Complete Genome Sequence of *Treponema pallidum*, the Syphilis Spirochete

Claire M. Fraser,* Steven J. Norris, George M. Weinstock, Owen White, Granger G. Sutton, Robert Dodson, Michelle Gwinn, Erin K. Hickey, Rebecca Clayton, Karen A. Ketchum, Erica Sodergren, John M. Hardham, Michael P. McLeod, Steven Salzberg, Jeremy Peterson, Hanif Khalak, Delwood Richardson, Jerrilyn K. Howell, Monjula Chidambaram, Teresa Utterback, Lisa McDonald, Patricia Artiach, Cheryl Bowman, Matthew D. Cotton, Claire Fujii, Stacey Garland, Bonnie Hatch, Kurt Horst, Kevin Roberts, Mina Sandusky, Janice Weidman, Hamilton O. Smith, J. Craig Venter

The complete genome sequence of *Treponema pallidum* was determined and shown to be 1,138,006 base pairs containing 1041 predicted coding sequences (open reading frames). Systems for DNA replication, transcription, translation, and repair are intact, but catabolic and biosynthetic activities are minimized. The number of identifiable transporters is small, and no phosphoenolpyruvate: phosphotransferase carbohydrate transporters were found. Potential virulence factors include a family of 12 potential membrane proteins and several putative hemolysins. Comparison of the *T. pallidum* genome sequence with that of another pathogenic spirochete, *Borrelia burgdorferi*, the agent of Lyme disease, identified unique and common genes and substantiates the considerable diversity observed among pathogenic spirochetes.

Venereal syphilis was first reported in Europe in the late 1400s (*1*), coincident with the return of Columbus from the New World. The disease quickly reached epidemic proportions in Europe and spread across the world during the early 16th century with the age of exploration. Syphilis was ubiquitous by the 19th century and has been called the acquired immune deficiency syndrome of that era (*2*). Syphilis is characterized by multiple clinical stages and long periods of latent, asymptomatic infection. The primary infection is localized, but organisms rapidly disseminate and cause manifestations throughout the body, including the cardiovascular and nervous systems (*3*). Although effective therapies have been available since the introduction of penicillin in the mid-20th century, syphilis remains an important global health problem.

Treponema pallidum is the causative agent of syphilis. It is a spirochete, a helical to sinusoidal bacterium with outer and cytoplasmic membranes, a thin peptidoglycan layer, and flagella that lie in the periplasmic space and extend from both ends toward the middle of the organism. Recent pulsed-field gel electrophoresis studies (*4*) have shown that *T. pallidum* contains a circular chromosome of about 1000 kilobase pairs, making it one of the smallest prokaryotic genomes. Despite its importance as an infectious agent, relatively little is known about *T. pallidum* in comparison with other bacterial pathogens (*5*). The organism is an obligate human parasite that cannot be cultured continuously in vitro (*6*). Mechanisms of *T. pallidum* pathogenesis are poorly understood. No known virulence factors have been identified, and the outer membrane is mostly lipid with a paucity of proteins (*7*). Consequently, existing diagnostic tests for syphilis are suboptimal, and no vaccine against *T. pallidum* is available.

Spirochetes represent a phylogenetically ancient and distinct bacterial group. Both *T. pallidum* and *Borrelia burgdorferi*, the causative agent of Lyme disease, are similar in having relatively small genomes and surviving only in association with a host. However, they are not closely related and probably evolved independently from a more complex ancestor by loss of unnecessary genes and acquisition of new functions that promoted survival in the host environment. Comparison of the *T. pallidum* and *B. burgdorferi* genomes (*8*) allows assessment of biological diversity within this group of bacteria.

Genome analysis. The genome of *T. pallidum* subsp. *pallidum* (Nichols) was sequenced by the whole genome random sequencing method as described (*8–10*). The *T. pallidum* genome is a circular chromosome of 1,138,006 base pairs with an average G + C content of 52.8% (Figs. 1 and 2). There are a total of 1041 predicted open reading frames (ORFs), with an average size of 1023 bp, representing 92.9% of total genomic DNA. Predicted biological roles were assigned to 577 ORFs (55%) by the classification scheme adopted from Riley (*11*); 177 ORFs (17%) match hypothetical proteins from other species, and 287 ORFs (28%) have no database match and presumably represent novel genes (Fig. 1 and Table 1). Ninety *T. pallidum* ORFs of unknown function match chromosome-encoded proteins in *B. burgdorferi* (*8*); however, no *T. pallidum* ORFs match *B. burgdorferi* plasmid-encoded proteins, suggesting that the plasmid proteins may be unique to *Borrelia* species (*8*). The average size of the predicted proteins in *T. pallidum* is 37,771 daltons, ranging from 3235 to 172,869 daltons, and the mean isoelectric point for all predicted proteins is 8.1, ranging from 3.9 to 12.3, values similar to those observed in other bacterial species (*8*, *9*).

Forty-two paralogous gene families containing a total of 129 ORFs (12%) were identified in *T. pallidum* (Fig. 1). Fifteen families contain 44 genes that have no assigned biological role. Thirty families have only two members. The largest family, with 14 members, consists of proteins with adenosine triphosphate (ATP)–binding cassettes in ABC transport systems. Within 13 gene families are 16 clusters of adjacent genes that may represent duplications in the *T. pallidum* genome.

All 61 triplet codons are used in *T. pallidum*. There is a bias for G or C in the third codon position in *T. pallidum*, in contrast to an A or T bias in this position in *B. burgdorferi*. This observation is consistent with the G + C content in the *T. pallidum* genome being almost twice that in the *B. burgdorferi* genome. The disparate G + C content between the spirochete genomes creates a bias in overall codon usage, resulting in a difference in amino acid composition in the predicted coding sequences.

C. M. Fraser, O. White, G. G. Sutton, R. Dodson, M. Gwinn, E. K. Hickey, R. Clayton, K. A. Ketchum, S. Salzberg, J. Peterson, H. Khalak, D. Richardson, T. Utterback, L. McDonald, P. Artiach, C. Bowman, M. D. Cotton, C. Fujii, S. Garland, B. Hatch, K. Horst, K. Roberts, M. Sandusky, J. Weidman, H. O. Smith, and J. C. Venter are with The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. S. J. Norris, G. M. Weinstock, E. Sodergren, J. M. Hardham, M. P. McLeod, J. K. Howell, and M. Chidambaram are at the University of Texas Health Science Center, Departments of Microbiology and Molecular Genetics, Pathology and Laboratory Medicine, and the Center for the Study of Emerging and Reemerging Pathogens, Post Office Box 20708, Houston, TX 77225, USA.

*To whom correspondence should be addressed. E-mail: tpdb@tigr.org

Origin of replication. Two criteria were used to identify a replication origin in *T. pallidum*: the co-localization of genes (*dnaA*, *dnaN*, *recF*, and *gyrA*) often found near the origin in prokaryotic genomes and GC skew (*12*) (Fig. 2). On the basis of these results, we designated base pair 1 of the *T. pallidum* genome in an intergenic region of the chromosome that is located within the putative origin of replication.

Sixty-four percent of the coding sequences in the *T. pallidum* genome are aligned in the direction of replication, with the point of transcriptional divergence located near the putative origin between *clpP* and *dnaA* (Figs. 1 and 2). A number of codons occur in coding sequences aligned in the direction of replication at significantly higher frequencies than expected ($P < 5.3e^{-27}$), including TTG (Leu), GCG (Ala), CGT (Arg), GTG (Val), and TGT (Cys). Codons that are overrepresented are also found in the most highly skewed oligomers (GGAGCGTG, TGTGTGTG, GTGTGTGC, TTTTTTGT, and GGTGTGTG).

