



Complete Genome Sequence of a Virulent Isolate of *Streptococcus pneumoniae*
Hervé Tettelin *et al.*
Science **293**, 498 (2001);
DOI: 10.1126/science.1061217

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of January 2, 2013):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
<http://www.sciencemag.org/content/293/5529/498.citation>

Supporting Online Material can be found at:
<http://www.sciencemag.org/content/suppl/2001/07/19/293.5529.498.DC1.html>

A list of selected additional articles on the Science Web sites **related to this article** can be found at:
<http://www.sciencemag.org/content/293/5529/498.citation#related>

This article has been **cited by** 549 article(s) on the ISI Web of Science

This article has been **cited by** 100 articles hosted by HighWire Press; see:
<http://www.sciencemag.org/content/293/5529/498.citation#related-urls>

This article appears in the following **subject collections**:
Microbiology
<http://www.sciencemag.org/cgi/collection/microbio>

43. Lysates from frozen brain human tissue were prepared as in (24). Radioactive RT-PCR was performed in a total volume of 50 μ l containing cDNA synthesized from 0.25 μ g RNA, 20 mM Tris-HCl, pH 8.4, 50 mM KCl, 1.5 mM MgCl₂, 0.2 mM dNTPs, 1.7 μ M [α -³²P]CTP, and 0.4 μ M of the primers as follows: hBDNF5', 5'-AGCCA-GAATCGGAACACCA-3'; hBDNF3', 5'-GCACACCT-GGCTAGGCCAAG-3'. PCR amplification was carried out for 30 cycles. Each cycle consisted of the following steps: 94°C for 30 s, 57°C for 30 s and 72°C for 30 s. The same amount of each cDNA was also amplified, independently, with SNAP-25 (synaptosomal associated protein 25, a presynaptic membrane-associated protein localized in grown cones, axons and presynaptic terminals) specific primers. SNAP-25 5', 5'-CAAATGATGC-CCGAGAAAT-3'; SNAP25 3', 5'-GGAATCAGCCT-TCTCCATGA-3'. PCR products were separated by non-denaturing 8% polyacrylamide gel electrophoresis and visualized by autoradiography. BDNF levels were quantified and normalized relative to SNAP-25 levels.
44. V. O. Ona, et al. *Nature* **399**, 263 (1999).
45. Total cellular lysates from conditionally immortalized CNS cells (13, 27) were obtained in a buffer containing Tris 50 mM pH 7.4, 5 mM NaCl, Triton X100 1%, 1 mM DTT, 15 mM EGTA supplemented with 1:100 of Protease Inhibitor Cocktail (Sigma). Immunoprecipitates were obtained by incubating the total cellular lysate (from 4×10^6 cells) with Mab2166 (1:1000) following conventional immunoprecipitation protocols and loaded. The blotted proteins were exposed to antibody to Htt Mab2166 (dilution 1:5000; Chemicon, Temecula, CA). RNA was reverse-transcribed into single-stranded cDNA using Superscript II RNase H⁻ Reverse Transcriptase (Life Technologies) as described by the manufacturer. PCR was performed in a total volume of 50 μ l containing 1 μ g cDNA, 20 mM Tris-HCl, pH 8.4, 50 mM KCl, 1.5 mM MgCl₂, 0.2 mM dNTPs, 5% dimethyl sulfoxide (DMSO), 0.4 μ M of Htt-specific primers (5'-CGAC-CCTGGAAAAGCTGATGAA-3' and 5'-CACACG-GTCTTCTTGGTAGCTGA-3'), 2 U Taq polymerase (Life Technologies). Amplification was carried out for 25 cycles. Each cycle consisted of the following steps: 94°C for 30 s, 56°C for 30 s, and 72°C for 60 s. PCR products were separated by electrophoresis on 2% agarose gel and visualized by staining with ethidium bromide.
46. E. Cattaneo et al., *Trends Neurosci.* **24**, 182 (2001).
47. A. C. Bachoud-Levi et al., *Lancet* **356**, 1975 (2000).
48. The research described in this manuscript was entirely developed at the Department of Pharmacological Sciences, University of Milano. Supported by grants from the Huntington's Disease Society of America (HDSA, New York), Telethon (Italy #E840) and Ministero dell'Universita' e della Ricerca Scientifica (Italy, Murst#MM06278849-005), and in part by a grant from the Hereditary Disease Foundation (HDF, Santa Monica) (E.C.) and by funds from Associazione Amici Centro "Dino Ferrari," Milano, Italy (V.S.). T.T. was supported by grants from the Swedish Medical Research Council and Life 2000 Program of the Academy of Finland. We thank R. Molteni for help in setting the RNase Protection Assays. E.C., M.E.M., R.M.F., and M.R.H. are members of the "Coalition for the Cure" (HDSA) and of the "Cure HD Initiative" (HDF).

5 February 2001; accepted 1 June 2001
Published online 14 June 2001;
10.1126/science.1059581
Include this information when citing this paper.

Complete Genome Sequence of a Virulent Isolate of *Streptococcus pneumoniae*

Hervé Tettelin,¹ Karen E. Nelson,¹ Ian T. Paulsen,^{1,2} Jonathan A. Eisen,^{1,2} Timothy D. Read,¹ Scott Peterson,^{1,3} John Heidelberg,¹ Robert T. DeBoy,¹ Daniel H. Haft,¹ Robert J. Dodson,¹ A. Scott Durkin,¹ Michelle Gwinn,¹ James F. Kolonay,¹ William C. Nelson,¹ Jeremy D. Peterson,¹ Lowell A. Umayam,¹ Owen White,¹ Steven L. Salzberg,^{1,4} Matthew R. Lewis,¹ Diana Radune,¹ Erik Holtzapple,¹ Hoda Khouri,¹ Alex M. Wolf,¹ Terry R. Utterback,¹ Cheryl L. Hansen,¹ Lisa A. McDonald,¹ Tamara V. Feldblyum,¹ Samuel Angiuoli,¹ Tanja Dickinson,¹ Erin K. Hickey,¹ Ingeborg E. Holt,¹ Brendan J. Loftus,¹ Fan Yang,¹ Hamilton O. Smith,^{1*} J. Craig Venter,^{1*} Brian A. Dougherty,⁵ Donald A. Morrison,⁶ Susan K. Hollingshead,⁷ Claire M. Fraser^{1,3†}

The 2,160,837-base pair genome sequence of an isolate of *Streptococcus pneumoniae*, a Gram-positive pathogen that causes pneumonia, bacteremia, meningitis, and otitis media, contains 2236 predicted coding regions; of these, 1440 (64%) were assigned a biological role. Approximately 5% of the genome is composed of insertion sequences that may contribute to genome rearrangements through uptake of foreign DNA. Extracellular enzyme systems for the metabolism of polysaccharides and hexosamines provide a substantial source of carbon and nitrogen for *S. pneumoniae* and also damage host tissues and facilitate colonization. A motif identified within the signal peptide of proteins is potentially involved in targeting these proteins to the cell surface of low-guanine/cytosine (GC) Gram-positive species. Several surface-exposed proteins that may serve as potential vaccine candidates were identified. Comparative genome hybridization with DNA arrays revealed strain differences in *S. pneumoniae* that could contribute to differences in virulence and antigenicity.

Streptococcus pneumoniae (pneumococcus) has played a pivotal role in the fields of genetics and microbiology. The pioneering studies of Avery, MacLeod, and McCarty in 1944 (1) demonstrated that DNA is the true hereditary material by transforming a nonencapsulated, avirulent *S. pneu-*

moniae strain with DNA from a capsulated virulent strain. This work highlighted the importance of the bacterial polysaccharide capsule as a key pathogenicity factor.

