

# BAC Library of *T. pallidum* DNA in *E. coli*

David Šmajš,<sup>1,2</sup> Matthew McKeivitt,<sup>3</sup> Ling Wang,<sup>1,2</sup> Jerrilyn K. Howell,<sup>4</sup>  
Steven J. Norris,<sup>4</sup> Timothy Palzkill,<sup>3</sup> and George M. Weinstock<sup>1,2,5,6</sup>

<sup>1</sup>Department of Microbiology and Molecular Genetics, University of Texas–Houston Medical School, Houston, Texas 77030, USA; <sup>2</sup>Center for the Study of Emerging and Re-emerging Pathogens, University of Texas–Houston Medical School, Houston, Texas 77030, USA; <sup>3</sup>Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, Texas 77030, USA; <sup>4</sup>Department of Pathology and Laboratory Medicine, University of Texas–Houston Medical School, Houston, Texas 77030, USA; <sup>5</sup>Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA

*Treponema pallidum* subspecies *pallidum* (Nichols) chromosomal DNA was used to construct a large insert bacterial artificial chromosome (BAC) library in *Escherichia coli* DH10B using the pBeloBACII cloning vector; 678 individual insert termini of 339 BAC clones (13.9 x coverage) were sequenced and the cloned chromosomal region in each clone was determined by comparison to the genomic sequence. A single 15.6-kb region of the *T. pallidum* chromosome was missing in the BAC library, between bp 248727 and 264323. In addition to the 12 open reading frames (ORFs) coded by this region, one additional ORF (TP0596) was not cloned as an intact gene. Altogether, 13 predicted *T. pallidum* ORFs (1.25% of the total) were incomplete or missing in the library. Three of 338 clones mapped by restriction enzyme digestion had detectable deletions and one clone had a detectable insertion within the insert. Of mapped clones, 19 were selected to represent the minimal set of *E. coli* BAC clones covering 1026 of the total 1040 (98.7%) predicted *T. pallidum* ORFs. Using this minimal set of clones, at least 12 *T. pallidum* proteins were shown to react with pooled sera from rabbits immunized with *T. pallidum*, indicating that at least some *T. pallidum* genes are transcribed and expressed in *E. coli*.

*Treponema pallidum* subspecies *pallidum* (Nichols), causative agent of the sexually transmitted disease syphilis, cannot be continuously grown under in vitro conditions. It also cannot cause syphilis in animals. As a result, there is limited genetic data about the *T. pallidum* spirochete and its interactions with its human host. However, the genome of *T. pallidum* (1.14 Mbp) was completely sequenced and 1040 open reading frames (ORFs) were predicted (Fraser et al. 1998; Weinstock et al. 1998), opening the door for new approaches. Because *T. pallidum* cannot be continuously cultured in the laboratory, and is usually purified from infected rabbit testes, there are still challenges in taking advantage of this genomic information.

Construction of genomic libraries represents an important approach in the study of pathogenic bacteria that are difficult to culture. Screening of genomic libraries of *T. pallidum* was used for identification of genes coding for antigens (Bailey et al. 1989), exported proteins (Hardham et al. 1995), and genes able to complement *Escherichia coli* mutants (Gherardini et al. 1990). For these purposes, libraries with relatively small inserts were prepared, each clone coding for several genes. It is known for these libraries that problems of biased representation of genes and clone instability occur (Brayton et al. 1999; Hindle et al. 1994).

For stable large insert libraries, bacterial artificial chromosome (BAC) vectors have been used for eukaryotic, as well as for bacterial, species. In contrast to eukaryotic BAC libraries, prokaryotic inserts may express genes using endogenous

signals resembling those of *E. coli*. Such production of cognate foreign proteins may interfere with *E. coli* growth, for example, by assembling into complexes of reduced function. Gene expression from bacterial inserts in BACs was detected (Rondon et al. 1999; Xu et al. 1998) and a reduced maximum insert size was observed compared with nonprokaryotic inserts. The F-plasmid derived copy number control of the BAC vector allows one to two copies of BAC DNA per cell, which is crucial in cloning genes that are toxic when overexpressed. Moreover, better growth of the *E. coli* host and a reduced rate of DNA rearrangements are more likely with BAC clones.

Because of the difficulty in obtaining *T. pallidum* DNA, we undertook to construct a set of large insert clones covering the whole genome. Here we report the construction and characterization of a large insert genomic library of *T. pallidum* in a BAC vector in *E. coli*. The detailed analysis of the resulting clones allows us to test the hypothesis that the DNA cloning efficiency depends on specific gene content. In addition, large blocks of the *T. pallidum* chromosome propagated in *E. coli* will allow the use of genetic approaches to study *T. pallidum* genes, including methods of functional genomics, strain comparisons, and postgenomic applications.

## RESULTS AND DISCUSSION

### Construction of a *T. pallidum* BAC Vector Library in *E. coli* DH10B

To provide a convenient source of *T. pallidum* DNA in as few clones as possible, a large insert library was constructed. *T. pallidum* chromosomal DNA was partially digested with *Hind* III restriction enzyme, and digested DNA was size-selected using PFGE. The 259 *Hind* III target sites were randomly distributed throughout the *T. pallidum* chromosome with the largest fragments being 31 kb comprising ORFs TP0273–TP0304 and

**Present address: Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Alkek N1519, Houston, TX 77030, USA.**

**Corresponding author.**

**E-MAIL gwstock@bcm.tmc.edu; FAX (713) 798-5741.**

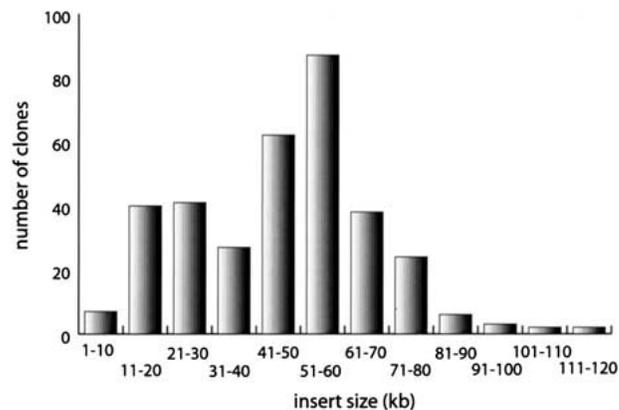
Article and publication are at <http://www.genome.org/cgi/doi/10.1101/gr.207302>.