Codon adaptation index (CAI), which is designed to be a relative measure of translational efficiency (*13*), was computed for *T. pallidum* with the codon frequencies from the ribosomal proteins, the translation elongation factors, and glyceraldehyde-3-phosphate dehydrogenase. Proteins with a high CAI are presumably highly expressed in exponential growth (*13*). The distribution of CAI scores in *T. pallidum* ORFs (*13*) exhibits a strand-dependent switch in magnitude around the origin of replication (Fig. 2). In both *T. pallidum* and *B. burgdorferi*, there is a marked difference in CAI values (high versus low) for genes on opposite strands of the chromosome, with genes transcribed in the direction of replication exhibiting a high CAI.

Transcription and translation. *Treponema pallidum* contains a basic set of genes for transcription and translation that includes homologs to the α, β, and β′ subunits of the core RNA polymerase, five sigma factors (σ^{24}, σ^{28}, σ^{43}, σ^{54}, and σ^{70}), and five genes that encode proteins involved in transcript elongation and termination (*nusA*, *nusB*, *nusG*, *greA*, and *rho*). *Treponema pallidum* is missing both a recognizable σ^{38} (*rpoS*), which is the major sigma factor in stationary phase activated in response to oxidative and osmotic stress, and a σ^{32}, which is involved in transcription of heat shock proteins.

Forty-four tRNA species, organized into eight clusters containing 25 genes plus 19 single genes, were identified (Figs. 1 and 2). Two ribosomal RNA (rRNA) operons are present in the genome. Their organization is the same as that commonly found in eubacteria (16*S*-tRNA-23*S*-5*S*) (*14*), in contrast with the unusual arrangement seen in *B. burgdorferi* (*8*, *15*). Both *T. pallidum* rRNA operons are transcribed in the direction of replication.

All tRNA synthetase genes were identified with the exception of glutaminyl-tRNA synthetase, similar to *B. burgdorferi* (*8*). It is likely that glutamyl-tRNA synthetase aminoacylates $tRNA^{Gln}$ with glutamate followed by transamidation by Glu-tRNA amidotransferase (*16*). Two distinct lysyl-tRNA synthetase (LysS) species are present in *T. pallidum*, a class I type most similar to those in euryarchaea and *B. burgdorferi* (*17*) and a class II type most similar to those in eubacteria and eukaryotes. The class II LysS in *T. pallidum* represents a COOH-terminal fragment of *Escherichia coli* LysS. A region near the NH_2-terminus of LysS binds the anticodon of the tRNA and is crucial for its activity (*18*). Thus, it is likely that the class II LysS is nonfunctional and may be in the process of being lost from the genome.

Replication, repair, recombination, and restriction-modification systems. The complement of genes for DNA replication in *T. pallidum* is similar to that in other minimal genomes such as *Mycoplasma genitalium* and *B. burgdorferi*. Orthologs for the α, β, ε, γ, and τ subunits of *E. coli* DNA polymerase III are present. *Treponema pallidum* has homologs of one type I topoisomerase (*topA*) and one type II topoisomerase (*gyrAB*), but unlike *B. burgdorferi*, it is missing topoisomerase IV, which is involved in chromosome segregation. However, chromosome segregation in *T. pallidum* may proceed by an alternative mechanism that involves the binding of hemimethylated DNA to the cytoplasmic membrane. This idea is supported by the presence of DNA adenine methyltransferase (*dam*) in *T. pallidum* but not in *B. burgdorferi* (*19*).

DNA repair in *T. pallidum* includes the major known pathways of *uvr* excision repair, *mutL/mutS* mismatch repair, *mutY*, and *dat*. The *T. pallidum* genome encodes homologs of the *recF* pathway of recombination (*recFGJNR*) but lacks homologs to *sbcB* (*exoI*) as well as *recB*, *recC*, and *recD*. Thus, homologous recombination resembles the recF pathway of *E. coli*. The converse is true in *B. burgdorferi*, where there are homologs of *recBCD* but not the *recF* pathway genes (*8*). *Treponema pallidum* contains an A- or G-specific adenine glycosylase (*mutY*), recognizes GA mismatches in duplex DNA, and excises adenine. No enzyme with similar activity has been identified in either *M. genitalium* or *B. burgdorferi*. This difference may, in part, explain the lower G + C content of the *B. burgdorferi* and *M. genitalium* genomes as compared with *T. pallidum*. No recognizable genes encoding restriction or modification enzymes were found.

Biosynthetic pathways. *Treponema pallidum* is an obligate parasite of humans. Consistent with this property, previous physiologic studies have shown that it has limited biosynthetic capabilities and requires multiple nutrients from the host (*20*) (Fig. 3). The *T. pallidum* genome encodes a pathway for the conversion of phosphoenolpyruvate or pyruvate through oxaloacetate to aspartate (at the expense of glutamate), in accordance with the previous observation that most of the [^{14}C]glucose incorporated into amino acids was in the form of aspartate (*21*). Predicted pathways for the interconversion of aspartate and glutamine to glutamate, aspartate to asparagine, glutamate to proline, and serine to glycine are also present. *Treponema pallidum* is unable to synthesize enzyme co-factors, fatty acids, and nucleotides de novo, similar to *M. genitalium* and *B. burgdorferi*. Deoxyribonucleotides can be obtained by reduction of ribonucleoside diphosphates through the action of ribonucleotide diphosphate reductase and thioredoxin reductase.

Transport. An organism such as *T. pallidum*, with limited biosynthetic capabilities, must have a repertoire of transport proteins with broad substrate specificity to obtain the necessary nutrients from the environment. The *T. pallidum* genome contains 57 ORFs (5% of the total) that encode 18 distinct transporters with predicted specificity for amino acids, carbohydrates, and cations (Fig. 3 and Table 1). For the most part, these transport systems are of similar specificity to those found in *M. genitalium* and *B. burgdorferi* (*8*); however, several important differences are seen.

Treponema pallidum has a broad spectrum of amino acid transporters, although these transporters are different from those in *B. burgdorferi*. For example, a transporter for glutamate or aspartate in *T. pallidum* is most similar to mammalian glutamate transporters. There are no phosphoenolpyruvate:phosphotransferase (PTS) systems in *T. pallidum* for the import of carbohydrates, in contrast to other bacterial species whose genome sequences have been determined (*8*, *9*). Genome analysis predicts that *T. pallidum* has three ATP-binding cassette transporters with specificity for galactose (*mglBAC*) (*22*, *23*), ribose (*rbsAC*), and multiple sugars (*y4oQRS*), respectively; however, these three transporters may display a broader substrate specificity. In *E. coli*, the *mgl* transporter displays affinity not only for galactose but also for glucose (*24*), and its expression is up-regulated in glucose-limiting conditions but repressed at high glucose concentrations (*24*). *Treponema pallidum* may also require an environment with limiting glucose concentrations for maximal expression of this transporter. *Treponema pallidum* has no recognizable inorganic phosphate (P_i) uptake system, unlike other bacteria studied by whole-genome analysis to date; therefore, up-

take of glycerol-3-phosphate through the multiple sugar transporter may represent the primary means whereby *T. pallidum* obtains P_i (Fig. 3).

Treponema pallidum contains an ATP-binding cassette transporter with specificity for thiamine. Both thiamine and thiamine pyrophosphate (TPP) are substrates for the thiamine transporter in *E. coli* (*25*). This finding is of interest because *T. denticola*, *T. vincentii*, and *Leptospira* species require TPP for growth in vitro (*26*), which suggests that *T. pallidum* may also exhibit a growth dependency on TPP. The only recognizable TPP-dependent enzyme present in *T. pallidum* is transketolase, which creates a link between the pentose phosphate pathway and glycolysis.