As a human pathogen, *S. pneumoniae* is the most common bacterial cause of acute respira-

tory infection and otitis media and is estimated to result in over 3 million deaths in children every year worldwide from pneumonia, bacteremia, or meningitis (2). Even more deaths occur among elderly people, among whom *S. pneumoniae* is the leading cause of community-acquired pneumonia and meningitis (3). Since 1990, the number of penicillin-resistant strains has increased from 1 to 5% to 25 to 80% of isolates, and many strains are now resistant to commonly prescribed antibiotics such as penicillin, macrolides, and fluoroquinolones (4).

The complete genome sequence of a capsular serotype 4 isolate of *S. pneumoniae* [designated TIGR4 (5); TIGR indicates The Institute for Genomic Research] was determined by the random shotgun sequencing strategy (6) (GenBank accession number AE005672; see www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl). This clinical isolate was taken from the blood of a 30-year-old male patient in Kongsvinger, Norway, and is highly invasive and virulent in a mouse model of infection (7).

The genome consists of a single circular chromosome of 2,160,837 base pairs (bp) with a G + C content of 39.7%. Base pair 1 of the chromosome was assigned within the putative origin of replication. Of the 2236 genes identified (8), 1155 are located on the right of the

¹The Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA.

²Johns Hopkins University, Charles and 34th Streets, Baltimore, MD 21218, USA.

³George Washington University Medical Center, 2300 Eye Street, NW, Washington, DC 20037, USA.

⁴Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA.

⁵Bristol-Myers Squibb PRI, 5 Research Parkway, Wallingford, CT 06492, USA.

⁶University of Illinois at Chicago, 900 South Ashland Avenue, Chicago, IL 60607, USA.

⁷University of Alabama at Birmingham, 845 19th Street South, Birmingham, AL 35294, USA.

*Present address: Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA.

†To whom correspondence should be addressed. E-mail: cmfraser@tigr.org

REPORTS

origin of replication, and 916 (79%) of these are transcribed in the same direction as DNA replication; similarly, 1081 genes are on the left of the origin of replication, and 857 (79%) of them

transcribed in the same direction [Fig. 1 and Web fig. 1 (9)]. This type of gene orientation bias appears to be a common feature of low-GC Gram-positive organisms (10).

Although the *S. pneumoniae* genome was reported to contain six ribosomal RNA (rRNA) operons (11), the TIGR4 isolate contains only four rRNA operons. Only 12 of the 58 tRNAs

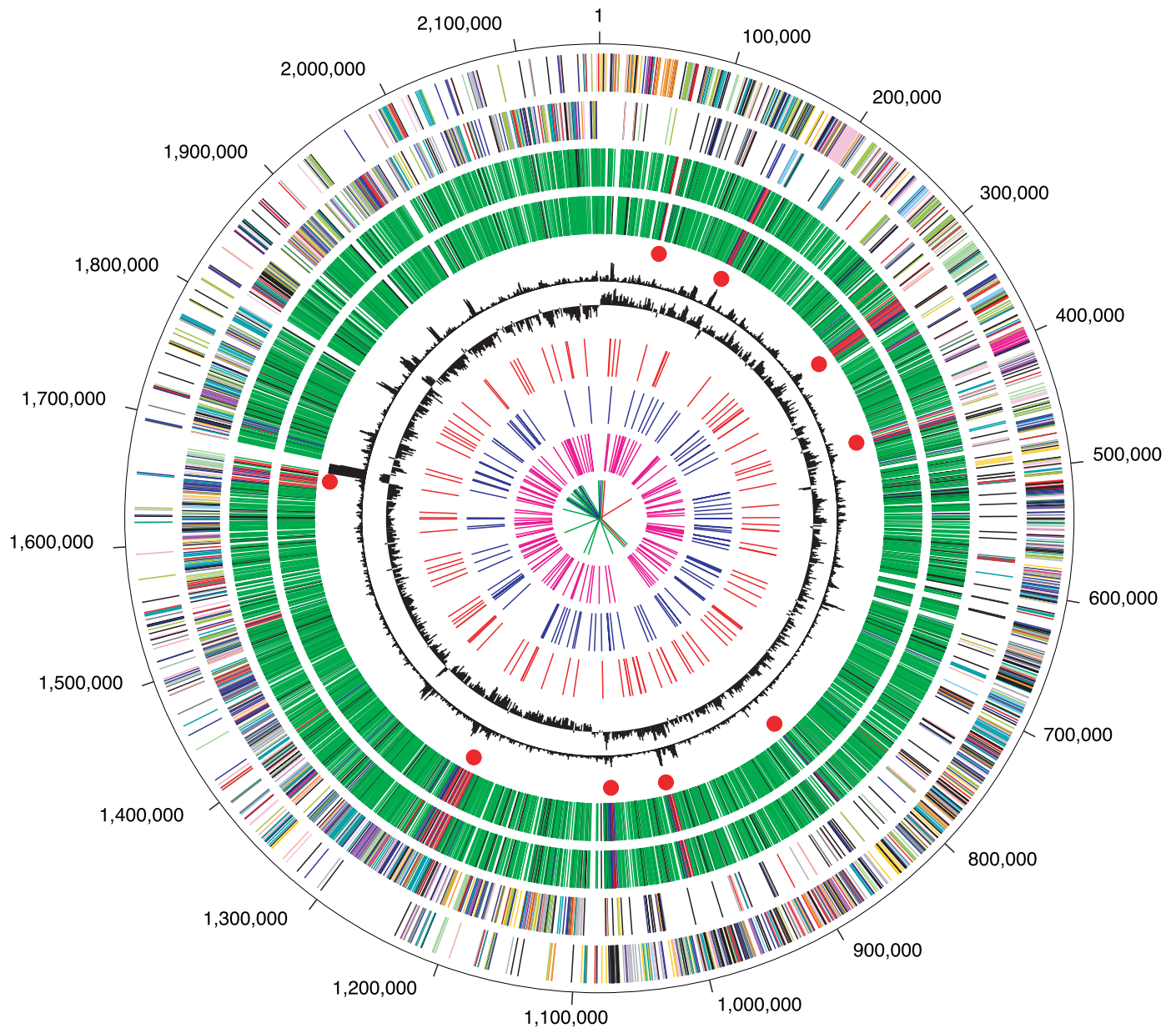


Fig. 1. Circular representation of the *S. pneumoniae* TIGR4 genome and comparative genome hybridizations using microarrays. Comparative genome hybridizations are used to identify genomic differences between the TIGR4 isolate and strains R6 and D39, using a preliminary microarray. Results are displayed on the third and fourth circles. Genes were classified in four groups: (i) gene not present on the array and not analyzed (black) (394 genes, 17% of total); (ii) ortholog present in the test strain (green); (iii) ortholog absent in the test strain (red); and (iv) ambiguous result (blue). The Cy3/Cy5 ratio (TIGR4 signal/test strain) cutoffs for each category were determined subjectively as $\text{Cy3/Cy5} = 1.0$ to 3.0, green; 3.0 to 10.0, blue; and >10.0 , red. There were a number of loci for which hybridization ratios fell between what is expected for gene presence or absence (Cy3/Cy5 ratios between 3.0 to 10.0). Ambiguous results (blue bars) can be explained in at least two ways: (i) The gene may be highly diverged in R6 and/or D39 relative to the TIGR4 isolate. (ii) Alternatively, the gene may be absent in R6 and/or D39 but still be able to produce a hybridization signal, because the TIGR4 isolate gene is a member of a

paralogous gene family or a repetitive element. The outer circle shows predicted coding regions on the plus strand, color-coded by role categories: salmon, amino acid biosynthesis; light blue, biosynthesis of cofactors and prosthetic groups and carriers; light green, cell envelope; red, cellular processes; brown, central intermediary metabolism; yellow, DNA metabolism; green, energy metabolism; purple, fatty acid and phospholipid metabolism; pink, protein fate/synthesis; orange, purines, pyrimidines, nucleosides, and nucleotides; blue, regulatory functions; grey, transcription; teal, transport and binding proteins; black, hypothetical and conserved hypothetical proteins. The second circle shows predicted coding regions on the minus strand, color-coded by role categories. The third circle shows strain R6 genes. The fourth circle shows strain D39 genes. The fifth circle shows an atypical nucleotide composition curve; the nine gene clusters that are absent in strains R6 and D39 are indicated by red bullets. The sixth circle shows the GC-skew curve. The seventh circle shows IS elements. The eighth circle shows RUP elements. The ninth circle shows BOX elements. The tenth circle shows rRNAs in blue, tRNAs in green, and structural RNAs in red.



are not found adjacent to a rRNA operon [Fig. 1 and Web fig. 1 (9)]. Three structural RNAs were identified: a tRNA-like/mRNA-like (tm) RNA (www.indiana.edu/~tmrna/), a signal recognition particle RNA (12), and a ribonuclease P RNA (13).