25 kb comprising ORFs TP00448–TP0471. Four different agarose blocks were cut out of the gel containing DNA fragments between 40 kb and 80 kb, 80 kb and 120 kb, 120 kb and 160 kb, and 160 kb and 200 kb, respectively. To maximize the number of resulting clones, the digested DNA was not subjected to tight size selection (e.g., pre-electrophoresis). Moreover, using these conditions could result in clones with smaller inserts from regions of the *T. pallidum* chromosome that cannot be efficiently cloned on large inserts. DNA was electroeluted from agarose and ligated into the pBeloBAC11 cloning vector. The dialyzed ligation mixture was used for electroporation of *E. coli* DH10B cells, and white colonies were further characterized. The number of transformants (white colonies) was dependent on the size of DNA used for cloning. The 40–80-kb fragments resulted in most of the white colonies isolated ( $\sim 10^3$ ), with only  $\sim 10\%$  of the colonies isolated from 80 kb to 120 kb fragments and none for 120 kb to 160 kb and 160 kb to 200 kb inserts. Of the white colonies,  $>20\%$  represented empty clones and were discarded. Parallel construction of a similar large insert library for the culturable bacterium *Treponema denticola* gave significantly better results for the 80–120-kb fraction but still showed a strong bias against large inserts. Experiments with human and mycobacterial DNA showed that the maximum size of inserts is dependent on the source of the DNA. For prokaryotic DNA, the insert length is considerably restricted when compared with eukaryotic DNA (Brosch et al. 1998). Using the pBeloBAC11 cloning vector, the maximum insert size achieved for *Mycobacterium tuberculosis* genomic DNA was 104 kb. On the other hand, the maximum insert sizes reported were  $>180$  kb for *Bacillus cereus* DNA (Rondon et al. 1999), 250 kb for the opportunistic human pathogen *Ochrobactrum anthropi* (Tomkins et al. 1999), and 290 kb for *Pseudomonas aeruginosa* (Dewar et al. 1998).

### Sequencing BAC Clones Containing *T. pallidum* Chromosomal DNA

To accurately characterize the inserts, the ends of 339 inserts in BAC clones were sequenced using priming sites on both sides of the multiple cloning site of pBeloBAC11. The sequences were compared with the *T. pallidum* genomic sequence (Fraser et al. 1998) to characterize the inserts. Insert sizes varied from 6.4 kb to 120.4 kb with an average of 46.7 kb. No noncontiguous insert ends were detected. The distribution of insert lengths is shown in Figure 1. A major peak was seen for insert sizes of 51–60 kb with 87 clones in this category. This finding is consistent with the predominant 40–80-kb length of DNA fragments giving positive transformants. Another peak for 11–30-kb inserts was also observed. This may be explained by preferential cloning and higher transformation efficiency of smaller clones. DNA eluted from the agarose gel slice is likely to contain smaller contaminating DNA fragments despite size selection.

The 339 clones contained 15,781 kb of *T. pallidum* DNA, representing 13.9 x clone coverage of the genome. A single 15.6-kb gap was found between bp 248727 and 264323 of the *T. pallidum* chromosome. There are 12 ORFs predicted in this interval (TP0241–TP0252). In addition, the *Hind* III site at position 647900 was not present in any clone internal to the insert DNA. The corresponding ORF (TP0596) containing this site was thus not completely present in any of the clones. Altogether, 13 ORFs (TP0241–TP0252; TP0596) were not cloned intact or were completely missing. The properties of



**Figure 1** Distribution of insert lengths of the 339 end-sequenced clones of the *T. pallidum* library. The average insert size was 46.7 kb and the insert length ranged from 6.4 to 120.4 kb as calculated by comparison to the genomic sequence. The major peak corresponds to insert lengths of 51–60 kb and a second, minor peak was observed for insert lengths 11–30 kb.

the missing genes with known or predicted function are shown in Table 1. With the exception of TP0242, all other missing ORFs were cloned in pUniD/V5-His-TOPO (Invitrogen; data not shown). However, no promoter sequence was present upstream of the cloned ORFs in this vector. In addition, TP0241, TP0243, TP0244, TP0246–TP0248, TP0251, and TP0252 were cloned as fusions to glutathione S-transferase gene in the pHB2-GST expression vector. The *E. coli* TP0596 ortholog, *pcnB*, codes for a poly(A) polymerase directing mRNA polyadenylation. It is known that the function of this gene is dose-dependent and even moderate overexpression of *pcnB* is lethal to *E. coli* (Cao and Sarkar 1992). Thus, it is likely that TP0596 is expressed in *E. coli* and negatively selected. With respect to predicted functions of TP0241–TP0252, it is possible that *T. pallidum* ORFs coding for components of protein complexes like DNA-dependent RNA polymerase or the ribosome were difficult to clone in *E. coli* because of their interference with *E. coli* components. However, other explanations for the absence of those genes are possible, for example, the presence of unclonable DNA sequences or the proximity to rRNA gene clusters (Fig. 2). In the *M. tuberculosis* library, a single  $\sim 150$ -kb gap within 420 BAC clones was identified (Brosch et al. 1998). In this case, the missing DNA fragment resulted from the missing *Hind* III target sites in this region when *Hind* III was used for construction of the library. However, this explanation cannot apply to the missing *T. pallidum* ORFs.