Energy metabolism. The complement of transport proteins in *T. pallidum* suggests that it may use several carbohydrates as energy sources, including glucose, galactose, maltose, and glycerol. Experimental evidence has demonstrated that only glucose, mannose, and maltose support the multiplication of *T. pallidum* in a tissue culture system (*27*). It is not known whether *T. pallidum* can use amino acids as a source of carbon and energy; however, the lack of necessary catabolic and anabolic pathway genes suggests that it would not be able to use such alternative compounds.

Metabolic pathway analysis reveals that genes encoding all of the enzymes of the glycolytic pathway are present in *T. pallidum*, including hexokinase, which phosphorylates glucose and other hexose sugars (Fig. 3). Both *M. genitalium* and *B. burgdorferi* lack hexokinase; however, in these organisms, phosphorylation of hexoses is an integral part of the PTS uptake mechanism. Instead of the typical eubacterial phosphofructokinase and pyruvate kinase, *T. pallidum* contains homologs of these enzymes that use pyrophosphate. Similar inorganic pyrophosphate (PP_i)–dependent enzymes have been described in some bacteria, protists, protozoa, and plants (*28*). None of the genes encoding components of the tricarboxylic acid cycle or oxidative phosphorylation were identified, contrary to previous reports of the presence of cytochromes, flavoproteins, and some of the tricarboxylic acid cycle enzymes (*20*, *29*); these may have represented contaminating rabbit components. Reducing power is probably generated through the oxidative branch of the pentose phosphate pathway. This simplified metabolic strategy is similar to that seen in both *M. genitalium* (*9*) and *B. burgdorferi* (*8*).

Treponema pallidum, like *B. burgdorferi*, lacks a respiratory electron transport chain; therefore, ATP production must be accomplished by substrate-level phosphorylation. As a result, membrane potential must be established by the reverse reaction of the ATP

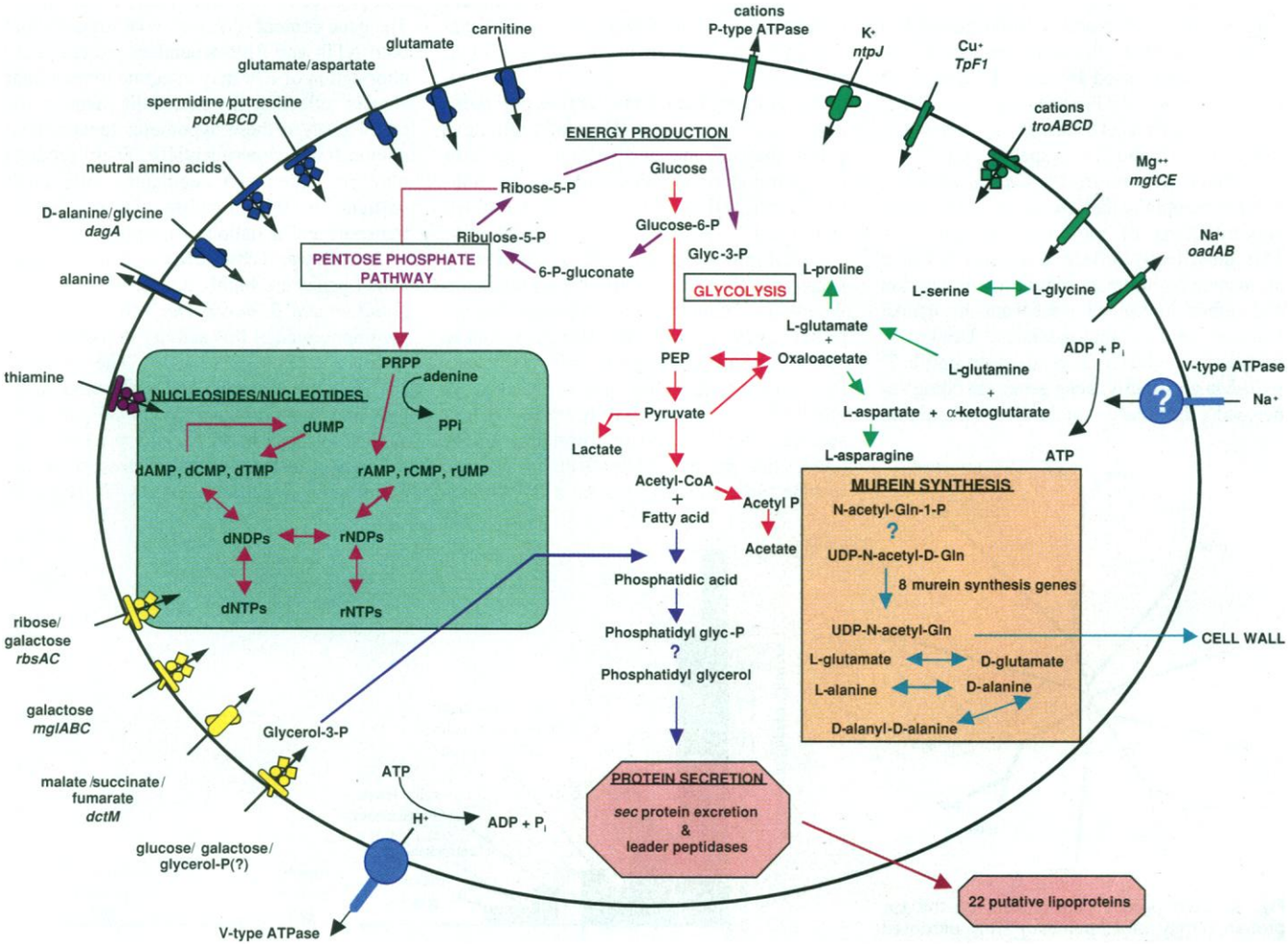

Fig. 3. Solute transport and metabolic pathways in *T. pallidum*. A schematic diagram of a *T. pallidum* cell providing an integrated view of the transporters and the main components of the metabolism of this organism, as deduced from the genes identified in the genome. Presumed transporter specificity is indicated. Question marks indicate where particular uncertainties exist or expected activities were not found. r, ribo; d, deoxy; AMP, adenosine monophosphate; CMP, cytosine monophosphate; NDP, nucleotide diphosphate; NTP, nucleotide triphosphate; TMP, thymidine monophosphate; UMP, uridine monophosphate; ADP, adenosine diphosphate; CoA, coenzyme A; UDP, uridine diphosphate; PRPP, phosphoribosyl-pyrophosphate.

synthase. In both spirochetes, the ATP synthase is of the V_1V_0 type, most similar to those found in eukaryotic vacuoles and in archaea (*30*). *Treponema pallidum* has two V_1V_0-type ATP synthase operons, each containing seven genes (Table 1). The gene order in one operon (subunit E–ORF–subunit A–subunit B–subunit D–subunit I–subunit K) is identical to that seen in the ATP synthase operon in *B. burgdorferi* (*8*). The second operon in *T. pallidum* contains ATP synthase subunits A, B, D, E, F, I, and K. The difference in subunit gene composition between these operons suggests that the ATP synthases may have different functions in the cell.

One clue as to the functional role for the two ATP synthases is the presence of an oxaloacetate decarboxylase transporter that may be involved in extrusion of Na^+ from the cell, creating a Na^+ gradient (*31*). Such a gradient can be used to drive Na^+-dependent transporters similar to the amino acid transporters that are found in *T. pallidum*. Alternatively, the Na^+ gradient could be used to synthesize ATP in the same manner as the H^+ gradient is used by an F_1F_0-type ATP synthase. Two ATP synthases have been identified in *Enterococcus hirae*, with specificity for Na^+ and H^+, respectively (*32*).