Biological roles were assigned to 1440 (64%) of the predicted proteins according to the classification scheme adapted from Riley (14). Another 359 (16%) predicted proteins matched proteins of unknown function, and the remaining 437 (20%) had no database match. A total of 260 paralogous protein families were identified in the TIGR4 isolate (8), containing 823 predicted proteins (37% of the total).

Comparative genome analysis identified 258 genes in *S. pneumoniae* [Web table 1 (9)] that probably were duplicated after the divergence of this species from other evolutionary lineages for which complete genomes are available (8). Such lineage-specific gene duplications may reveal species-specific adaptations, because gene duplication is frequently accompanied by functional diversification and divergence. These duplications in *S. pneumoniae* include bacteriocin genes, choline-binding proteins, immunoglobulin A (IgA) proteases, immunity proteins, glycosyl transferases, and a large number of hypothetical and conserved hypothetical proteins. Comparison of the complete set of predicted proteins of *S. pneumoniae* with those of other completely sequenced organisms revealed 1219 proteins that are most similar to a protein from another low-GC Gram-positive species (*Lactococcus lactis* has the most with 905) [Web fig. 2 (9)]. Only 105 proteins have no similarity to low-GC Gram-positive proteins [Web table 2 (9)].

Two adjacent genes (SP1467 and SP1468) displayed a high degree of DNA sequence identity (76 and 88%, respectively) between *S. pneumoniae* and *Haemophilus influenzae*. Both pairs of genes, which may be involved in pyridoxine biosynthesis, are more closely related to each other than to orthologs in any other species, which suggests that they were horizontally transferred between these respiratory pathogens.

The *S. pneumoniae* genome is rich in insertion sequences (ISs), which make up ~5% (101,045 bp) of the TIGR4 chromosome [Table 1, Fig. 1, and Web fig. 1 (9)]. IS genes make up

>3.5% (84 out of 2236) of the genes in *S. pneumoniae*, in contrast to other published genomes in which the percentage ranges from 0 to 3% (see www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl). In addition to IS elements, there are two full-length group II introns and a 1400-bp fragment of the streptococcal conjugative transposon Tn5252. The TIGR4 isolate does not contain any large prophagelike structure or full-length conjugative transposon. The majority of IS elements appear to be non-functional because of insertions, deletions, and/or point mutations (Table 1) that result in frameshifted or degenerated transposase genes. However, programmed frameshifting may allow the expression of several of the frameshifted genes (15). Intact elements are typically families with 98 to 100% nucleotide sequence identity, probably reflecting "waves" of expansion of IS element isotypes. Despite the large number of IS elements, only two genes (encoding hypothetical proteins SP2178/SP2180 and SP0327/SP0329) are disrupted, and one gene (encoding lacX protein SP1194) is truncated by an IS insertion. This suggests selection against insertions into most of the *S. pneumoniae* genes, or some form of editing to remove these insertions, or both. Regarding the latter, it is possible that the complete DNA transformation system identified in the TIGR4 isolate [Web table 3 (9)] may allow conversion of IS disrupted genes by homologous recombination.

Two types of small, dispersed DNA repeats—the RUP and the BOX elements—were identified previously in *S. pneumoniae*. The 107-bp RUP element is thought to act like a nonautonomous insertion sequence that is mobilized by the transposase of IS630-Spn1 (16). The TIGR4 isolate contains 108 RUP elements, which insert preferentially into IS elements. The BOX element is a modular DNA repeat that is composed of three subunits: *boxA*, *boxB* (which can be present in multiple copies), and *boxC* (17). There are 127 BOX elements in the TIGR4 isolate; of these, 115 are intact ($A_1B_0C_1$) and 12 are incomplete. The BOX elements do not appear to be linked to competence or virulence genes, as was previously suggested (17).

There appears to be a system for generating polymorphic type I restriction enzymes in *S. pneumoniae* similar to that found in *Mycoplasma*

pulmonis (18). Shotgun sequencing revealed populations of clones from the TIGR4 isolate that were fusions of type I restriction-modification enzyme specificity subunit *hsdS* pseudogenes SP0505 and SP0507 with the nearby intact *hsdS* gene SP0508 [Web fig. 3 (9)]. These rearrangements, which are recombination events between conserved inverted repeats (IRs) within SP0508 and the pseudogenes, might be catalyzed by a nearby integrase (SP0506). Polymerase chain reaction (PCR) on chromosomal DNA using primers inside and outside the *hsdS* genes indicated that the chromosomal region between the IRs was invertible. The specificity subunit may therefore have up to four possible sequences, presumably altering the DNA site recognition of the restriction-modification system and reducing the efficiency of DNA exchange between bacteria in the same clone line.

Streptococcus pneumoniae has the widest substrate utilization range for sugars and substituted nitrogen compounds of the three completed genomes of near-commensal residents of the human upper respiratory tract (*H. influenzae*, *Neisseria meningitidis*, and *S. pneumoniae*). Genome analysis suggests that *S. pneumoniae* possesses pathways for catabolism of pentitols via the pentose phosphate pathway, as well as for cellobiose, fructose, fucose, galactose, galactitol, glucose, glycerol, lactose, mannitol, mannose, raffinose, sucrose, trehalose, and maltosaccharides, which can flow directly into the glycolytic pathway (Fig. 2). Ten amino acids and *N*-acetylglucosamine can potentially be used as nitrogen and carbon sources. Genome analysis also revealed a large number of pathways for the complete or partial synthesis of 14 amino acids and chorismate (Fig. 2).

Streptococcus pneumoniae contains a high percentage of ATP-dependent transporters, as has been seen in other organisms lacking an electron transfer chain (19). *Streptococcus pneumoniae* possesses both a complete F-type proton adenosine triphosphatase (ATPase) and a V-type ATPase that is probably sodium ion-specific. It also has a sodium ion/proton exchanger and several probable sodium ion-driven transporters (Fig. 2), whose activity would be dependent on the establishment of a sodium motive force. Thus, *S. pneumoniae* can probably interconvert the proton gradient, the sodium

Fig. 2. Overview of metabolism and transport in *S. pneumoniae*. Pathways for energy production, metabolism of organic compounds, and capsule biosynthesis are shown. There exist other genes in the capsule biosynthesis locus to which no specific function could be assigned. Transporters are grouped by substrate specificity as follows: inorganic cations (green), inorganic anions (pink), carbohydrates/carboxylates (yellow), amino acids/peptides/amines/purines and pyrimidines (red), and drug efflux and other (black). Question marks indicate uncertainty about the substrate transported. Export or import of solutes is designated by the direction of the arrow through the transporter. The energy-coupling mechanisms of the transporters are also shown: Solutes transported by channel proteins are shown with a double-headed arrow; secondary transporters are shown with two arrowed lines, indicating both the solute and the coupling ion; ATP-driven transporters are

indicated by the ATP hydrolysis reaction; and transporters with an unknown energy coupling mechanism are shown with only a single arrow. Components of transporter systems that function as multisubunit complexes that were not identified are outlined with dotted lines. Where multiple homologous transporters with similar substrate predictions exist, the number of that type of transporter is indicated in parentheses. Systematic gene numbers (SPXXXX) are indicated next to each pathway or transporter; those separated by a dash represent a range of consecutive genes. Details for the PTS transporters are indicated in Web fig. 4 (9). Abbreviations are as follows: ADP, adenosine diphosphate; UMP, uridine monophosphate; UDP, uridine diphosphate; FucNAc, *N*-acetylglucosamine; Gal, galactose; GalNAc, *N*-acetylgalactosamine; GluNAc, *N*-acetylglucosamine; ManNAc, *N*-acetylmannosamine; NeurNAc, *N*-acetylneuraminic acid; P, phosphate; PP, diphosphate; Pyr, pyruvate.