### Number of Individual Gene (ORF) Copies per Library and Corresponding Average Insert Length of Encoding BAC Clones

It became apparent that to achieve the complete BAC coverage of the chromosome, the total genome coverage should exceed 10x. Statistics were calculated to study the hypothesis that the recovery of prokaryotic DNA in BAC vectors is region-specific because groups of genes that are transcribed and expressed collectively impair growth of *E. coli*. The number of copies of each complete ORF in the library was counted. The copy numbers varied widely throughout the chromosome, from 0 to 27 with an average of 13.7 copies per gene (Fig. 2A). Some *T. pallidum* chromosomal regions appeared to be cloned

**Table 1.** Under- and overrepresented genes in the *T. pallidum* BAC library

| Copy number of intact genes | Predicted function (gene) <sup>a</sup>   |
|-----------------------------|--|
| 0                           | DNA-directed RNA polymerase, beta subunit ( <i>rpoB</i> , TP0241); RNA polymerase, beta' subunit ( <i>rpoC</i> , TP0242); ribosomal protein S12 ( <i>rpsL</i> , TP0243); ribosomal protein S7 ( <i>rpsG</i> , TP0244); N-acetylmuramoyl-L-alanine amidase ( <i>amiA</i> , TP0247); flagellar filament outer layer protein ( <i>flaA-1</i> , TP0249); DNA-binding protein II ( <i>dbh</i> , TP0251); apolipoprotein N-acyltransferase ( <i>cutE</i> , TP0252); polynucleotide adenyltransferase ( <i>pcnB</i> , TP0596) |
| 1                           | ribosomal protein L10 ( <i>rplJ</i> , TP0239); ribosomal protein L7/L12 ( <i>rplL</i> , TP0240)  |
| 2                           | primosomal protein N ( <i>priA</i> , TP0230); ribosomal protein L1 ( <i>rplA</i> , TP0238); oligopeptide ABC transporter ( <i>oppA</i> , TP0585); leucyl-tRNA synthetase ( <i>leuS</i> , TP0586); phosphocarrier protein HPr ( <i>ptsH</i> , TP0589); ribosomal protein L36 ( <i>rpmJ-2</i> , TP0590); HPr kinase ( <i>ptsK</i> , TP0591); adenylate kinase ( <i>adk</i> , TP0595)   |
| 3                           | heat shock protein 70 ( <i>dnaK</i> , TP0216); anti-sigma F factor antagonist (TP0233); ribosomal protein L33 ( <i>rpmG</i> , TP0234); preprotein translocase subunit ( <i>secE</i> , TP0235); transcription antitermination protein ( <i>nusG</i> , TP0236); ribosomal protein L11 ( <i>rplK</i> , TP0237); ribosomal protein L1 ( <i>rplA</i> , TP0238)  |
| 23                          | long-chain-fatty-acid-CoA ligase (TP0145); thioredoxin ( <i>trx</i> , TP0919); flagellar protein ( <i>fliS</i> , TP0943); DNA helicase II ( <i>uvrD</i> , TP1028); tpr protein L ( <i>tprL</i> , TP1031)   |
| 24                          | K <sup>+</sup> transport protein ( <i>ntpI</i> , TP0168); methylated-DNA-protein-cysteine S-methyltransferase ( <i>dat</i> , TP0141); thiamine ABC transporter, ATP-binding protein (TP0142); thiamine ABC transporter, permease protein (TP0143); thiamine ABC transporter, thiamine-binding periplasmic protein (TP0144); prolyl-tRNA synthetase ( <i>proS</i> , TP0160); Mg <sup>2+</sup> transport protein ( <i>mgfE</i> , TP0917); pyruvate oxidoreductase (TP0939)   |
| 25                          | alpha-amylase 1 (TP0147); nitrogen fixation protein ( <i>mfc</i> , TP0152); oligoendopeptidase F (TP1026)  |
| 27                          | tex protein ( <i>tex</i> , TP0924); flavodoxin (TP0925); signal peptidase I (TP0926); UDP-N-acetylmuramoylalanyl-D-glutamate-2,6-diaminopimelate ligase ( <i>murE</i> , TP0933); glutamate transporter (TP0934); N-acetylphosphinothricin-tripetide-deacetylase (TP0935); hemolysin (TP0936)   |

<sup>a</sup>Further information available at <http://www.tigr.org> and <http://www.stdgen.lanl.gov>

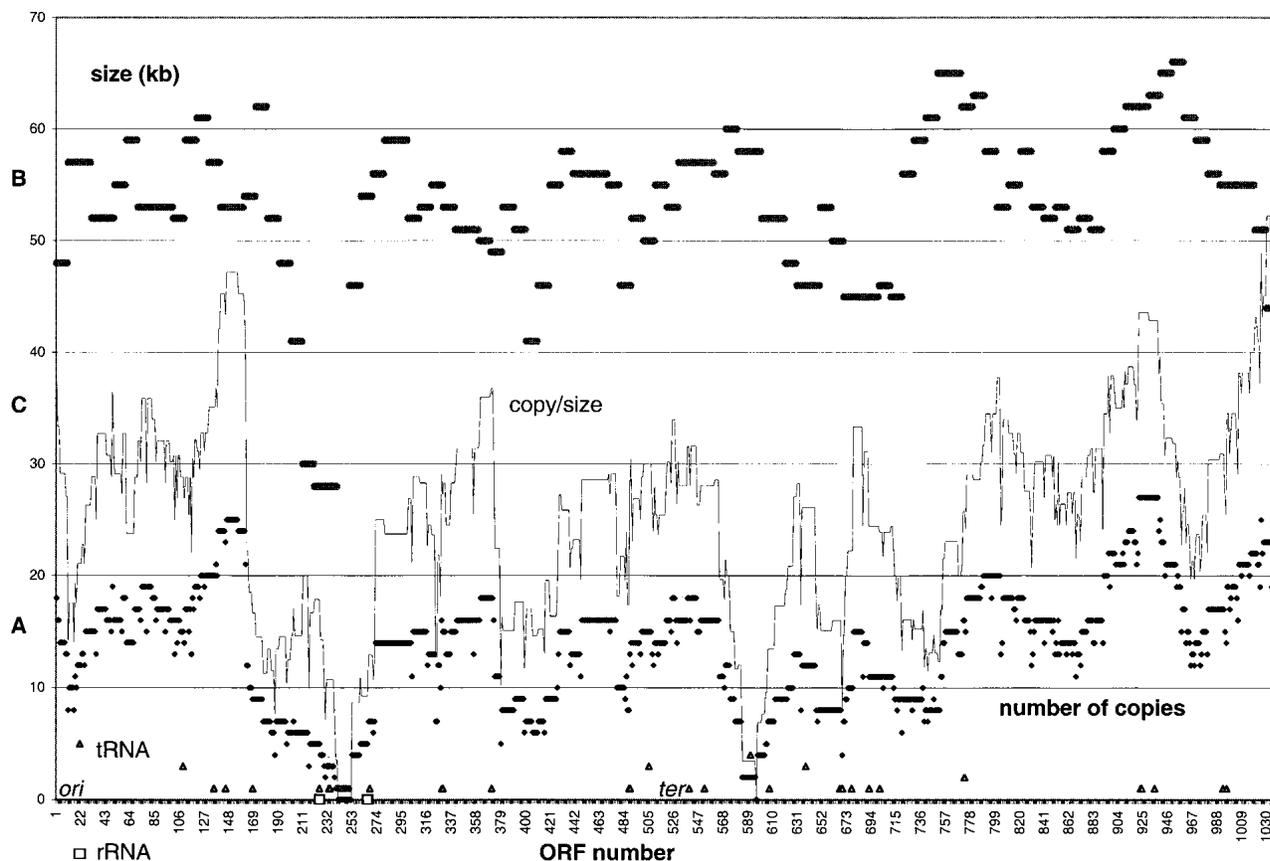
and isolated more often than others. Although for most ORFs the number of copies ranged between 8–20 (759 genes, 72.9%), three regions comprising 113 ORFs (TP0137–TP162, TP0896–TP0953, and TP1007–TP1034) were found at >20 copies. Three regions, containing 146 ORFs (TP0177–TP0272, TP0399–TP415, and TP0578–TP611) were found to be present in less than eight copies. In addition, 10 other ORFs (TP0324, TP325, TP0379, TP0648, TP0669–TP0671, TP0720, TP0739, and TP0742) were found in less than eight copies. All 10 of these ORFs contained *Hind* III restriction site(s) within their sequence, often used for cloning, and thus may be underrepresented when complete ORFs were counted. Overrepresented regions contained 59 (52%) hypothetical ORFs (i.e., genes with unknown function and no database match). Underrepresented regions were enriched for genes with database matches with 52 (37%) hypothetical ORFs. Although the significance of this bias is not clear, genes with database matches are likely to have cognates in *E. coli* and may interfere with cell function if expressed. Moreover, hypothetical ORFs might not be as likely to be transcribed and thus more likely to be functionally inactive. The most under- and overrepresented genes are shown in Table 1. Underrepresented genes are mainly involved in general cellular processes including transcription, translation, and protein transport and modification, whereas overrepresented genes code for enzyme and transport functions.