Cellular processes. *Treponema pallidum* is microaerophilic and grows only at reduced concentrations of molecular oxygen (*33*). This most likely reflects a balance between an oxygen requirement for energy production and defects in protective mechanisms against reactive oxygen intermediates. Unlike *B. burgdorferi*, which is also microaerophilic, *T. pallidum* apparently lacks genes encoding superoxide dismutase, catalase, or peroxidase activities that protect against oxygen toxicity. NADH oxidase is the only enzyme identified thus far that can account for O_2 utilization by *T. pallidum*.

Treponema pallidum contains a basic set of heat shock proteins but lacks σ^{32}, which is responsible for transcription of heat shock genes in other bacteria. This lack is consistent with previous reports that *T. pallidum* lacks a detectable heat shock response. There is no change in the amounts of GroEL or other proteins at increased temperatures (*34*). At least two heat shock proteins in *T. pallidum* (GroEL and DnaK) appear to be constitutively expressed at high levels, which may mitigate the need for a typical heat shock response (*35*). However, the observed thermal sensitivity of *T. pallidum* (*6*) may reflect the absence of a robust heat shock response in this organism. It is of interest that *B. burgdorferi*, which also lacks a recognizable σ^{32}, exhibits a heat shock response. The differential response of the two spirochetes to increased temperatures suggests that a protein or proteins of unknown biological function may be involved in this process in *B. burgdorferi*.

Regulatory functions. *Treponema pallidum* contains a minimal set of regulatory genes that encode two response-regulator two-component systems and several putative transcriptional repressors of unknown specificity.

Although *T. pallidum* does not have a sugar-specific PTS system, it does contain a homolog of enzyme I (*ptsI*), a phosphocarrier protein HPr (*ptsH*), an HPr(Ser) kinase (*ptsK*), and two *ptsN* genes, which suggests that these proteins may function mainly as regulators. Gram-positive bacteria have a specific ATP-dependent protein kinase (*ptsK*) that phosphorylates HPr on a serine residue (*36*). HPr(Ser~P) and a DNA-binding protein then interact to mediate repression by binding specifically to DNA sequences, catabolite responsive elements found in the control regions of catabolite-sensitive operons (*36*). These proteins in *T. pallidum* may function in a manner similar to that observed in Gram-positive bacteria.

Escherichia coli and other Gram-negative organisms coordinate nitrogen and carbon utilization so that mechanisms of carbon repression do not block the uptake and use of organic nitrogen sources. Nitrogen-carbon utilization in *E. coli* is modulated by a regulatory protein, PtsN, that displays similarity to the PTS enzymes IIA specific for fructose and mannitol (*37*). Biochemical data suggest that PtsN does not phosphorylate carbohydrates but instead serves as a positive regulator of organic nitrogen metabolism. Under such conditions, phosphoenolpyruvate (PEP)-dependent phosphorylation of PtsN occurs through the transfer of a phosphate group from PEP to enzyme I, then to a histidine residue on HPr, and finally to PtsN (*37*). The gene content of *T. pallidum* suggests that both ATP- and PEP-dependent protein phosphorylation of HPr may integrate intracellular signals reflecting the metabolic state of the cell. However, these hypotheses remain to be demonstrated experimentally. These proteins may play alternative regulatory roles in *T. pallidum* as this organism displays limited transport and metabolic capacities.

Motility and chemotaxis. Motility-associated genes are highly conserved in both *T. pallidum* and *B. burgdorferi*, consistent with the importance of this activity in these highly invasive spirochetes (*23*, *38*). The 36 genes encoding proteins involved in flagellar structure and function in *T. pallidum* are most similar to those in *B. burgdorferi* (*8*). They differ only in the number of proteins in the periplasmic flagellar filaments; *T. pallidum*

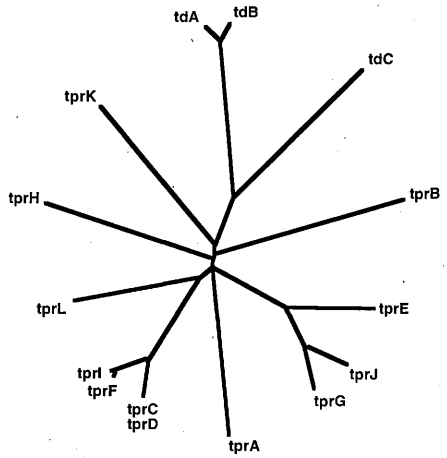

Fig. 4. Dendrogram of members of the tpr protein family of *T. pallidum*. This unrooted distance dendrogram was generated from a multiple sequence alignment of the 12 *T. pallidum* tpr paralogs described herein and Msp sequences from three *Treponema denticola* strains (tdA, *T. denticola* OTK; tdB, *T. denticola* 35405; tdC, *T. denticola* 33520) (*42*).

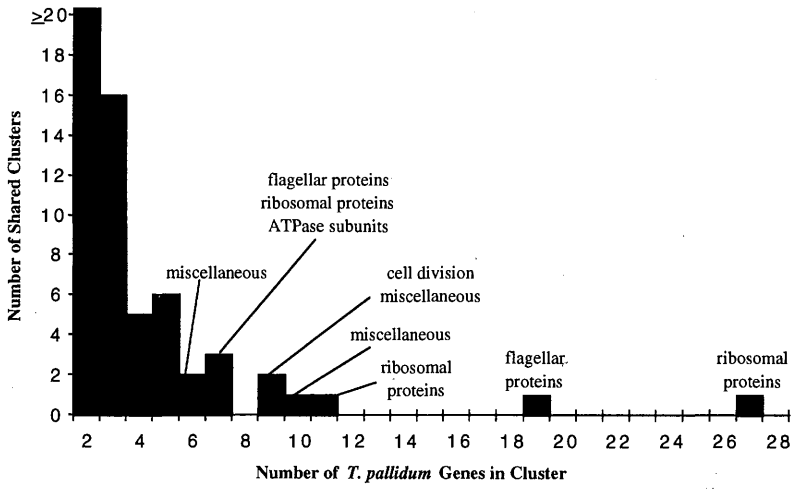

Fig. 5. Gene clusters found in both the *T. pallidum* and *B. burgdorferi* genomes. Graph values represent the number of *T. pallidum* gene clusters in which orthologous genes have the same organization in *B. burgdorferi*.


1 kb
Amino acid biosynthesis
Biosynthesis of cofactors, prosthetic groups, carriers
Cell envelope
Cellular processes
Central intermediary metabolism
Energy metabolism
Fatty acid/Phospholipid metabolism
Purines, pyrimidines, nucleosides and nucleotides
Regulatory functions
DNA metabolism
Transport/binding protei
Translation
Transcription
Other categories
Conserved hypothetical
Unknown


1 kb
Transport/binding proteins
Translation
Transcription
Other categories
Conserved hypothetical
Unknown
Signal peptide
Lipoprotein
Transporter
GES region
Paralogous gene family
Authentic Frame Shift
23S
16S
5S