REPORTS

ion gradient, and ATP as energy sources, using its F- and V-type ATPases and its sodium ion/proton exchanger. This is somewhat similar to the activity of *Treponema pallidum*, which possesses two V-type ATPases, probably for protons and sodium ions, but no exchanger (20).

Over 30% of the transporters in *S. pneumoniae* were predicted to be sugar transporters (Fig. 2), which is the highest percentage observed to date in any sequenced prokaryote (19). Other completely sequenced respiratory tract organisms, *H. influenzae* and *N. meningitidis*, have a paucity of sugar transporters and are much more reliant on carboxylates and other compounds for their carbon needs. This suggests that *S. pneumoniae* may occupy a distinct microenvironment within the respiratory tract. Host glycoproteins and murein polysaccharides, as well as its own capsular polysaccharides, may be major sources of sugars for *S. pneumoniae*. Reliance on sugar transport and metabolism appears to be a common feature of streptococci, based on their abundance in sugar-rich environments such as the oral cavity (21).

The *S. pneumoniae* sugar transporters primarily consist of phosphoenolpyruvate (PEP)-dependent phosphotransferase system (PTS) transporters and ATP-binding cassette (ABC) transporters. *Streptococcus pneumoniae* has 21 PTS sugar-specific enzyme II complexes with a variety of gene and domain arrangements [Web fig. 4 (9)], more than twice as many as any other sequenced organism relative to genome size, again emphasizing the importance of sugars to the life-style of *S. pneumoniae*. It also possesses single copies of the general PTS enzymes enzyme I and histidine-containing protein (HPr), as well as a HPr serine kinase for regulatory purposes. The *S. pneumoniae* PTS includes systems specific for fructose, glucose, lactose, mannose, mannitol, trehalose, *N*-acetylglucosamine, and sucrose, as well as a variety of PTS systems whose sugar specificities remain to be determined. One PTS system (SP2161 to SP2164) is encoded within a gene cluster including all of the genes necessary for fucose metabolism. *N*-acetylglucosamine is a constituent of the capsule of the TIGR4 isolate, and it is therefore possible that this system may be a PTS for the uptake of *N*-acetylglucosamine or other fucose derivatives. In addition to the PTS, there are seven ABC sugar uptake systems, most of which do not have cytoplasmic ATP-binding components encoded with the other components (Fig. 2).

Streptococcus pneumoniae also possesses a variety of ATP- and ion-driven amino acid transporters, as well as transporters for polyamines, uracil, and xanthine. A single ABC transporter lacking a binding protein was found for choline, an important requirement for the streptococcal cell wall. In contrast to the emphasis on sugar transport, only a single transporter was found for monocarboxylates and one for dicarboxylates. *Streptococcus pneumoniae* has a

relatively limited repertoire of transporters for inorganic anion and cations, although this includes a manganese ABC transporter (SP1648

to SP1650) and a zinc transporter (SP2169 to SP2171), which have been associated with virulence (22), as well as three ferric iron and three

Table 1. *S. pneumoniae* IS families.

IS family*	Name (isotype)	IS size (nt)†	Intact transposase	Truncated or frameshifted	Species with homologous elements‡
IS3	IS3-Spn	1359	0	14	Sp Ec My Sg Ne Ha La Ba
IS5	IS1381-Spn	854–860	0	12	La
IS5	IS1515	861	0	1§	Sp Fr Cy La
IS30	IS1239	1046	0	2	Sp So Cl St Ae Le
IS66	IS66	2484–2498	0	7	
IS110	–	?	0	2	
IS605	IS200	747	2	1	Ec Sa Ye En Cl Ha Vi Wo Th De
IS630	IS630-Spn1	896	0	12	Sp Sy Ne
IS1380	IS1380-Spn	1703	11	1	Ab Sp Ba Xa Kl Sm
ISL3	IS1167	1414–1432	8	14	Sp Sh Sd En La St Le Mi
Unknown			0	17	
Total			21	84	

*According to the Mahillon and Chandler classification [J. Mahillon, M. Chandler, *Microbiol. Mol. Biol. Rev.* **62**, 725 (1998)]. †Distance between inverted repeats flanking intact or nontruncated IS elements. ‡Species with the most similar elements in GenBank. BlastP hits with an *E* value <10^{–20} were included. Key: Ab, *Acetobacter*; Ae, *Aeromonas*; Ba, *Bacillus*; Cl, *Clostridium*; Cy, *Cyanobacterium*; De, *Deinococcus*; Ec, *E. coli*; En, *Enterococcus*; Fr, *Fremyella*; Ha, *Haemophilus*; Kl, *Klebsiella*; La, *Lactobacillus*; Le, *Leuconostoc*; Mi, *Microcystis*; My, *Mycoplasma*; Ne, *Neisseria*; Sa, *Salmonella*; Sg, *S. agalactiae*; Sd, *S. gordonii*; Sh, *S. thermophilus*; Sm, *Sphingomonas*; So, *S. pyogenes*; Sp, *S. pneumoniae*; St, *Staphylococcus*; Sy, *Synechocystis*; Th, *Thermotoga*; Vi, *Vibrio*; Wo, *Wolbachia*; Xa, *Xanthobacter*; Ye, *Yersinia*. §*S. pneumoniae* element demonstrates functional activity [R. Munoz, R. Lopez, E. Garcia, *J. Bacteriol.* **180**, 1381 (1998)].

Table 2. Subset of *S. pneumoniae* genes related to virulence containing stretches of iterative DNA that could induce phase-variation. Iterative DNA motifs, including homopolymeric tracts, were searched in the TIGR4 genome [see (29)]. The iterative motifs identified in genes related to virulence are displayed. Abbreviations under "location" are as follows: 5', the motif is in the 5' third of the gene; M, the motif is in the middle third; 3', the motif is in the 3' third; P, the motif is within 50 nt upstream of the translation start site. For SP1772, repeats occur in all three parts of the protein.

ORF	Description	Repeat	Location
SP0071	Immunoglobulin A1 protease	(AT) ₄ , (TA) ₄	M, 3'
SP0102	Glycosyl transferase	(G) ₆	M
SP0168	Putative macrolide efflux protein	(TTA) ₄	5'
SP0346	Capsular polysaccharide biosynthesis protein (Cps4A)	(TATT) ₃	5'
SP0349	Capsular polysaccharide biosynthesis protein (Cps4D)	(A) ₈	5'
SP0350	Capsular polysaccharide biosynthesis protein (Cps4E)	(AG) ₄	M
SP0351	Capsular polysaccharide biosynthesis protein (Cps4F)	(A) ₈ , (A) ₉	5', 5'
SP0352	Capsular polysaccharide biosynthesis protein (Cps4G)	(AT) ₄ , (T) ₈	5', M
SP0353	Capsular polysaccharide biosynthesis protein (Cps4H)	(A) ₈	5'
SP0462	Cell wall surface anchor family protein	(GA) ₄	M
SP0664	Putative zinc metalloprotease (ZmpB)	(CAAAA) ₃	5'
SP0689	UDP- <i>N</i> -acetylglucosamine- <i>N</i> -acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol <i>N</i> -acetylglucosamine transferase	(G) ₆ , (G) ₆	5'
SP0907	Putative capsular polysaccharide biosynthesis protein	(G) ₆	5'
SP0966	Adherence and virulence protein A	(A) ₈	5'
SP1267	LicC protein	(ATG) ₄ , (AG) ₄	5', M
SP1272	Putative polysaccharide biosynthesis protein	(CT) ₄ , (CT) ₄	M, 3'
SP1274	LicD2 protein	(A) ₈	5'
SP1492	Cell wall surface anchor family protein	(CT) ₄	3'
SP1693	Neuraminidase A, authentic frameshift	(T) ₈	5'
SP1769	Glycosyl transferase, authentic frameshift	(C) ₉ , (CT) ₄	5', M
SP1772	Cell wall surface anchor family protein	(TCAGCGTCGACAA GTGCGTCGGCC) ₅₄₀	
SP1950	Putative bacteriocin formation protein	(T) ₉	P
SP2136	Choline-binding protein (PcpA)	(T) ₈ , (T) ₈	5'
SP2145	Antigen, cell wall surface anchor family	(G) ₆	5'
SP2190	Choline-binding protein A (CbpA)	(T) ₈ , (T) ₈	5', M

REPORTS

Table 3. *S. pneumoniae* proteins likely to be exposed on the surface, based on computer predictions [see (33)].