For each window comprising ten adjacent ORFs, an average length of BAC clones containing at least one ORF within this window was calculated and the results are shown in Figure 2B. The median value of 104 windows was 53 kb. The average insert length varied from 28 kb to 66 kb along the *T. pallidum* chromosome. The patterns of the number of ORF (gene) copies and average length of corresponding inserts were generally similar, indicating that the regions present in most copies in the library could be cloned on larger inserts and vice versa. This is consistent with regional selection against blocks of expressed genes. However, with the excep-

tion of TP0896–TP0953 and TP0399–TP415, the distributions of the number of gene copies and average insert lengths often did not correspond exactly and the copy number/size ratio varied from 0 to 0.52 (Fig. 2C). These data indicate that additional factors are involved in the ability to recover cloned regions. This situation seems to apply also for other prokaryotic BAC libraries. Regions of the *M. tuberculosis* chromosome were significantly underrepresented and additional selection was needed to isolate BAC clones containing these regions from the library pool (Brosch et al. 1998). Regions in the *B. cereus* BAC library are also known to be underrepresented. When this library was screened for six particular genes, the number of positive clones ranged between one and two per gene, although the coverage of the library was estimated to be 5.75-fold (Rondon et al. 1999). To more closely address this question, the average number of gene copies per library was calculated for groups of genes coding for proteins predicted to have similar functions based on the published annotation (Fraser et al. 1998). Although the differences in copy numbers among the group of genes were not statistically significant, genes coding for protein secretion, transcription, and RNA processing, and genes for ribosomal proteins were present in fewer copies than genes involved in energy metabolism and transport functions (not shown). Because *T. pallidum* regions containing dozens of genes, and not the single genes, were selected during the library construction, these data are consistent with the hypothesis that regions containing genes for components of multiprotein complexes regulating gene expression and protein trafficking will interfere with the cell viability.

### Verification of BAC Clone Sizes and Insert Continuity

Although the sequences at the ends of the inserts were determined, the possibility remains that the clones may have alterations internal to the inserts. To test if the clones contain



**Figure 2** A (lower plot): Plot of the number of ORF (gene) copies present in the library calculated for each individual ORF. The copy numbers varied from 0 to 27 with an average of 13.7 copies per library. Altogether, 13 genes were missing in the library (TP0241–TP0252, TP0596). B (upper plot): Plot of the average insert size calculated for 104 10-ORF windows of the *T. pallidum* chromosome. BAC clones harboring at least one ORF from each 10-ORF window were used for calculation of average insert size within this window. C (middle plot): Plot of the ratio between gene copy number and window average insert size ranging between 0 and 0.52. Open triangles represent number (1–5) and position of tRNA genes, whereas open squares represent the positions of two 3-gene rRNA clusters. The possible origin (*ori*) and terminus (*ter*) of replication regions are indicated.

the desired DNA, 338 clones were digested with *Hind* III and, in addition, some of them with *Eco*R I, *Xba* I, and *Xho* I, and the lengths of the fragments were compared with the predicted lengths derived from the genome sequence. For most clones, the experimentally measured fragments matched the predicted lengths and no deletions, insertions, or aberrant bands were found. However, 2 of 338 clones tested, DSTP133 and DSTP313, showed large deletions of >40 kb in the insert. On the other hand, DSTP133 DNA was used as a template for individual amplification by PCR of 74 of 88 ORFs predicted to be encoded by this clone. This contradiction is consistent with the hypothesis that the deletion occurred subsequent to cloning of the *T. pallidum* DNA fragment into pBeloBAC11 vector. A small subpopulation of clones probably contained undelivered DNA sufficient for the PCR amplifications. The deleted region in clone DSTP133 comprised the region coding for ribosomal proteins. In clone DSTP313, the region around ORF TP0596 (Table 1) was deleted. There are no other BACs spanning the deleted regions in one piece. Moreover, in DSTP313 no other clone with a complete ORF TP0596 was found. Thus, the *pcnB* gene (TP0596) is likely to cause the BAC insert instability. The deletion in the DSTP133 clone may be explained by the presence of two or more genes contributing to the insert instability. The decreased number of ORF (gene)

copies present in the library within this region (around the ORF TP0200; Fig. 2B) supports this idea. These results indicate that some BAC clones are unstable when cloned in *E. coli* and that this instability is likely attributable to the specific gene content in the insert DNA. A small deletion (1 kb) was found in the clone DSTP276 in the region comprising ORFs TP0044–TP0048, and a 1.9-kb insertion was found in the clone DSTP130 within the gene coding for ribosomal protein L5 (TP0201). No deletions and insertions were found in an additional 14 and 5 clones covering similar *T. pallidum* chromosome regions, respectively. In addition, 6 of 21 clones harboring TP0126 were found to contain the same 1.3-kb insertion within this ORF, indicating that at least two subpopulations of *T. pallidum* Nichols strain were present in infected rabbits. Intrastrain genetic heterogeneity has been recently shown in *T. pallidum* strains (Stamm and Bergen 2000). However, it is not known if this explanation may apply also for DSTP130 and/or DSTP276. Taken together, 2 to 4 clones of 338 (0.59%–1.18%) have been shown to contain aberrations within the inserts.