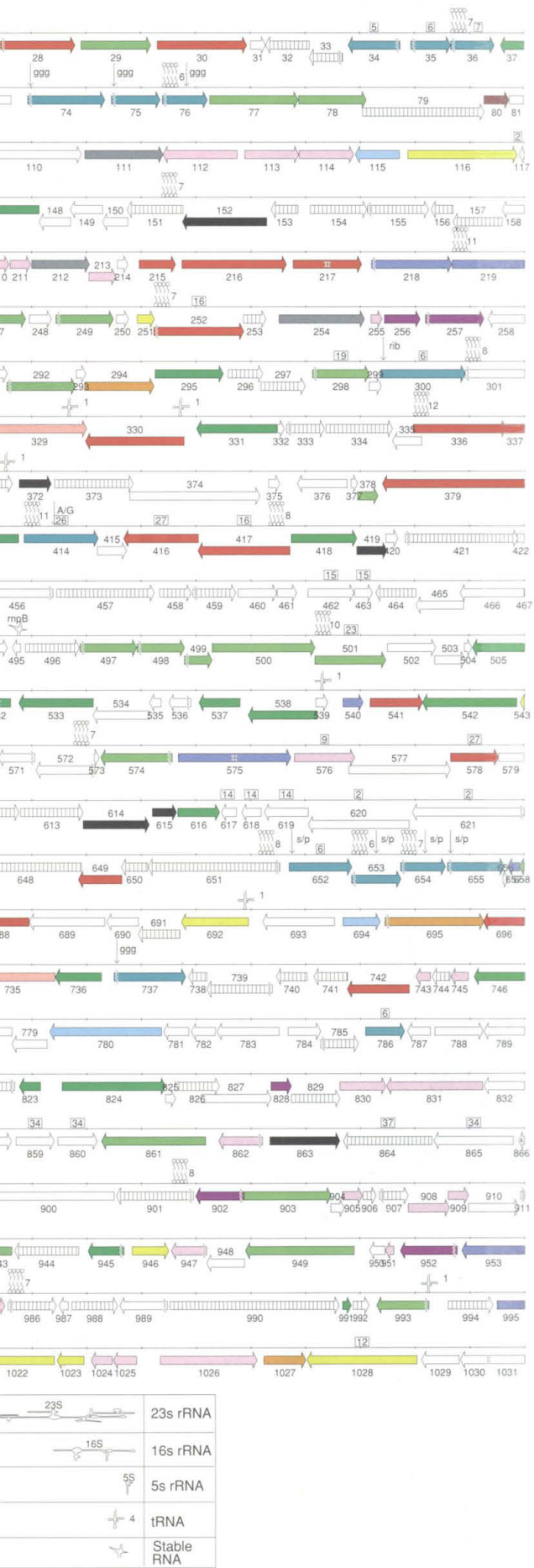

Fig. 1. Linear representation of the *T. pallidum* chromosome. The locations of predicted coding regions color-coded by biological role, RNA genes, and tRNAs are indicated. Arrows represent the direction of transcription for each predicted coding region. Numbers associated with tRNA symbols represent the numbers of tRNAs at a given locus. Numbers associated with GES represent the number of putative MSDs according to the Goldman, Engleman, and Steitz scale as predicted by TopPred (*10*). Only proteins with five or more GES are indicated. Members of paralogous gene families are identified by family number above the predicted coding region. Presumed transporter specificity is indicated. Transporter abbreviations: aa, amino acids; aaX, oligopeptides; car, carnitine; fum, malate/succinate/fumarate; ggg, glucose/galactose/glycerol-3-P; glu, glucose/galactose; rib, ribose/galactose; s/p, spermidine/putrescine; B1, thiamine; A, alanine; A/G, alanine/glycine; E, glutamate; E/D, aspartate/glutamate; ++, cations.

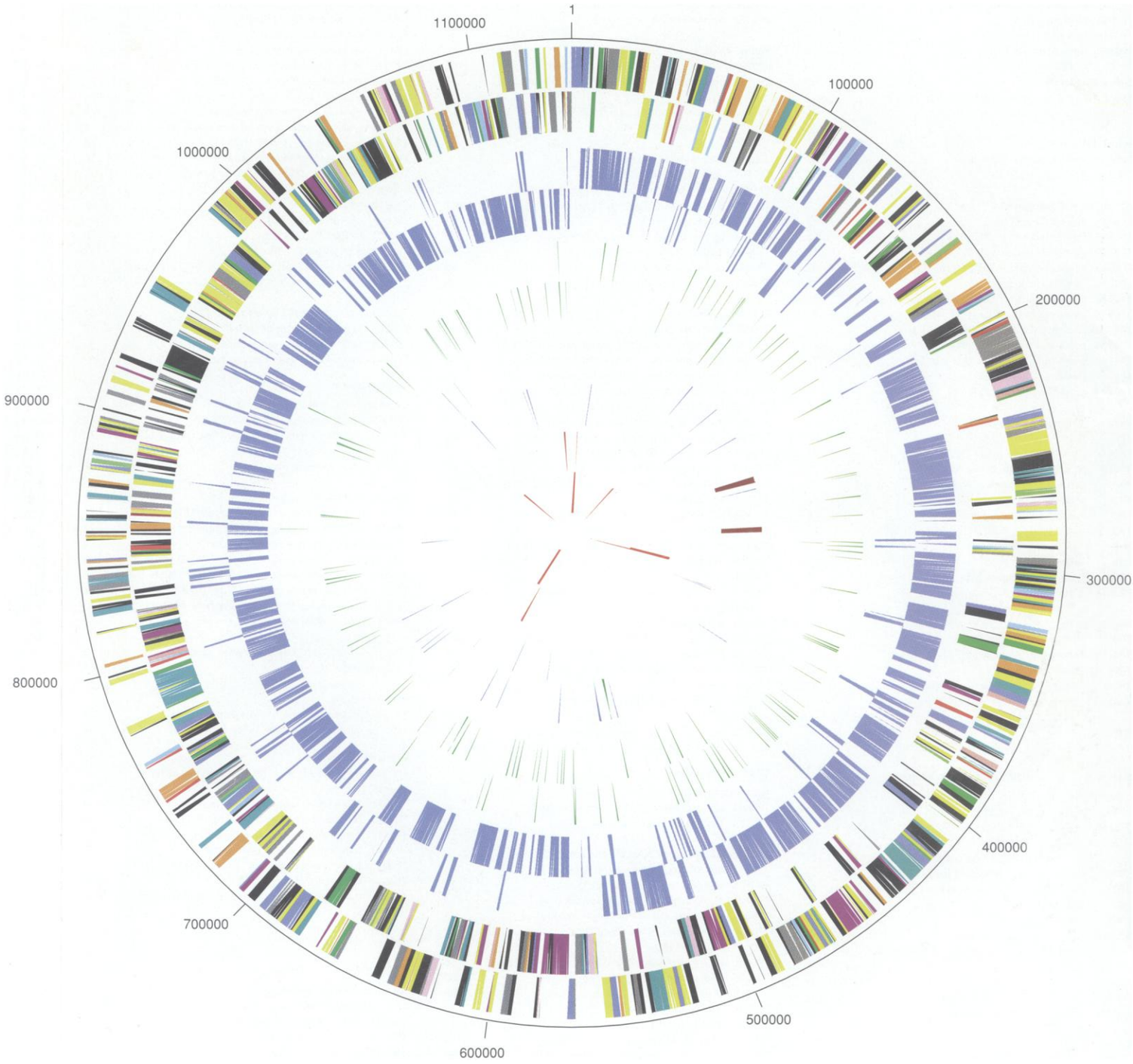

Fig. 2. Circular representation of the *T. pallidum* chromosome illustrating the location of each predicted coding region and RNA in the genome. Outer concentric circle: predicted coding regions on the + strand classified as to role according to the color code in Fig. 1 (with the exception that unknowns and hypotheticals are colored black). Second concentric circle: predicted coding regions on the - strand. Third and fourth concentric circles: CAI on the + and - strand, respectively. Fifth and sixth concentric circles: oligonucleotide skew on the + and - strand, respectively. Seventh and eighth concentric circles: tRNAs (blue), rRNAs (red), and small nuclear RNAs (green) on the + and - strand, respectively. Ninth and tenth concentric circles: *tpr* genes on the + and - strand, respectively.

has three core proteins (FlaB1, FlaB2, and FlaB3), a sheath protein (FlaA), and two uncharacterized proteins (*39*), and *B. burgdorferi* has a single core protein and sheath protein, whereas most other bacteria have only a core protein. Both spirochetes contain two copies of the flagellar motor switch protein, FliG; however, the importance of this gene duplication is not known. Most of the flagellar genes in *T. pallidum* are found in four operons that contain between 2 and 16 genes, most similar to the arrangement seen in *B. burgdorferi*. *Treponema pallidum* has retained a σ^{28} ortholog and class II and class III motility promoters, whereas motility genes in *B. burgdorferi* appear to be transcribed through σ^{70} initiation (*8*, *40*). *Treponema pallidum* contains 13 chemotaxis genes that include four methyl-accepting chemotaxis proteins with putative specificity for amino acids (aspartate, glutamate, and histidine) or carbohydrates (glucose, ribose, and galactose).

