider, and A. E. Hromockyj.** 2002. Differential fluorescence induction analysis of *Streptococcus pneumoniae* identifies genes involved in pathogenesis. Infect. Immun. **70:**1422–1433.

72. **Marra, A., S. Lawson, J. S. Asundi, D. Brigham, and A. E. Hromockyj.** 2002. In vivo characterization of the *psa* genes from *Streptococcus pneumoniae* in multiple models of infection. Microbiology **148:**1483–1491.

73. **Martinussen, J., and K. Hammer.** 1998. The *carB* gene encoding the large subunit of carbamoylphosphate synthetase from *Lactococcus lactis* is transcribed monocistronically. J. Bacteriol. **180:**4380–4386.

74. **Mascher, T., D. Zahner, M. Merai, N. Balmelle, A. B. de Saizieu, and R. Hakenbeck.** 2003. The *Streptococcus pneumoniae cia* regulon: CiaR target sites and transcription profile analysis. J. Bacteriol. **185:**60–70.

75. **McAllister, L. J., H. J. Tseng, A. D. Ogunniyi, M. P. Jennings, A. G. McEwan, and J. C. Paton.** 2004. Molecular analysis of the psa permease complex of *Streptococcus pneumoniae*. Mol. Microbiol. **53:**889–901.

76. **McCarty, M.** 1986. The transforming principle: discovering that genes are made of DNA. W. W. Norton & Company, New York, NY.

77. **Merino, E., and C. Yanofsky.** 2005. Transcription attenuation: a highly conserved regulatory strategy used by bacteria. Trends Genet. **21:**260–264.

78. **Morlot, C., A. Zapun, O. Dideberg, and T. Vernet.** 2003. Growth and division of *Streptococcus pneumoniae*: localization of the high molecular weight penicillin-binding proteins during the cell cycle. Mol. Microbiol. **50:**845–855.

79. **Mrazek, J., L. H. Gaynon, and S. Karlin.** 2002. Frequent oligonucleotide motifs in genomes of three streptococci. Nucleic Acids Res. **30:**4216–4221.

80. **Mulholland, K.** 1999. Strategies for the control of pneumococcal diseases. Vaccine **17**(Suppl. 1)**:**S79–S84.

81. **Musher, D. M.** 2000. *Streptococcus pneumoniae*, p. 2128–2146. *In* G. L. Mandell, J. E. Bennett, and R. Dolin (ed.), Mandell, Douglas, and Bennett's principles and practice of infectious diseases, vol. 2. Churchill Livingstone, Philadelphia, PA.

82. **National Research Council.** 1996. Guide for the care and use of laboratory animals. National Academy Press, Washington, DC.

83. **Neidhardt, F. C., R. Curtiss, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.).** 1996. *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed. ASM Press, Washington, DC.

84. **Ng, W. L., G. T. Robertson, K. M. Kazmierczak, J. Zhao, R. Gilmour, and M. E. Winkler.** 2003. Constitutive expression of PcsB suppresses the requirement for the essential VicR (YycF) response regulator in *Streptococcus pneumoniae* R6. Mol. Microbiol. **50:**1647–1663.

85. **Novak, R., J. S. Braun, E. Charpentier, and E. Tuomanen.** 1998. Penicillin tolerance genes of *Streptococcus pneumoniae*: the ABC-type manganese permease complex Psa. Mol. Microbiol. **29:**1285–1296.

86. **Oggioni, M. R., F. Iannelli, and G. Pozzi.** 1999. Characterization of cryptic plasmids pDP1 and pSMB1 of *Streptococcus pneumoniae*. Plasmid **41:**70–72.

87. **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.

88. **Orihuela, C. J., G. Gao, M. McGee, J. Yu, K. P. Francis, and E. Tuomanen.** 2003. Organ-specific models of *Streptococcus pneumoniae* disease. Scand. J. Infect. Dis. **35:**647–652.

89. **Ottolenghi, E., and R. D. Hotchkiss.** 1960. Appearance of genetic transforming activity in pneumococcal cultures. Science **132:**1257–1258.

90. **Paik, J., I. Kern, R. Lurz, and R. Hakenbeck.** 1999. Mutational analysis of the *Streptococcus pneumoniae* bimodular class A penicillin-binding proteins. J. Bacteriol. **181:**3852–3856.

91. **Pericone, C. D., D. Bae, M. Shchepetov, T. McCool, and J. N. Weiser.** 2002. Short-sequence tandem and nontandem DNA repeats and endogenous hydrogen peroxide production contribute to genetic instability of *Streptococcus pneumoniae*. J. Bacteriol. **184:**4392–4399.

92. **Pericone, C. D., S. Park, J. A. Imlay, and J. N. Weiser.** 2003. Factors contributing to hydrogen peroxide resistance in *Streptococcus pneumoniae* include pyruvate oxidase (SpxB) and avoidance of the toxic effects of the Fenton reaction. J. Bacteriol. **185:**6815–6825.

93. **Pohlschroder, M., E. Hartmann, N. J. Hand, K. Dilks, and A. Haddad.** 2005. Diversity and evolution of protein translocation. Annu. Rev. Microbiol. **59:**91–111.

94. **Pozzi, G., and P. M. J. Lievens.** 1989. Vectors for cloing in streptococci derived from cryptic pneumococcal plasmid pDP1/pSMB1, p. 393–395. *In* L. O. Butler, C. Harwood, and B. E. B. Moseley (ed.), Genetic transformation and expression. Intercept, New York, NY.