### Minimal Set of Clones Covering the *T. pallidum* Chromosome

A set of 19 clones was the smallest number of clones covering

**Table 2.** Minimal Set of Clones Covering the *T. pallidum* Chromosome

| Name    | Start coordinate    | Stop coordinate | No. of complete ORFs | Length of insert (bp) | Complete ORFs encoded | No. of ORFs amplified <sup>a</sup> |
|---------|---------------------|-----------------|----------------------|-----------------------|-----------------------|------------------------------------|
| DSTP001 | 158990 <sup>b</sup> | 93406           | 55                   | 65585                 | TP0084-TP0137         | 40                                 |
| DSTP003 | 543253              | 454629          | 81                   | 88625                 | TP0427-TP0507         | 74                                 |
| DSTP021 | 248727              | 211853          | 37                   | 36875                 | TP0203-TP0240         | 27                                 |
| DSTP029 | 264323              | 340728          | 71                   | 76406                 | TP0253-TP0323         | 60                                 |
| DSTP046 | 349734              | 402678          | 50                   | 52945                 | TP0329-TP0378         | 78                                 |
| DSTP055 | 1047306             | 1120241         | 61                   | 72936                 | TP0966-TP1026         | 58                                 |
| DSTP076 | 974953              | 1047790         | 68                   | 72838                 | TP0898-TP0965         | n.t.                               |
| DSTP085 | 1120242             | 95971           | 99                   | 113742                | TP1028-TP0085         | 95                                 |
| DSTP094 | 647900              | 579485          | 58                   | 68416                 | TP0538-TP0595         | 11                                 |
| DSTP109 | 379311              | 282570          | 89                   | 96742                 | TP0267-TP0355         | 29                                 |
| DSTP147 | 622396              | 533036          | 75                   | 89361                 | TP0499-TP0573         | 52                                 |
| DSTP155 | 926938              | 1047305         | 110                  | 120368                | TP0855-TP0964         | 98                                 |
| DSTP173 | 891478              | 787554          | 103                  | 103925                | TP0721-TP0823         | 97                                 |
| DSTP198 | 684778              | 754634          | 59                   | 69857                 | TP0628-TP0686         | 51                                 |
| DSTP201 | 943409              | 886572          | 47                   | 56838                 | TP0818-TP0864         | 42                                 |
| DSTP216 | 813434              | 742636          | 69                   | 70799                 | TP0679-TP0747         | 44                                 |
| DSTP274 | 134173              | 211852          | 87                   | 77680                 | TP0117-TP0203         | n.t.                               |
| DSTP288 | 383107              | 464889          | 76                   | 81783                 | TP0362-TP0437         | 62                                 |
| DSTP334 | 694312              | 647900          | 37                   | 46413                 | TP0597-TP0633         | n.t.                               |

<sup>a</sup>Number of ORFs amplified from each individual clone used as a template; n.t., not tested.

<sup>b</sup>Position on *T. pallidum* chromosome based on published coordinates (Fraser et al. 1998).

the most *T. pallidum* ORFs. The 19 clones covered 1462 kb and with the exception of one 15.6-kb region (1.4% of total DNA) covered the whole *T. pallidum* chromosome (1138 kb). These clones coded for 1026 of 1040 ORFs representing 98.7% of the total ORFs. The minimal clone set had 1.3 x clone coverage of the genome and an insert length ranging from 36.9 to 120.4 kb, with an average length of 76.9 kb. A detailed description and positions of inserts are given in Table 2 and Figure 3. In addition to ORFs TP0241–TP0252 and TP0596, TP1027 was not cloned intact and thus considered as missing. The number of ORFs amplified by PCR from individual clones of the minimal set is shown in Table 2. The remaining ORFs of each clone were either not tested or the amplification was negative.

### Screening of the *T. pallidum* Library with Anti-*T. pallidum* Rabbit Sera

The minimal set of clones was used for antigenic screening with immunized rabbit sera. At least 12 positive gene products were identified and screening with additional overlapping clones was used to narrow the DNA region coding for individual antigens (Fig. 4). The results are summarized in Table 3. For most of the antigens, the number of possible genes coding for the positively reacting protein was restricted to 1 to 12 genes. The corresponding genes were then predicted based on the detected molecular weight of antigen (Table 3). Detection of 12 *T. pallidum* proteins reacting with rabbit sera indicates that at least some of the *T. pallidum* genes are transcribed and expressed in *E. coli* and that this library can be used for functional studies. However, it is not known if the detected anti-

gens were identified because of specific immunity of rabbits to *T. pallidum* or represent a nonspecific cross-reactivity of rabbit sera with *T. pallidum* proteins. At least 7 of 12 proteins identified as antigens can be found within 15 *T. pallidum* polypeptides identified as antigens using a set of 41 antibody reagents (Norris 1993).

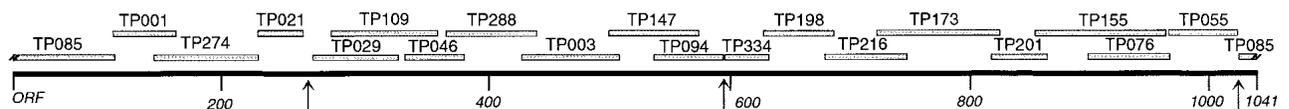
### Screening for Hemolytic Activity

The minimal set of clones were inoculated on the 5% sheep blood agar plates and screened for hemolytic activity. None of the 19 clones were identified to cause hemolysis, although the clones contained five genes that have been predicted to code for hemolysins. Similar screening of 81 *T. denticola* BAC clones resulted in identification of two hemolytic clones. Both clones contained 60–66-kb inserts and overlapped each other (data not shown).