Membrane proteins and lipoproteins. Freeze fracture studies (*7*) have shown that the outer membrane of *T. pallidum* contains a relatively small number of integral membrane proteins, a feature that may permit the organism to evade the human immune response. Two candidate outer membrane proteins have been identified, but the cellular location and function of these proteins are a subject of some controversy (*41*). Although it is difficult to identify outer membrane proteins with certainty, genome analysis of *T. pallidum* indicates the presence of only 22 putative lipoproteins, as compared with 105 in *B. burgdorferi*, consistent with results from ultrastructural studies.

Potential virulence factors. *Treponema pallidum* contains a large family of duplicated genes (paralogs) (*tpr*A-L) that encode putative membrane proteins that may function as porins and adhesins (Figs. 2 and 4). This hypothesis is based on pair-wise and multiple sequence alignments of the *T. pallidum* gene family to a major outer sheath protein (Msp) from *T. denticola* that represents an abundant, highly immunogenic, pore-forming adhesin in the outer membrane (*42*). It is not yet known whether the *tpr* genes are expressed individually or coordinately or to what extent each gene is expressed. This gene family in *T. pallidum* is reminiscent of a 32-member paralogous gene family encoding outer membrane proteins in *Helicobacter pylori* (omp) (*9*). The two gene families share features besides possible porin and adhesin functions, the most striking being that both have members with regions of extensive sequence identity. However, in both organisms, the homologous regions do not always encompass the entire gene, so that some regions are identical, but others are variable. As in the *H. pylori* family, one of the *T. pallidum* genes (*tpr*A and L) contains a frameshift within a small dinucleotide repeat that might be corrected by slipped-strand mispairing. Multiple copies of the *tpr* genes may represent a mechanism for generation of antigenic variation in *T. pallidum* as is found in other pathogenic bacteria, including *Neisseria gonorrhoeae*, *M. genitalium*, relapsing fever borreliae, and *B. burgdorferi*. Identification of the *tpr* family of putative outer membrane proteins may provide new targets for vaccine development.

Previous studies have indicated that *T. pallidum* does not produce lipopolysaccharide or potent exotoxins, although cytotoxic activity against neuroblasts and other cell types has been observed at extremely high concentrations of the bacterium (*43*). Genome analysis has revealed five genes encoding proteins similar to bacterial hemolysins. These putative hemolysin orthologs also share varying degrees of amino acid sequence similarity with *B. burgdorferi* predicted proteins (*8*). None of the predicted hemolysins with similarity to *T. pallidum* sequences have been shown to be cytolytic in their purified state, and it will be necessary to perform such studies with the *T. pallidum* proteins before a cytotoxic function can be assigned definitively. A *B. burgdorferi* protein with hemolytic activity (BlyA) was described recently (*44*), but orthologous sequences are not present in *T. pallidum*.

Comparative genomics. Four hundred seventy-six ORFs in *T. pallidum* (46%) have orthologs in *B. burgdorferi*; 76% of these ORFs have a predicted biological function. More than 40% of the orthologous genes in *T. pallidum* and *B. burgdorferi* are highly conserved in other bacteria (*8*, *9*) and are involved in housekeeping functions such as transcription, translation, DNA replication, basic energy metabolism, flagellar structure and function, cell division, and protein secretion. Some of the genes of unknown function that are conserved in the spirochetes but not recognized in the other available genome sequences are likely to represent "spirochete-specific" genes that contribute to the unusual structural properties of these bacteria.

One hundred fifteen ORFs shared by *T. pallidum* and *B. burgdorferi* encode proteins of unknown biological function; and almost 50% of these appear to be unique to the spirochete group. This set of proteins with a limited phylogenetic distribution may include important determinants of spirochete structure and physiology and may, for example, be involved in the ability of both *T. pallidum* and *B. burgdorferi* to infect humans and cause chronic, disseminated disease.

Three hundred four of the ORFs shared by the two spirochetes are located in gene clusters with conserved gene order (Fig. 5). Several conserved clusters contain ORFs encoding ribosomal proteins, including the largest cluster containing 27 ORFs, whereas other important clusters encode proteins for the flagella, adenosine triphosphatases (ATPases), and cell division, as well as groups of proteins that are not obviously related. Further study of the arrangement of these clusters in the two genomes may provide insight into the evolution of the chromosomes of these organisms.

Of the 572 *T. pallidum* ORFs (54% of total) that are not shared with *B. burgdorferi*, more than 80% are of unknown biological function. This finding lends support to the concept of diversity within a single group of bacteria and underscores the fact that a considerable amount of *T. pallidum* biology is yet to be elucidated.

Conclusion. *Treponema pallidum* has been a difficult organism to study experimentally because of its absolute dependence on a mammalian host for sustained growth and viability. The genomic sequence of *T. pallidum* offers a wealth of basic information that would be difficult, if not impossible, to obtain by any other approach. A more complete understanding of the biochemstry of this organism derived from genome analysis may provide a foundation for the development of a culture medium for *T. pallidum*, which opens up the possibility of future genetic studies.

References and Notes

1. F. F. Cartwright, *Disease and History* (Dorset, New York, 1972).
2. S. H. Barondes, *Molecules and Mental Illness* (Scientific American Library, New York, 1993).
3. E. N. Robinson Jr. *et al.*, in *Syphilis: Disease with a History. Mechanisms of Microbial Disease*, M. Schaechter, G. Medoff, B. I. Einstein, Eds. (Williams & Wilkins, Baltimore, MD, 1993), pp. 334–342; S. Sell and S. J. Norris, *Int. Rev. Exp. Pathol.* **24**, 204 (1983).
4. E. M. Walker, J. K. Arnett, J. D. Heath, S. J. Norris, *Infect. Immun.* **59**, 2476 (1991); E. M. Walker *et al.*, *J. Bacteriol.* **177**, 1797 (1995).
5. S. J. Norris *et al.*, *Microbiol. Rev.* **57**, 750 (1993).
6. A. H. Fieldsteel, D. L. Cox, R. A. Moeckli, *Infect. Immun.* **32**, 908 (1981); S. J. Norris and D. G. Edmondson, *ibid.* **53**, 534 (1986); D. L. Cox, *Methods Enzymol.* **236**, 390 (1994).
7. J. D. Radolf, M. V. Norgard, W. W. Shulz, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2051 (1989); E. M. Walker, L. A. Borenstein, D. R. Blanco, J. N. Miller, M. A. Lovett, *J. Bacteriol.* **173**, 5585 (1991); D. L. Cox, P. Chang, A. W. McDowall, J. D. Radolf, *Infect. Immun.* **60**, 1076 (1992).
8. C. M. Fraser *et al.*, *Nature* **390**, 580 (1997).
9. R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995); C. M. Fraser *et al.*, *ibid.* **270**, 397 (1995); C. J. Bult *et al.*, *ibid.* **273**, 1058 (1996); J.-F. Tomb *et al.*, *Nature* **388**, 539 (1997); H.-P. Klenk *et al.*, *ibid.* **390**, 364 (1997).
10. The Nichols strain of *T. pallidum* subsp. *pallidum* was originally isolated from the cerebrospinal fluid of a neurosyphilis patient in 1912 [H. A. Nichols and W. H. Hough, *J. Am. Med. Assoc.* **60**, 108 (1913)]. This strain has retained its virulence and appears to be closely related to more recent isolates. Freshly extracted *T. pallidum* subsp. *pallidum* Nichols was inoculated into the testes of adult male New Zealand white rabbits ($\sim 5 \times 10^7$ *T. pallidum* per testis) that were housed at 16° to 18°C and were provided antibiotic-free feed and water. Animals were treated on days 0 and 7 with triamcinolone acetonide (6 mg/kg) to inhibit the host immune response and thereby increase bacterial yields. Rabbits were killed on day 11, and the testes were removed aseptically,