ORF	Description	LPxTG*	Choline [†] binding	Lipoprotein [‡]	SignalP [§]	YSIRK	Atypical [¶]	Repeat [#]
SP0057	Beta-N-acetylhexosaminidase (StrH)	+			+	+		
SP0069	Choline-binding protein I (CbpI)		+					
SP0071	Immunoglobulin A1 protease (Iga)	+			+	+		++
SP0082	Cell wall surface anchor family protein	+			+	+		
SP0092	ABC transporter, substrate-binding protein			+	+			
SP0112	Amino acid ABC transporter, periplasmic amino acid-binding protein, putative			+	+			
SP0117	Pneumococcal surface protein A (PspA)		+		+			
SP0148	ABC transproter, substrate-binding protein			+	+			
SP0149	Lipoprotein			+	+			
SP0191	Hypothetical protein			+	+			
SP0198	Hypothetical protein			+	+			
SP0268	Alkaline amylopullulanase, putative	+			+	+		
SP0314	Hyaluronidase	+			+			
SP0368	Cell wall surface anchor family protein, authentic frameshift	+			+	+		
SP0377	Choline-binding protein C (CbpC)		+		+			
SP0378	Choline-binding protein J (CbpJ)		+		+			
SP0390	Choline-binding protein G (CbpG)		+					
SP0391	Choline-binding protein F (CbpF)		+		+			
SP0462	Cell wall surface anchor family protein	+			+			+
SP0463	Cell wall surface anchor family protein	+			+			
SP0464	Cell wall surface anchor family protein	+			+			
SP0468	Sortase, putative			+	+			
SP0498	Endo-beta-N-acetylglucosaminidase, putative	+			+	+		
SP0620	Amino acid ABC transproter, amino acid-binding protein, putative			+	+			
SP0629	Conserved hypothetical protein			+	+			
SP0641	Serine protease, subtilase family	+			+			+++
SP0648	Beta-galactosidase (BgaA)	+			+	+		
SP0659	Thioredoxin family protein			+	+			
SP0664	Zinc metalloprotease ZmpB, putative	+			+		+	+
SP0667	Pneumococcal surface protein, putative		+		+			
SP0771	Peptidyl-prolyl cis-trans isomerase, cyclophilin-type			+	+			
SP0845	Lipoprotein			+	+			
SP0899	Conserved hypothetical protein			+	+			
SP0930	Choline-binding protein E (CbpE)		+		+			
SP0965	Endo-beta-N-acetylglucosaminidase (LytB)		+		+			
SP0981	Protease maturation protein, putative			+	+			+
SP1000	Thioredoxin family protein			+	+			
SP1002	Adhesion lipoprotein			+	+			
SP1032	Iron-compound ABC transporter, iron compound-binding protein			+	+		+	
SP1154	Immunoglobulin A1 protease (Iga)	+			+	+		
SP1394	Amino acid ABC transporter, amino acid-binding protein			+	+			
SP1400	Phosphate ABC transporter, phosphate-binding protein, putative			+	+			
SP1417	PspC-related protein, degenerate		+					+
SP1492	Cell wall surface anchor family protein	+						+
SP1500	Amino acid ABC transporter, amino acid-binding protein (AatB)			+	+			
SP1527	Oligopeptide ABC transporter, oligopeptide-binding protein (AliB)			+	+			
SP1573	Lysozyme (LytC)		+		+			+
SP1650	Manganese ABC transporter, manganese-binding adhesion liprotein			+	+			
SP1683	Sugar ABC transporter, sugar-binding protein			+	+			
SP1690	ABC transporter, substrate-binding protein			+	+			
SP1772	Cell wall surface anchor family protein	+					+	+(540)
SP1796	ABC transporter, substrate-binding protein			+	+			
SP1826	ABC transporter, substrate-binding protein			+	+			

(Continued on page 504)

Table 3. (Continued)

ORF	Description	LPxTG*	Choline [†] binding	Lipoprotein [‡]	SignalP§	YSIRK	Atypical¶	Repeat#
SP1833	Cell wall surface anchor family protein	+				+	+	
SP1870	Iron-compound ABC transporter, permease protein			+	+			
SP1872	Iron-compound ABC transporter, iron-compound binding protein			+	+			+
SP1891	Oligopeptide ABC transporter, oligopeptide-binding protein (AmiA)			+	+			+
SP1897	Sugar ABC transporter, sugar-binding protein (MsmE)			+				
SP1937	Autolysin (LytA)		+					
SP1975	SpolI family protein			+	+			
SP1992	Cell wall surface anchor family protein	+			+			
SP2041	SpolI family protein			+	+			
SP2084	Phosphate ABC transporter, phosphate-binding protein (PstS)			+	+			
SP2108	Maltose/maltodextrin ABC transporter, maltose/maltodextrin-binding protein (MalX)			+	+			
SP2136	Choline-binding protein (PcpA)		+				+	++
SP2169	Zinc ABC transporter, zinc-binding lipoprotein (AdcA)			+	+			
SP2190	Choline-binding protein A (CbpA)	+	+		+	+		++
SP2197	ABC transporter, substrate-binding protein, putative			+	+			
SP2201	Choline-binding protein D (CbpD)		+		+			

*Sortase motif. [†]Choline-binding motif. [‡]Lipid attachment motif. [§]Signal peptide; a Y-score lower limit of 0.3 was used as the cutoff. ||Signal peptide YSIRK for Gram-positive cell wall-attached proteins. ¶ORFs present in regions of atypical nucleotide composition [see (40)]. #ORFs containing iterative DNA motifs that could induce repeat-associated phase variation; one plus sign is shown per motif (exception: SP1772 contains 540 copies of a 24-nt motif).

phosphate ABC transporters. Overcoming iron and phosphate limitation may also be important for virulence. *Streptococcus pneumoniae* possesses an ABC efflux system involved in competence (SP0042 and SP0043). The characterized macrolide efflux proteins MefE and MefA (23) are absent from the TIGR4 isolate.

Analysis of the genome sequence suggests that extracellular enzyme systems for the metabolism of polysaccharides and hexosamines are important for providing carbon and nitrogen for this organism and may be important for the synthesis of the capsule and the virulence of this species. Enzyme systems based on *N*-acetylglucosaminidases, α - and β -galactosidases, endoglycosidases, hydrolases, hyaluronidases, and neuraminidases are present in *S. pneumoniae*. These enzymes probably enable degradation of host polymers, including mucins, glycolipids, and hyaluronic acid, as well as degradation of the organism's own capsule. These enzymatic activities may serve to increase substrate availability to *S. pneumoniae* by converting larger polymers to products that can be transported into the cell, while at the same time damaging host tissues and facilitating colonization.