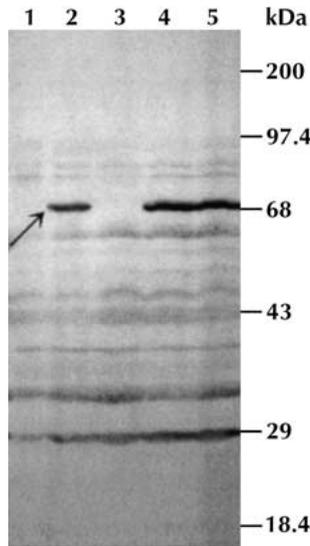
## METHODS

### Media

Bacterial strains were grown at 37°C in TY medium containing 8 g Bacto-tryptone (Difco Laboratories), 5 g yeast extract, and 5 g NaCl per liter (pH 7). For selection and maintenance of plasmids, 12.5–25 µg of chloramphenicol per mL of liquid medium or 1.5% TY agar (w/v) were added. Isopropylthio-β-D-galactoside (IPTG) and 5-bromo-4-chloro-3-indolyl-β-galactoside (X-Gal) were used at 0.5 mM and 40 µg mL<sup>-1</sup>, respectively. Ribonuclease A (Sigma) was used in resuspension buffer P1.



**Figure 3** The minimal set of 19 clones covering >98% of the *T. pallidum* chromosome. Positions of individual clones are shown and arrows indicate gaps in the coverage. Altogether, 14 *T. pallidum* ORFs are missing or incomplete in this minimal set: TP0241–TP0252, TP0596, and TP1027.



**Figure 4** Western blot analysis of *T. pallidum* clones. Strains were grown overnight and protein samples were run in 10% SDS-PAGE gels. Gene products were visualized using pooled sera from rabbits infected with *T. pallidum* and horseradish peroxidase labeled antirabbit secondary antibody after chemiluminescent detection. Lanes: 1, DH10B pBeloBAC11; 2, DH10B pDSTP021; 3, DH10B pDSTP012; 4, DH10B pDSTP108; and 5, DH10B pDSTP179. pDSTP021 codes for 37 ORFs (TP0203–TP0240), pDSTP012 for 13 (TP0217–TP0229), pDSTP108 codes for 50 (TP0188–TP0237), and pDSTP179 codes for 28 (TP0198–TP0225).

**Bacterial Strains and Plasmids**

The *E. coli* DH10B strain (Grant et al. 1990) was used for electroporation and the VCS257 strain (Stratagene) was used for isolation of vector DNA. For construction of a library, pBeloBAC11 vector was used (Kim et al. 1996). *T. pallidum* subspecies *pallidum* (Nichols) was used for isolation of chromosomal DNA.

**Isolation of *T. pallidum* Chromosomal DNA**

*T. pallidum* was grown in rabbit testes, harvested, and the cells purified using sodium diatrizoate gradient centrifugation

(Baseman and Hayes 1974; Hanff et al. 1984). Plugs containing *T. pallidum* DNA were prepared according to Walker et al. (1995). An equal volume of *T. pallidum* cells ( $2 \times 10^{10}$  mL<sup>-1</sup>) in Tris/EDTA (TE) buffer (10 mM Tris, 1 mM EDTA; pH 8.0) was mixed with molten 1.6% low-melting-point InCert agarose (FMC BioProducts) and 200  $\mu$ L was applied into the plug molds. The resulting  $15 \times 9 \times 1.5$  mm plugs were then gently removed and put into 30 mL of TE buffer supplemented with 0.5% SDS and incubated overnight at 37°C. Subsequently, proteinase K (Sigma) was added to a final concentration of 100  $\mu$ g mL<sup>-1</sup> and plugs were incubated at 55°C for an additional 48 h. Plugs were then washed four times with TE buffer for each wash step.

**Preparation of pBeloBACII Vector**

*E. coli* VCS257 carrying pBeloBAC11 was inoculated into 1 L of TY medium containing 25  $\mu$ g of chloramphenicol per mL. Cells were harvested and plasmid DNA was isolated by a modified QIAGEN Plasmid Protocol for isolation of BAC DNA (QIAGEN). Cells were carefully resuspended to 100 mL of P1 buffer (50 mM Tris-Cl, pH 8.0; 10 mM EDTA; 100  $\mu$ g mL<sup>-1</sup> RNase A), lysed with P2 buffer (200 mM NaOH, 1% SDS), and cell components were precipitated with buffer P3 (3.0 M potassium acetate, pH 5.5). Cell debris was removed by two-step centrifugation at 20,000g at 4°C for 20 min. The clear supernatant was applied to a QBT-equilibrated QIAGEN-tip 500 column, washed twice with QC buffer, and vector DNA was eluted with QF buffer according to the manufacturer’s recommendations. The QBT, QC, and QF buffers were supplied with the QIAGEN plasmid purification kit. Eluted DNA was precipitated with isopropanol, washed with 70% ethanol, and resuspended to 100  $\mu$ L of distilled water. The DNA concentration was determined in a fluorometer.

**Partial Digestion of *T. pallidum* Chromosomal DNA with *Hind* III and Size Selection**

Partial digestion was performed as described previously (Brosch et al. 1998). Chromosomal DNA-containing plugs were equilibrated in three 1-h steps in 10 mL of *Hind* III buffer 2 (New England Biolabs) supplemented with 0.1% Triton X-100. Subsequently, each plug was transferred into 1 mL of *Hind* III-containing buffer 2 (20 U mL<sup>-1</sup>) and incubated 2 h on ice. After this equilibration step, plugs were incubated at 37°C for 30 min. After 30 min, further digestion was stopped by adding 0.5 mL of 50 mM EDTA (pH 8.0) to 1 mL of *Hind*