rinsed in phosphate-buffered saline (PBS), and minced. The tissue was extracted for 30 min with constant stirring under an atmosphere of 1.5% O_2, 5% CO_2, balance N_2 in 30 ml of PBS containing 10% heat-inactivated normal rabbit serum, 1.0 mM dithiothreitol, and heparin (1.5 U/ml). The extract was removed and saved and the extraction process repeated. The resulting suspensions were centrifuged twice at 500*g* for 7 min at 4°C to remove tissue debris, and the resulting supernatant was centrifuged at 20,000*g* for 30 min at 4°C to sediment the *T. pallidum*. Pellets from each extraction were washed once with PBS, resuspended in 2 ml of PBS, and layered onto a 34-ml 10 to 60% continuous gradient of Hypaque in PBS [J. B. Baseman, J. C. Nichols, O. Rumpp, N. S. Hayes, *Infect. Immun.* **10**, 1062 (1974)]. Gradients were centrifuged at 87,000*g* for 80 min at 20°C. Bands corresponding to purified *T. pallidum* were removed and washed once with 30 ml of PBS (20,000*g* for 30 min at 4°C) and twice in a microcentrifuge (1.5 ml, 13,000*g* for 10 min at 4°C). About 1.9×10^{10} purified organisms were recovered. *Treponema pallidum* in 1 ml of PBS was lysed by addition of 1 ml of lysis solution [tris-EDTA (TE) 10:100 0.5% SDS, proteinase K (100 mg/ml)] and incubation at 55°C for 2 hours. The lysate was repeatedly extracted with an equal volume of phenol: chloroform:isoamyl alcohol until the interface was free of visible debris. After a final extraction with chloroform and addition of 1 M ammonium acetate, the nucleic acid was precipitated from the aqueous phase by addition of two volumes of ethanol, pelleted by centrifugation, washed once with 70% ethanol, air-dried for 15 min at 37°C, resuspended in 100 μl of TE buffer (10:1), and stored at 4°C. DNA concentration was determined by fluorometry. Electrophoresis of the DNA in 0.6% agarose gel (0.5× tris–boric acid–EDTA buffer) indicated that the DNA was >23 kb in size. Purity of the *T. pallidum* DNA was assessed by treatment of 3 μl of the sample with Sal I, electrophoresis, Southern (DNA) blotting, and hybridization with a rabbit mitochondrial DNA probe. Although hybridizing bands were detectable, the amount of contaminating rabbit mitochondrial DNA was estimated to be <1% of the total DNA. For preparation of a small insert library in pUC 18 vector, ~10 μg of purified *T. pallidum* DNA was sheared by nebulization at 1.5×10^{-3} N/m² for 2 min and size fractionated on a 1.2% agarose gel. A library with an average insert size between 1.6 and 2.0 kb was prepared, and the *T. pallidum* genome was sequenced by a whole-genome random sequencing method previously applied to other microbial genomes (*8*, *9*). An about sevenfold genome coverage was achieved by generating 14,464 sequences with an average edited length of 525 bases. Sequences were assembled with TIGR Assembler, resulting in a total of 130 assemblies containing at least two sequences, which were clustered into 30 groups on the basis of linking information from forward and reverse sequence reads. All *T. pallidum* sequences that had been mapped previously were searched against the assemblies in an attempt to order the contig groups. Sequence and physical gaps for the chromosome were closed as previously described (*8*, *9*). A total of 676 sequencing reactions were necessary to complete genome closure. At the completion of the project, less than 1% of the chromosome had single-fold coverage. Coding regions (ORFs) were identified with compositional analysis that used an interpolated Markov model based on variable length oligomers [S. Salzberg, A. Delcher, S. Kasif, O. White, *Nucleic Acids Res.* **26**, 544 (1998)]. ORFs of >600 bp were used to train the Markov model, as well as *T. pallidum* ORFs from GenBank. Once trained, the model was applied to the complete *T. pallidum* genome sequence. ORFs that overlapped were visually inspected and, in some cases, removed. All putative ORFs were searched against a nonredundant amino acid database as previously described (*8*, *9*). ORFs were also analyzed with 527 hidden Markov models constructed for a number of conserved protein families (pfam v2.0) with hmmer [E. L. Sonhammer, S. R. Eddy, R. Durbin, *Proteins* **28**, 405 (1997)]. Families of paralogous genes were constructed by pair-wise searches of proteins with FASTA. Matches that spanned at least 60% of the smaller of the protein pair were retained and visually inspected. A total of 42 paralogous gene families containing 129 genes were identified (Fig. 1). TopPred was used to identify potential membrane-spanning domains (MSD) in proteins [M. G. Claros and G. vonHeijne, *Comput. Appl. Biosci.* **10**, 685 (1994)]. Five hundred thirty-five proteins containing at least one putative MSD were identified, of which 246 were predicted to have more than one MSD. The presence of signal peptides and the probable position of a cleavage site in secreted proteins were detected with Signal-P [H. Neilsen, J. Engelbrecht, S. Brunak, G. von Heijne, *Protein Eng.* **10**, 1 (1997)]; 155 proteins were predicted to have a signal peptide. Lipoproteins were identified by scanning for a lipobox in the first 30 amino acids of every protein. Twenty-two putative lipoproteins were identified, and five of these were ORFs of unknown function. All features identified with these programs were visually inspected to eliminate false positives.