Pathogenesis and virulence in *S. pneumoniae* are associated with the inflammation and colonization of host tissues and with bypass of the host immune system [Web table 4 (9)] (24). The polysaccharide capsule is considered to be the primary pneumococcal virulence determinant, allowing for the evasion of the host immune response (25). Although no pathway

has been biochemically characterized for the synthesis of the type 4 capsular polysaccharide, a proposed pathway for capsular biosynthesis derived from the genome analysis is shown in Fig. 2. A 13-gene cluster (SP0346 to SP0360) was identified that is likely to be involved in capsular biosynthesis and secretion. This region of the genome has an atypical nucleotide composition and is flanked by two IS elements on each side. Outside of the IS elements are the *aliA* (also called *plpA*) (SP0366) and *dexB* (SP0342) genes, which also flank the capsule loci in other *S. pneumoniae* strains (26). This gene cluster may not represent the complete pathway for capsular biosynthesis, because several other capsular polysaccharide biosynthesis genes are dispersed elsewhere in the genome. An operon of genes involved in the incorporation of phosphorylcholine into teichoic acid is also present in this genome (SP1267 to SP1274), as are all the genes required for peptidoglycan synthesis.

Phase variation has been described in *S. pneumoniae* and shown to involve variation of multiple cell-surface structures that contribute to the ability of the organism to interact with its host (27). One of the mechanisms involves reversible, high-frequency molecular switching of genes through slippagelike mechanisms at iterative DNA motifs, especially homopolymeric tracts (28). Such motifs were identified in the TIGR4 genome (29), and their location was correlated to predicted genes and their promoters. In total, 397 genes (18%) contain iterative

DNA motifs [Web table 5 (9)] and 25 of these are directly related to virulence (Table 2), including genes from the teichoic acid and capsule pathways that are associated with colony opacity variation (30). In contrast to other pathogenic species, most of the nucleotide repeat-containing genes in *S. pneumoniae* are not frameshifted. This might reflect the presence of general mismatch repair in *S. pneumoniae* (31), a process absent in many pathogens (32).

Sixty-nine proteins that are likely to be exposed on the surface of this organism were identified (Table 3) (33). Genomewide analysis of all predicted signal sequences (34) revealed two discernable clusters. The first cluster contains most of the lipoproteins for which the lipid attachment motif (33) extends beyond the covalently modified cysteine and the membrane-spanning region. This suggests some reuse of lipoprotein signal sequences as evolutionary cassettes. The second cluster, composed of proteins anchored in the cell wall through their sortase motif (33), revealed a previously uncharacterized pentapeptide motif (Y/F)SIRK (35), starting usually at residue 12 (Table 3). A large fraction of the surface proteins of various species of *Streptococcus* and *Staphylococcus* display this motif in their signal peptides. The near-perfect conservation of glycine and serine at the fourth and seventh positions past the pentapeptide, within the predicted transmembrane helix, suggests a specific functional interaction and may reflect a step in cell wall attachment in *S. pneumoniae* and related species.

Among the newly identified surface-exposed genes are a putative alkaline amylopullulanase (SP0268) and a putative endo- β -*N*-acetylglucosaminidase (SP0498). These two genes could be involved in the degradation of host polysaccharides. Several cell-wall surface anchor family proteins and lipoproteins are also possibly involved in adherence to host cells. An unusual surface-associated component in this genome is a 4776-amino acid protein (SP1772) that contains 540 imperfect repeats of the amino acid motif SASTSASA (35). This protein is similar to the *Lactobacillus brevis* surface layer protein (36) and to proteins from *S. gordonii* and *S. cristatus*. It is adjacent to seven glycosyl transferases (SP1758, SP1764 to SP1767, SP1770, and SP1771) that could make O-linked glycosylations on the serines in SP1772. This would produce a structure similar to mucins that might also coat the surface of the bacterium or interact with host cellular mucins, although some strains of *S. pneumoniae* have been shown not to interact with mucins (37).

Comparative genome hybridizations on DNA microarrays were performed (38) between the TIGR4 isolate and both the R6 noncapsulated laboratory strain and the closely related D39 serotype 2 capsulated strain (39). Nine gene clusters in the TIGR4 isolate did not hybridize with the other two strains [Fig. 1 and Web table 6 (9)], which suggests that they are absent or significantly divergent in strains R6 and D39. Six of these regions display an atypical nucleotide composition [Fig. 1 and Web table 7 (9)] (40), which suggests that they were horizontally acquired by the TIGR4 isolate. These include the capsule biosynthesis locus (SP0347 to SP0353), the V-type ATPase locus (SP1315 to SP1322), a gene cluster encoding a cell wall surface anchor protein (SP1772) and seven glycosyl transferases, and a putative macrolide efflux protein (SP0168). In addition to these regions, strains R6 and D39 also lack three putative sortases and two sortase motif proteins (SP0463 to SP0468), as well as choline-binding protein I (SP0069) and an IgA1 protease paralog (SP0071). Similar differences in the capsule locus, IgA1 protease, and choline-binding protein were identified by Hakenbeck *et al.* (41) by means of an oligonucleotide-based microarray. The majority of the loci that differ between the three strains are surface-exposed and/or related to pathogenesis, and these differences may contribute to differences in virulence and antigenicity between these strains.

The complete genome sequence of *S. pneumoniae* has revealed new insights into the complexity of its biology and metabolism, particularly with regard to the dual role of extracellular enzyme systems to provide essential nutrients while at the same time facilitating the colonization of host tissues. Recent experimental studies based on the preliminary genome sequence of the TIGR4 isolate have revealed new candidate vaccine targets for this species (42). The avail-

ability of the complete genome sequence will provide additional avenues for followup studies on the basic biology and pathogenicity of *S. pneumoniae*.