**Table 3.** Screening of the *Treponema pallidum* BAC Genomic Library with Pooled Rabbit Sera

| Clone | Antigen identified | Possible ORFs   | Most probable ORF | Function of predicted antigen                            |
|-------|--------------------|-----------------|-------------------|--|
| TP085 | 55 kDa             | 0024–0034       | TP0030            | GroEL <sup>a</sup> —heat shock protein 60 (58kDa)        |
| TP021 | 68 kDa             | 0208–0216       | TP0216            | DnaK <sup>b</sup> —heat shock protein 70 (68kDa)         |
| TP029 | 38 kDa             | 0253–0259       | TP0257            | GlpQ—glycerophosphoryldiester phosphodiesterase (41 kDa) |
|       | 33 kDa             | 0304; 0317–0323 | TP0319            | TmpC—membrane lipoprotein (37.8 kDa)                     |
| TP201 | 30 kDa             | 0860–0864       | TP0862            | peptidyl-prolyl cis-trans isomerase, FKBP-type (28 kDa)  |
| TP076 | 36 kDa             | 0965            | TP0965            | membrane fusion protein, putative (35 kDa)               |
| TP094 | 47 kDa             | 0574–0595       | TP0574            | carboxypeptidase (48 kDa)                                |
| TP003 | 90 kDa             | 0429–0437       | TP0435            | lipoprotein, 17 kDa (tpp 17)                             |
|       | 55 kDa             | 0477–0486       | TP0486            | antigen, p83/100 (53 kDa)                                |
| TP173 | 75 kDa             | 0747–0755       | TP0748            | CfpA—cytoplasmic filament protein A (79 kDa)             |
|       | 43, 35, 25 kDa     | 0768–0773       | TP0768            | TmpA—membrane protein (37 kDa)                           |
|       |                    |                 | TP0769            | TmpB—outer membrane protein (37 kDa)                     |

<sup>a</sup>GroEL is closely related to heat shock protein homologs from more than 60 different bacteria.

<sup>b</sup>Sequence of DnaK proteins are highly conserved among many organisms, e.g., *T. pallidum* DnaK is more than 58% identical to *E. coli* DnaK (Bardwell and Craig 1984).

III-containing buffer 2. Plugs with partially digested DNA were placed in the wells of a 1% agarose (I.D.NA agarose; Bio-Whittaker Molecular Applications) gel and subjected to pulsed field gel electrophoresis (PFGE) using the CHEF DR II apparatus (Bio-Rad Laboratories). Gels were run in 0.5x Tris/acetate (TAE) buffer (20 mM Tris-acetate, 0.5 mM EDTA, pH 8.3) at 14°C and 6V cm<sup>-1</sup> for 16 h with a 5–45-sec pulse time at a 120° angle. Lanes containing digested genomic DNA were excised in regions corresponding to DNA fragment sizes of 40–200 kb and gel slices were stored in 0.5 M EDTA (pH 8.0) at 4°C.

### Electroelution, Ligation, Dialysis, and Electroporation

Electroelution of digested genomic DNA from gel slices was performed according to Strong et al. (1997). The gel slice was first equilibrated with 50 mL of 1.0 x TAE buffer at 4°C for 3 h and subsequently transferred into dialysis tubing with one-fourth- to three-quarter-inch diameter (Life Technologies) with 200–400 µL of fresh 1.0 x TAE buffer. The DNA was eluted from the gel at 2.5 V cm<sup>-1</sup> for 2 h and at the end of elution, the polarity of current was reversed for 30 sec. The eluted DNA was either directly used for ligation or stored at 4°C. Then, 10 ng of size-selected *T. pallidum* DNA was ligated to 1 ng of *Hind* III-digested and dephosphorylated pBeloBAC11 DNA at insert to vector molar ratio of 1 : 10. The pBeloBAC11 plasmid was digested with *Hind* III at 37°C for 2 h and dephosphorylated for an additional 30 min with calf intestinal phosphatase (New England Biolabs). Ligation was performed overnight at 16°C with 20 U of T4 DNA ligase (New England Biolabs). T4 DNA ligase was inactivated at 65°C for 10 min and the ligation solution was then drop-dialyzed against TE buffer using VSWP 0.025 µm membranes (Millipore). Fifty µL of electrocompetent cells (*E. coli* DH10B) were mixed with 1 µL of ligation mixture in a 0.2-cm gap electrode cuvette on ice. A Gene Pulse Controller II apparatus (Bio-Rad Laboratories) set to 2.5 kV, 25 µF, and 100 Ω was used. Immediately after electroporation, 0.6 mL SOC medium (2% bacto-tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM MgCl<sub>2</sub>, 10 mM MgSO<sub>4</sub>, 20 mM glucose) was added and cells were grown at 37°C for 1 h. Cells were then plated on chloramphenicol TY plates supplemented with IPTG and X-Gal. Plates were incubated at 37°C for up to 48 h, and white colonies were isolated and used for further investigations.

### Isolation of BAC DNA and Sequencing

For isolation of BAC DNA, the same procedure as for isolation of pBeloBAC11 was used with the exception that a 10 mL volume of overnight culture, 2 mL of P1, P2, and P3 buffers, QIAGEN-tip 20, and 20 µL of distilled water for final resuspension of DNA were used. The isolated BAC DNA (9 µL) was used as a template for DNA sequencing reactions. DNA was sequenced using the *Taq* Dye-deoxy Terminator method and a model 377 DNA sequencing system (Applied Biosystems). Two PCR primers with target sites on pBeloBAC11 were used to sequence both insert termini, GW386: 5'-ttgtaaaacgacggccatg-3' and GW387: 5'-ttacccaagctatttagtgac-3'.

### Restriction Analysis of BAC Clones

Standard methods were used for restriction endonuclease analysis and agarose gel electrophoresis (Sambrook et al. 1989).

### Western Blot Analysis

Western blot analysis was performed after semidry electrotransfer of proteins from the SDS-polyacrylamide slab gel to PVDF membranes (Millipore). Bacterial proteins were detected with 1 : 1000 diluted pooled rabbit anti-*T. pallidum* serum and goat 1 : 1000 antirabbit antibodies conjugated with

horseradish peroxidase (Rockland Immunochemicals). Proteins were visualized with a chemiluminescent detection kit (ECL, Amersham Pharmacia Biotech).

### Computer-Assisted Sequence Analysis

Computer-assisted sequence analysis was performed using the LASERGENE package (DNASTAR).