11. M. Riley, *Microbiol. Rev.* **57**, 862 (1993).
12. J. R. Lobry, *Mol. Biol. Evol.* **13**, 660 (1996); M. P. Francino and H. Ochman, *Trends Genet.* **13**, 240 (1997).
13. P. M. Sharp and W.-H. Li, *Nucleic Acids Res.* **15**, 1281 (1987); J. O. McInerney, *Microbiol. Comp. Genomics* **2**, 89 (1997).
14. B. J. Paster, F. E. Dewhirst, W. G. Weisburg, *J. Bacteriol.* **173**, 6101 (1991); A. Centurion-Lara, C. Castro, W. C. van Voorhis, S. Lukehart, *FEMS Microbiol. Lett.* **143**, 235 (1996).
15. C. Ojaimi, B. E. Davidson, I. Saint Girons, I. G. Old, *Microbiology* **140**, 2931 (1994).
16. W. Freist, D. H. Gauss, M. Ibba, D. Soll, *Biol. Chem.* **378**, 1103 (1997).
17. M. Ibba, J. L. Bono, P. A. Rosa, D. Soll, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 14383 (1997); M. Ibba *et al.*, *Science* **278**, 1119 (1997).
18. S. Commans, P. Plateau, S. Blanquet, F. Dardel, *J. Mol. Biol.* **253**, 100 (1995); S. Cusack, A. Yaremchuk, M. Tukalo, *EMBO J.* **15**, 6321 (1996).
19. L. V. Stamm, S. R. Greene, N. Y. Barnes, H. L. Bergen, J. M. Hardham, *FEMS Microbiol. Lett.* **155**, 115 (1997).
20. C. D. Cox, in *Pathogenesis and Immunology of Treponemal Infection*, R. F. Schell and D. M. Musher, Eds. (Dekker, New York, 1983), pp. 57–70.
21. J. T. Barbieri, F. E. Austin, C. D. Cox, *Infect. Immun.* **31**, 1071 (1981).
22. L. V. Stamm, N. R. Young, J. G. Frye, J. M. Hardham, *DNA Sequence* **6**, 293 (1996).
23. S. F. Porcella *et al.*, *Gene* **177**, 115 (1996).
24. A. Death and T. Ferenci, *Res. Microbiol.* **144**, 529 (1993); T. Ferenci, *FEMS Microbiol. Rev.* **18**, 301 (1996).
25. A. Matsuura, A. Iwashima, Y. Nose, *J. Vitaminol.* **18**, 29 (1972).
26. K. G. Van Horn and R. M. Smibert, *Can. J. Microbiol.* **29**, 1141 (1983); R. C. Johnson, in *The Prokaryotes*, M. P. Starr *et al.*, Eds. (Springer-Verlag, Berlin, 1981), pp. 582–591.
27. S. J. Norris *et al.*, unpublished data.
28. W. E. O'Brien, S. Bowien, H. G. Wood, *J. Biol. Chem.* **250**, 8690 (1975); R. E. Reeves, R. Serrano, D. J. South, *ibid.* **251**, 2958 (1976); R. G. Kemp and R. L. Tripathi, *J. Bacteriol.* **175**, 5723 (1993); I. Bruchhaus, T. Jacobs, M. Denart, E. Tannich, *Biochem. J.* **316**, 57 (1996).
29. P. G. Lysko and C. D. Cox, *Infect. Immun.* **21**, 462 (1978); *ibid.* **16**, 885 (1977); N. L. Schiller and C. D. Cox, *ibid.*, p. 60.
30. M. Forgac, *Physiol. Rev.* **69**, 765 (1989); P. L. Pederson and E. Carafoli, *Trends Biochem. Sci.* **12**, 146 (1987); J. Konishi, T. Wakagi, T. Oshima, M. Yoshida, *J. Biochem.* **102**, 1379 (1987).
31. E. Schwarz, D. Oesterhelt, H. Reinke, K. Beyreuther, P. Dimroth, *J. Biol. Chem.* **263**, 9640 (1988); G. Woehlke, K. Wifling, P. Dimroth, *ibid.* **267**, 22798 (1992).
32. C. Shibata, T. Ehara, K. Tomura, K. Igarashi, H. Kobayashi, *J. Bacteriol.* **174**, 6117 (1992); K. Takase *et al.*, *J. Biol. Chem.* **269**, 11037 (1994).
33. C. D. Cox and M. K. Barber, *Infect. Immun.* **10**, 123 (1974); W. H. Cover *et al.*, *Sex. Transm. Dis.* **9**, 1 (1982); D. L. Cox *et al.*, *Appl. Environ. Microbiol.* **56**, 3063 (1990).
34. L. V. Stamm, F. C. Gherardini, E. A. Parrish, C. R. Moomaw, *Infect. Immun.* **59**, 1572 (1991).
35. S. J. Norris, unpublished data.
36. J. Reizer *et al.*, *Mol. Microbiol.* **27**, 1157 (1998); J. Deutscher and M. H. Saier Jr., *Proc. Natl. Acad. Sci. U.S.A.* **80**, 6790 (1983); B. E. Jones *et al.*, *J. Biol. Chem.* **272**, 26530 (1997); J. Reizer *et al.*, *EMBO J.* **8**, 2111 (1989).
37. B. S. Powell *et al.*, *J. Biol. Chem.* **270**, 4822 (1995); J. Reizer, A. Reizer, M. H. Saier Jr., G. R. Jacobsen, *Protein Sci.* **1**, 722 (1992); M. H. Saier Jr. and J. Reizer, *Mol. Microbiol.* **13**, 755 (1994).
38. K. Doetsch, *Nature* **255**, 656 (1975); N. W. Charon, E. P. Greenberg, M. B. Koopman, R. J. Limberger, *Res. Microbiol.* **143**, 597 (1992); S. R. Greene, L. V. Stamm, J. M. Hardham, N. R. Young, J. G. Frye, *DNA Sequence* **7**, 267 (1997); K. E. Hagman, S. F. Porcella, T. G. Popova, M. V. Norgard, *Infect. Immun.* **65**, 1701 (1997).
39. S. J. Norris, N. W. Charon, R. G. Cook, M. D. Fuentes, R. J. Limberger, *J. Bacteriol.* **170**, 4072 (1988); R. D. Isaacs *et al.*, *Infect. Immun.* **57**, 3403 (1989); R. D. Isaacs and J. D. Radolf, *ibid.* **58**, 2025 (1990); L. Pallesen and P. Hindersson, *ibid.* **57**, 2166 (1989); C. I. Champion, J. N. Miller, M. A. Lovett, D. R. Blanco, *ibid.* **58**, 1697 (1990).
40. R. J. Limberger, L. L. Slivienski, M. C. El-Afandi, L. A. Dantuono, *J. Bacteriol.* **178**, 4628 (1996); Y. Ge and N. W. Charon, *FEMS Microbiol. Lett.* **153**, 425 (1997).
41. J. D. Radolf, *Mol. Microbiol.* **16**, 1067 (1995); D. R. Blanco *et al.*, *J. Bacteriol.* **177**, 3556 (1995); D. R. Blanco *et al.*, *ibid.* **178**, 6685 (1996); D. R. Blanco, J. N. Miller, M. A. Lovett, *Emerg. Infect. Dis.* **3**, 11 (1997); D. R. Akins *et al.*, *J. Bacteriol.* **179**, 5076 (1997); C. I. Champion *et al.*, *ibid.*, p. 1230; J. M. Hardham *et al.*, *Gene* **197**, 47 (1997).
42. J. C. Fenno *et al.*, *J. Bacteriol.* **179**, 1082 (1997); D. A. Mathers, W. K. Leung, J. C. Fenno, Y. Hong, B. C. McBride, *Infect. Immun.* **64**, 2904 (1996); J. C. Fenno, K. H. Muller, B. C. McBride, *J. Bacteriol.* **178**, 2489 (1996).
43. T. J. Fitzgerald, L. A. Repesh, S. G. Oakes, *Br. J. Vener. Dis.* **58**, 1 (1982); S. G. Oakes, L. A. Repesh, R. S. Pozos, T. J. Fitzgerald, *ibid.*, p. 220; G. H. Wong, B. M. Steiner, S. Graves, *Infect. Immun.* **41**, 636 (1983).
44. T. Guina and D. B. Oliver, *Mol. Microbiol.* **24**, 1201 (1997).

45. We thank A. R. Kerlavage for preparation of Fig. 2; D. Soll and M. H. Saier Jr. for helpful discussions; M. Heaney, J. Scott, and A. Saeed for software and database support; and B. Vincent, D. Maas, and V. Sapiro for computer system support. Supported by NIH grant AI31068 to GMW. Requests for reprints should be sent to C.M.F. (e-mail: tpdb@tigr.org). The annotated genome sequence and gene family alignments are available on the World Wide Web at www.tigr.org/tdb/mdb/tpdb/tpdb.html. Additional information regarding the *T. pallidum* genome is available at utmmg.med.uth.tmc.edu/treponema/tpall.html. The genome sequence has been deposited with GenBank under accession number AE000520.

6 March 1998; accepted 6 May 1998