References and Notes

- O. T. Avery, C. M. MacLeod, M. McCarty, *J. Exp. Med.* **79**, 137 (1944).
- B. Greenwood, *Philos. Trans. R. Soc. London Ser. B* **354**, 777 (1999).
- D. M. Musher, *Clin. Infect. Dis.* **14**, 801 (1992).
- A. Tomasz, *N. Engl. J. Med.* **333**, 514 (1995); G. V. Doern, A. B. Brueggemann, H. Huynh, E. Wingert, *Emerg. Infect. Dis.* **5**, 757 (1999).
- The TIGR4 isolate was previously referred to as JNR.7/87, the label of the clinical isolate [A. L. Bricker, A. Camilli, *FEMS Microbiol. Lett.* **172**, 131 (1999)]; as KNR.7/87 [A. de Saizieu *et al.*, *J. Bacteriol.* **182**, 4696 (2000)]; R. Hakenbeck *et al.*, *Infect. Immun.* **69**, 2477 (2001)]; and as N4 [T. M. Witzmann *et al.*, *Infect. Immun.* **69**, 1593 (2001)]. Midway through the sequencing project, it became evident that one particular bacterial stock was contaminated with *S. gordonii*, because reads from libraries made with DNA derived from this stock were composed entirely of non-*S. pneumoniae* sequences (assessed by using all available *S. pneumoniae* and *S. gordonii* sequences in GenBank) and would not assemble with the *S. pneumoniae* DNA. Because all aspects of the sequencing project are tracked through a relational database [R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995)], the problem was addressed by identifying and removing all the reads from the libraries in question from the project (*S. gordonii* sequences are available on TIGR's Web site www.tigr.org/tdb/s_gordonii.shtml). The *S. pneumoniae* single-colony isolate that was grown for use in all subsequent libraries was named TIGR4.
- Cloning, sequencing, and assembly were as described [W. C. Niernan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4136 (2001)]. Four small insert (~1.5 kb) shotgun libraries were constructed in pUC-derived vectors after random mechanical shearing (nebulization) of genomic DNA, and three large insert (~18 kb) shotgun libraries were constructed in λ -DASH II vectors (Stratagene) after partial Sau 3A digestion of genomic DNA. Sequencing of the small insert libraries was achieved at a success rate of 66%, with an average read length of 518 bp. The first library constructed was nonrandom, but improvement of the construction methods provided subsequent random libraries. In contrast, none of the large insert libraries appeared to be completely random. Sequencing of these yielded the following success rates per library: first, 366 nucleotides (nt) average length, with a success rate of 26%; second, 620 nt at 52%; and third, 597 nt at 66%. In the late stages of closure, the newly engineered TIGR vector pHO52 (a pBR derivative) was used to construct a new large insert (~9 kb) library. Sequencing rates were 508 nt at 48.5% success; these are low values, but the library was substantially more random than the lambda libraries. 40,839 small insert and 3449 large insert end sequences were jointly assembled into 390 contigs larger than 1.5 kb (with 220 sequencing gaps and 170 physical gaps) using TIGR Assembler [G. S. Sutton, O. White, M. D. Adams, A. R. Kerlavage, *Genome Sci. Technol.* **1**, 9 (1995)]. The coverage criteria were that every position required at least double-clone coverage (or sequence from a PCR product amplified from genomic DNA) and either sequence from both strands or with two different sequencing chemistries. The sequence was edited manually with the TIGR Editor, and additional PCR [H. Tettelin, D. Radune, S. Kasif, H. Khouri, S. L. Salzberg, *Genomics* **62**, 500 (1999)] and sequencing reactions were performed to close gaps, improve coverage, and resolve sequence ambiguities. Particularly difficult regions, including SP1772, which contains 540 copies of a 24-bp imperfect repeat, were covered by transposon-assisted sequencing (New England Biolabs pGPS Transposon Kit) and mapping of transposon insertions before assembly.
- I. S. Aaberge, J. Eng, G. Lermark, M. Lovik, *Microb. Pathog.* **18**, 141 (1995).
- Open reading frames (ORFs) likely to encode proteins were predicted by Glimmer [S. L. Salzberg, A. L. Delcher, S. Kasif, O. White, *Nucleic Acids Res.* **26**, 544 (1998); A. L. Delcher, D. Harmon, S. Kasif, O. White, S. L. Salzberg, *Nucleic Acids Res.* **27**, 4636 (1999)]. This program, based on interpolated Markov models, was trained with ORFs larger than 600 bp from the genomic sequence, as well as with the *S. pneumoniae* genes available in GenBank. All predicted proteins larger than 30 amino acids were searched against a nonredundant protein database, as previously described [R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995)]. Frameshifts and point mutations were detected and corrected where appropriate. Remaining frameshifts and point mutations are considered to be authentic and were annotated as "authentic frameshift" or "authentic point mutation." Protein membrane-spanning domains were identified by TopPred [M. G. Claros, G. von Heijne, *Comput. Appl. Biosci.* **10**, 685 (1994)]. The 5' regions of each ORF were inspected to define initiation codons using homologies, position of ribosomal binding sites, and transcriptional terminators. Two sets of hidden Markov models were used to determine ORF membership in families and superfamilies: pfam v5.5 [A. Bateman *et al.*, *Nucleic Acids Res.* **28**, 263 (2000)] and TIGRFAMs 1.0 [D. H. Haft *et al.*, *Nucleic Acids Res.* **29**, 41 (2001)]. Pfam v5.5 hidden Markov models were also used with a constraint of a minimum of two hits to find repeated domains within proteins and mask them. Domain-based paralogous families were then built by performing all-versus-all searches on the remaining protein sequences, using a modified version of a previously described method [W. C. Niernan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4136 (2001)]. The extent of potential lineage-specific gene duplications in this genome was estimated by identification of ORFs that are more similar to other ORFs within the TIGR4 genome than to ORFs from other complete genomes, including those of plasmids, organelles, and phages. All ORFs were searched with FASTA3 against all ORFs from the complete genomes, and matches with a FASTA *p* value of 10^{-5} were considered significant.
- Supplementary Web material is available on Science Online at www.sciencemag.org/cgi/content/full/293/5529/498/DC1.
- C. Fraser *et al.*, *Science* **270**, 397 (1995); F. Kunst *et al.*, *Nature* **390**, 249 (1997); L. Banerjee, personal communication; S. Gill, personal communication.
- A. M. Gasc, L. Kauc, P. Barraille, M. Sicard, S. Goodgal, *J. Bacteriol.* **173**, 7361 (1991).
- H. Lutcke, *Eur. J. Biochem.* **228**, 531 (1995).
- N. R. Pace, J. W. Brown, *J. Bacteriol.* **177**, 1919 (1995).
- M. Riley, *Microbiol. Rev.* **57**, 862 (1993).
- M. Chandler, O. Fayet, *Mol. Microbiol.* **7**, 497 (1993).
- M. R. Oggioni, J. P. Claverys, *Microbiology* **145**, 2647 (1999).
- B. Martin *et al.*, *Nucleic Acids Res.* **20**, 3479 (1992).
- K. Dybvig, R. Sitarman, C. T. French, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 13923 (1998).
- I. T. Paulsen, L. Nguyen, M. K. Sliwinski, R. Rabus, M. H. Saier Jr., *J. Mol. Biol.* **301**, 75 (2000).
- C. M. Fraser *et al.*, *Science* **281**, 375 (1998).
- R. G. Quivey, W. L. Kuhnert, K. Hahn, *Adv. Microb. Physiol.* **42**, 239 (2000).
- N. S. Jakubovics, A. W. Smith, H. F. Jenkinson, *Mol. Microbiol.* **38**, 140 (2000); A. Dintilhac, G. Alloing, C. Granadel, J. P. Claverys, *Mol. Microbiol.* **25**, 727 (1997); A. M. Berry, J. C. Paton, *Infect. Immun.* **64**, 5255 (1996).
- M. Santagati, F. Iannelli, M. R. Oggioni, S. Stefani, G. Pozzi, *Antimicrob. Agents Chemother.* **44**, 2585 (2000).
- S. K. Hollingshead, D. E. Briles, *Curr. Opin. Microbiol.* **4**, 71 (2001).
- W. B. Wood, M. R. Smith, *J. Exp. Med.* **90** (1949).
- G. Alloing, P. de Philip, J. P. Claverys, *J. Mol. Biol.* **241**, 44 (1994); J. P. Dillard, M. W. Vandersea, J. Yother, *J. Exp. Med.* **181**, 973 (1995); B. J. Pearce, A. M. Naughton, H. R. Masure, *Mol. Microbiol.* **12**, 881 (1994); E. Garcia, D. Llull, R. Munoz, M. Mollerach, R. Lopez, *Res. Microbiol.* **151**, 429 (2000).
- J. N. Weiser, in *Streptococcus pneumoniae—Molecular Biology and Mechanisms of Disease*, A. Tomasz, Ed. (Mary Ann Liebert, Larchmont, NY, 2000), pp. 245–252.
- N. J. Saunders *et al.*, *Mol. Microbiol.* **37**, 207 (2000).
- Iterative DNA motifs, including homopolymeric tracts, were searched in the TIGR4 genome sequence using the