### ACKNOWLEDGMENTS

This work was supported by grants from the U.S. Public Health service to G.M.W. (R01 DE12488 and R01 DE13759).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

### REFERENCES

- Bailey, M.J., Thomas, C.M., Cockayne, A., Strugnell, R.A., and Penn, C.W. 1989. Cloning and expression of *Treponema pallidum* antigens in *Escherichia coli*. *J. Gen. Microbiol.* **135**: 2365–2378.
- Bardwell, J.C. and Craig, E.A. 1984. Major heat shock gene of *Drosophila* and the *Escherichia coli* heat-inducible *dnaK* gene are homologous. *Proc. Natl. Acad. Sci.* **81**: 848–852.
- Baseman, J.B. and Hayes, N.S. 1974. Protein synthesis by *Treponema pallidum* extracted from infected rabbit tissue. *Infect. Immun.* **10**: 1350–1355.
- Brayton, K.A., De Villiers, E.P., Fehrson, J., Nxomani, C., Collins, N.E., and Allsopp, B.A. 1999. *Cowdria ruminantium* DNA is unstable in a SuperCos1 library. *Onderstepoort J. Vet. Res.* **66**: 111–117.
- Brosch, R., Gordon, S.V., Billault, A., Garnier, T., Eiglmeier, K., Soravito, C., Barrell, B.G., and Cole, S.T. 1998. Use of a *Mycobacterium tuberculosis* H37Rv bacterial artificial chromosome library for genome mapping, sequencing, and comparative genomics. *Infect. Immun.* **66**: 2221–2229.
- Cao, G.J. and Sarkar, N. 1992. Identification of the gene for an *Escherichia coli* poly(A) polymerase. *Proc. Natl. Acad. Sci.* **89**: 10380–10384.
- Dewar, K., Sabbagh, L., Cardinal, G., Veilleux, F., Sanschagrín, F., Birren, B., and Levesque, R.C. 1998. *Pseudomonas aeruginosa* PAO1 bacterial artificial chromosomes: Strategies for mapping, screening, and sequencing 100 kb loci of the 5.9 Mb genome. *Microb. Comp. Genomics* **3**: 105–117.
- Fraser, C.M., Norris, S.J., Weinstock, G.M., White, O., Sutton, G.G., Dodson, R., Gwinn, M., Hickey, E.K., Clayton, R., Ketchum, K.A., et al. 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**: 375–388.
- Gherardini, F.C., Hobbs, M.M., Stamm, L.V., and Bassford, P.J. 1990. Complementation of an *Escherichia coli* *proC* mutation by a gene cloned from *Treponema pallidum*. *J. Bacteriol.* **172**: 2996–3002.
- Grant, S.G., Jessee, J., Bloom, F.R., and Hanahan, D. 1990. Differential plasmid rescue from transgenic mouse DNAs into *Escherichia coli* methylation-restriction mutants. *Proc. Natl. Acad. Sci.* **87**: 4645–4649.
- Hanff, P.A., Norris, S.J., Lovett, M.A., and Miller, J.N. 1984. Purification of *Treponema pallidum*, Nichols strain, by Percoll density gradient centrifugation. *Sex. Transm. Dis.* **11**: 275–286.
- Hardham, J.M., Frye, J.G., and Stamm, L.V. 1995. Identification and sequences of the *Treponema pallidum* *fliM'*, *fliY*, *fliP*, *fliQ*, *fliR* and *fliH'* genes. *Gene* **166**: 57–64.
- Hindle, Z., Callis, R., Dowden, S., Rudd, B.A., and Baumberg, S. 1994. Cloning and expression in *Escherichia coli* of a *Streptomyces coelicolor* A3(2) *argCJB* gene cluster. *Microbiology* **140**: 311–320.
- Kim, U.J., Birren, B.W., Slepak, T., Mancino, V., Boysen, C., Kang, H.L., Simon, M.I., and Shizuya, H. 1996. Construction and characterization of a human bacterial artificial chromosome library. *Genomics* **34**: 213–218.
- Norris, S.J. 1993. Polypeptides of *Treponema pallidum*: Progress toward understanding their structural, functional, and immunologic roles. *Treponema Pallidum* Polypeptide Research Group. *Microbiol. Rev.* **57**: 750–779.
- Rondon, M.R., Raffel, S.J., Goodman, R.M., and Handelsman, J. 1999. Toward functional genomics in bacteria: Analysis of gene expression in *Escherichia coli* from a bacterial artificial

- chromosome library of *Bacillus cereus*. *Proc. Natl. Acad. Sci.* **96**: 6451–6455.
- Sambrook, J., Fritsch, E.F., and Maniatis, T. 1989. *Molecular cloning: A laboratory manual*, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Stamm, L.V. and Bergen H.L. 2000. The sequence-variable, single-copy *tpkK* gene of *Treponema pallidum* Nichols strain UNC and Street strain 14 encodes heterogeneous TprK proteins. *Infect. Immun.* **68**: 6482–6486.
- Strong, S.J., Ohta, Y., Litman, G.W., and Amemiya, C.T. 1997. Marked improvement of PAC and BAC cloning is achieved using electroelution of pulsed-field gel-separated partial digests of genomic DNA. *Nucleic Acids Res.* **25**: 3959–3961.
- Tomkins, J.P., Miller-Smith, H., Sasinowski, M., Choi, S., Sasinowska, H., Verce, M.F., Freedman, D.L., Dean, R.A., and Wing, R.A. 1999. Physical map and gene survey of the *Ochrobactrum anthropi* genome using bacterial artificial chromosome contigs. *Microb. Comp. Genomics* **4**: 203–217.
- Walker, E.M., Howell, J.K., You, Y., Hoffmaster, A.R., Heath, J.D., Weinstock, G.M., and Norris, S.J. 1995. Physical map of the genome of *Treponema pallidum* subsp. *pallidum* (Nichols). *J. Bacteriol.* **177**: 1797–1804.
- Weinstock, G.M., Hardham, J.M., McLeod, M.P., Sodergren, E.J., and Norris, S.J. 1998. The genome of *Treponema pallidum*: New light on the agent of syphilis. *FEMS Microbiol. Rev.* **22**: 323–332.
- Xu, Y., Murray, B.E., and Weinstock, G.M. 1998. A cluster of genes involved in polysaccharide biosynthesis from *Enterococcus faecalis* OG1RF. *Infect. Immun.* **66**: 4313–4323.

Received July 23, 2001; accepted in revised form December 20, 2001.