95. **Pracht, D., C. Elm, J. Gerber, S. Bergmann, M. Rohde, M. Seiler, K. S. Kim, H. F. Jenkinson, R. Nau, and S. Hammerschmidt.** 2005. PavA of *Streptococcus pneumoniae* modulates adherence, invasion, and meningeal inflammation. Infect. Immun. **73:**2680–2689.

96. **Robertson, G. T., W. L. Ng, J. Foley, R. Gilmour, and M. E. Winkler.** 2002. Global transcriptional analysis of *clpP* mutations of type 2 *Streptococcus pneumoniae* and their effects on physiology and virulence. J. Bacteriol. **184:**3508–3520.

97. **Robertson, G. T., W. L. Ng, R. Gilmour, and M. E. Winkler.** 2003. Essentiality of *clpX*, but not *clpP*, *clpL*, *clpC*, or *clpE*, in *Streptococcus pneumoniae* R6. J. Bacteriol. **185:**2961–2966.

98. **Saluja, S. K., and J. N. Weiser.** 1995. The genetic basis of colony opacity in *Streptococcus pneumoniae*: evidence for the effect of box elements on the frequency of phenotypic variation. Mol. Microbiol. **16:**215–227.

99. **Salzberg, S. L., A. L. Delcher, S. Kasif, and O. White.** 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. **26:** 544–548.

100. **Sambrook, J., and D. W. Russell.** 2001. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

101. **Sanchez-Beato, A. R., R. Lopez, and J. L. Garcia.** 1998. Molecular characterization of PcpA: a novel choline-binding protein of *Streptococcus pneumoniae*. FEMS Microbiol. Lett. **164:**207–214.

102. **Schmidt, K. L., N. D. Peterson, R. J. Kustusch, M. C. Wissel, B. Graham, G. J. Phillips, and D. S. Weiss.** 2004. A predicted ABC transporter, FtsEX, is needed for cell division in *Escherichia coli*. J. Bacteriol. **186:**785–793.

103. **Sibold, C., Z. Markiewicz, C. Latorre, and R. Hakenbeck.** 1991. Novel plasmids in clinical strains of *Streptococcus pneumoniae*. FEMS Microbiol. Lett. **61:**91–95.

104. **Silva, N. A., J. McCluskey, J. M. Jefferies, J. Hinds, A. Smith, S. C. Clarke, T. J. Mitchell, and G. K. Paterson.** 2006. Genomic diversity between strains of the same serotype and multilocus sequence type among pneumococcal clinical isolates. Infect. Immun. **74:**3513–3518.

105. **Smith, M. D., and W. R. Guild.** 1979. A plasmid in *Streptococcus pneumoniae*. J. Bacteriol. **137:**735–739.

106. **Spellerberg, B., D. R. Cundell, J. Sandros, B. J. Pearce, I. Idanpaan-Heikkila, C. Rosenow, and H. R. Masure.** 1996. Pyruvate oxidase as a determinant of virulence in *Streptococcus pneumoniae*. Mol. Microbiol. **19:**803–813.

107. **Sung, C. K., H. Li, J. P. Claverys, and D. A. Morrison.** 2001. An *rpsL* cassette, Janus, for gene replacement through negative selection in *Streptococcus pneumoniae*. Appl. Environ. Microbiol. **67:**5190–5196.

108. **Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale.** 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:**41.

109. **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. J. Mitchell, D. A. Morrison, and B. G. Spratt (ed.), The pneumococcus. ASM Press, Washington, DC.

110. **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.

111. **Throup, J. P., K. K. Koretke, A. P. Bryant, K. A. Ingraham, A. F. Chalker, Y. Ge, A. Marra, N. G. Wallis, J. R. Brown, D. J. Holmes, M. Rosenberg, and M. K. Burnham.** 2000. A genomic analysis of two-component signal transduction in *Streptococcus pneumoniae*. Mol. Microbiol. **35:**566–576.

112. **Tomasz, A.** 2000. *Streptococcus pneumoniae*: molecular biology and mechanisms of disease. Mary Ann Liebert, Larchmont, NY.

113. **Tuomanen, E. I., and H. R. Masure.** 1997. Molecular and cellular biology of pneumococcal infection. Microb. Drug Resist. **3:**297–308.

114. **Tuomanen, E. I., T. J. Mitchell, D. A. Morrison, and B. G. Spratt (ed.).** 2004. The pneumococcus. ASM Press, Washington, DC.

115. **White, S. W., J. Zheng, Y. M. Zhang, and Rock.** 2005. The structural biology of type II fatty acid biosynthesis. Annu. Rev. Biochem. **74:**791–831.

116. **Yang, X., and C. W. Price.** 1995. Streptolydigin resistance can be conferred by alterations to either the beta or beta′ subunits of *Bacillus subtilis* RNA polymerase. J. Biol. Chem. **270:**23930–23933.

117. **Yother, J.** 2004. Capsules, p. 30–48. *In* E. I. Tuomanen, T. J. Mitchell, D. A. Morrison, and B. G. Spratt (ed.), The pneumococcus. ASM Press, Washington, DC.

118. **Yu, L., J. Mack, P. J. Hajduk, S. J. Kakavas, A. Y. Saiki, C. G. Lerner, and E. T. Olejniczak.** 2003. Solution structure and function of an essential CMP kinase of *Streptococcus pneumoniae*. Protein Sci. **12:**2613–2621.