- REPEATS program [G. Benson, M. S. Waterman, *Nucleic Acids Res.* **22**, 4828 (1994)]. The minimum length of homopolymeric tracts was set at eight for A and T and at six for G and C; four tandem copies of di- and trinucleotides; and three copies of tetra-, penta-, and hexanucleotides. Heptanucleotides and above were not found in three or more copies, except for the imperfect repeats in SP1772. The ratio of the observed frequency of homopolymeric tracts to their expected frequency was determined by means of Markov chain analysis, as described [N. J. Saunders et al., *Mol. Microbiol.* **37**, 207 (2000)]. It revealed that G or C tracts of 8 bp and A or T tracts of 10 and 11 bp are slightly overrepresented.
30. J. O. Kim et al., *Infect. Immun.* **67**, 2327 (1999).
 31. O. Humbert, M. Prudhomme, R. Hakenbeck, C. G. Dowson, J. P. Claverys, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 9052 (1995).
 32. J. A. Eisen, P. C. Hanawalt, *Mutat. Res.* **435**, 171 (1999).
 33. Putative choline-binding motifs [J. L. Garcia, A. R. Sanchez-Beato, F. J. Medrano, R. Lopez, in *Streptococcus pneumoniae—Molecular Biology and Mechanisms of Disease*, A. Tomasz, Ed. (Mary Ann Liebert, Larchmont, NY, 2000), pp. 231–244] were identified using Pfam hidden Markov model (HMM) PF01473 [A. Bateman et al., *Nucleic Acids Res.* **28**, 263 (2000)]. LPXG-type Gram-positive anchor regions [M. J. Pallen, A. C. Lam, M. Antonio, K. Dunbar, *Trends Microbiol.* **9**, 97 (2001)] were detected by Pfam HMM PF00746 and by a new HMM built with HMMER 2.1.1 [S. R. Eddy, *Bioinformatics* **14**, 755 (1998)] from a new, curated alignment of the surrounding region in *S. pneumoniae*. Candidate lipoprotein signal peptides [S. Hayashi, H. C. Wu, *J. Bioenerg. Biomembr.* **22**, 451 (1990)] were flagged by NH₂-terminal exact matches to the pattern [DERK](6)-[LVIMFW-STAG](2)-[LVIMFYSTAGCQ]-[AGS]-C (35), culled of hypothetical proteins and cytosolic proteins, aligned manually, and used to generate a new HMM. Proteins matching both the HMM and the regular expression are predicted lipoproteins. Putative signal peptides were identified with SignalP [H. Nielsen, J. Engelbrecht, S. Brunak, G. von Heijne, *Protein Eng.* **10**, 1 (1997)].
 34. The NH₂-terminal regions of all proteins predicted to have signal sequences were collected for clustering and alignment with ClustalW and were scrutinized. A HMM based on an edited alignment of 40-residue segments around the (Y/F)SIRK motif found several hundred hits to a nonredundant amino acid database. A more general motif, based on the larger family of YSIRK proteins, is (Y/F)(S/A)(I/L)(R/K)(R/K)xxxGxxS (35).
 35. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
 36. G. Vidgren, I. Palva, R. Pakkanen, K. Lounatmaa, A. Palva, *J. Bacteriol.* **174**, 7419 (1992).
 37. J. Davies et al., *Infect. Immun.* **63**, 2485 (1995).
 38. This method is used to identify genomic differences between the TIGR4 strain and strains R6 and D39. All the predicted genes from the TIGR4 strain were amplified by PCR and arrayed on glass microscope slides as previously described [S. Peterson, R. T. Cline, H. Tettelin, V. Sharov, D. A. Morrison, *J. Bacteriol.* **182**, 6192 (2000)]. Genomic DNA for comparative genome hybridization studies was labeled according to protocols provided by J. DeRisi (www.microarrays.org/pdfs/GenomicDNALabel_B.pdf), except that genomic DNA was not digested or sheared before labeling. Arrays were scanned with a GenePix 4000B scanner from Axon (Union City, CA), and individual hybridization signals were quantitated with TIGR SPOTFINDER [P. Hegde et al., *Biotechniques* **29**, 548 (2000)].
 39. M. D. Smith, W. R. Guild, *J. Bacteriol.* **137**, 735 (1979).
 40. Regions of atypical nucleotide composition were identified by the χ^2 analysis: The distribution of all 64 trinucleotides (trimers) was computed for the complete genome in all six reading frames, followed by the trimer distribution in 2000-bp windows. Windows overlapped by 1500 bp. For each window, the χ^2 statistic on the difference between its trimer content and that of the whole genome was computed. The most atypical regions, with a score of 600 and above, were considered in this analysis.

41. R. Hakenbeck et al., *Infect. Immun.* **69**, 2477 (2001).
42. T. M. Witzemann et al., *Infect. Immun.* **69**, 1593 (2001).
43. We thank M. Heaney, J. Scott, M. Holmes, V. Sapiro, B. Lee, and B. Vincent for software and database support at TIGR; M. Ermolaeva and M. Perlea for specific computer analyses; the TIGR faculty and sequencing core for expert advice and assistance; I. Aaberge (National Institute of Public Health, Oslo, Norway) for providing the initial

clinical isolate labeled JNR.7/87; and G. Zysk and A. Polissi for sharing specific sequence data not deposited in GenBank. Supported in part by the National Institutes of Allergy and Infectious Diseases (grant R01 AI40645-01A1) and the Merck Genome Research Institute (grant MGRI72).

2 April 2001; accepted 4 June 2001

NPAS2: An Analog of Clock Operative in the Mammalian Forebrain

Martin Reick,¹ Joseph A. Garcia,² Carol Dudley,¹ Steven L. McKnight^{1*}

Neuronal PAS domain protein 2 (NPAS2) is a transcription factor expressed primarily in the mammalian forebrain. NPAS2 is highly related in primary amino acid sequence to Clock, a transcription factor expressed in the suprachiasmatic nucleus that heterodimerizes with BMAL1 and regulates circadian rhythm. To investigate the biological role of NPAS2, we prepared a neuroblastoma cell line capable of conditional induction of the NPAS2:BMAL1 heterodimer and identified putative target genes by representational difference analysis, DNA microarrays, and Northern blotting. Coinduction of NPAS2 and BMAL1 activated transcription of the endogenous *Per1*, *Per2*, and *Cry1* genes, which encode negatively activating components of the circadian regulatory apparatus, and repressed transcription of the endogenous *BMAL1* gene. Analysis of the frontal cortex of wild-type mice kept in a 24-hour light-dark cycle revealed that *Per1*, *Per2*, and *Cry1* mRNA levels were elevated during darkness and reduced during light, whereas *BMAL1* mRNA displayed the opposite pattern. In situ hybridization assays of mice kept in constant darkness revealed that *Per2* mRNA abundance did not oscillate as a function of the circadian cycle in NPAS2-deficient mice. Thus, NPAS2 likely functions as part of a molecular clock operative in the mammalian forebrain.

Locomotor activity, body temperature, endocrine hormones, and metabolic rate fluctuate cyclically with a period of 24 hours. The regulatory apparatus that controls circadian rhythm consists of a transcriptional feedback cycle that is evolutionarily conserved in a wide variety of metazoans (1). In mammals, the activating arm of this cycle is executed by a heterodimeric transcription factor composed of the *Clock* and *BMAL1* gene products (2). The Clock:BMAL1 heterodimer binds directly to regulatory sequences of the genes comprising the negative arm of the transcriptional feedback cycle. The negative components of the regulatory apparatus include three period (*Per*) genes and two cryptochrome (*Cry*) genes (3–11), whose products function in a poorly understood manner to inactivate the Clock:BMAL1 heterodimer. The duration of *Per* and *Cry* activity may be modified by a serine-threonine kinase variously termed casein kinase I ϵ or Tau in mam-

mals and Doubletime in flies (12–14). In the absence of entraining influences, this regulatory apparatus oscillates rhythmically at or near the 24-hour light-dark cycle (i.e., 12 hours light, 12 hours dark). Entrainment derived from light, food, temperature, and metabolic activity can advance or delay the central regulatory apparatus such that it is properly adapted to the summation of these external zeitgebers.

The master pacemaker of circadian rhythm resides in the suprachiasmatic nucleus (SCN), a small group of neurons located at the base of the optic chiasma within the central nervous system (15). Classical transplantation experiments have demonstrated that the SCN is necessary and sufficient to specify circadian rhythm (16, 17). Surprisingly, the same molecular clock is operative in sites peripheral to the SCN (11, 18), including cultured mammalian cells of non-neural origin (19).

Neuronal PAS domain protein 2 (NPAS2, also termed MOP4) is a member of the basic helix-loop-helix (bHLH)-PAS domain family of transcription factors. The gene encoding NPAS2 is expressed in a stereotypic pattern of brain nuclei located within the mammalian forebrain (20, 21). Upon positional cloning of

¹Department of Biochemistry, ²Department of Internal Medicine, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390, USA.

*To whom correspondence should be addressed. E-mail: smckni@biochem.swmed.edu