2003

Analysis of Conserved Non-rRNA Genes of *Tropheryma whipplei*

Matthias Maiwald
Stanford University School of Medicine, maiwald@cmgm.stanford.edu

Paul W. Lepp
Stanford University School of Medicine, paul.lepp@minotstateu.edu

David A. Relman
Stanford University School of Medicine, relman@stanford.edu

System. Appl. Microbiol. 26, 3–12 (2003)
© Urban & Fischer Verlag
http://www.urbanfischer.de/journals/sam

SYSTEMATIC AND APPLIED MICROBIOLOGY

Analysis of Conserved Non-rRNA Genes of *Tropheryma whipplei*

Matthias Maiwald[1], Paul W. Lepp[1,2], and David A. Relman[1,2,3]

Departments of Microbiology and Immunology[1] and of Medicine[2], Stanford University School of Medicine, Stanford, California, USA
Veterans Affairs Palo Alto Health Care System, Palo Alto, California, USA[3]

Received: January 7, 2003

Summary

The causative agent of Whipple's disease, *Tropheryma whipplei*, is a slow-growing bacterium that remains poorly-understood. Genetic characterization of this organism has relied heavily upon rRNA sequence analysis. Pending completion of a complete genome sequencing effort, we have characterized several conserved non-rRNA genes from *T. whipplei* directly from infected tissue using broad-range PCR and a genome-walking strategy. Our goals were to evaluate its phylogenetic relationships, and to find ways to expand the strain typing scheme, based on rDNA sequence comparisons. The genes coding for the ATP synthase beta subunit (*atpD*), elongation factor Tu (*tuf*), heat shock protein GroEL (*groEL*), beta subunit of DNA-dependent RNA polymerase (*rpoB*), and RNase P RNA (*rnpB*) were analyzed, as well as the regions upstream and downstream of the rRNA operon. Phylogenetic analyses with all non-rRNA marker molecules consistently placed *T. whipplei* within the class, *Actinobacteria*. The arrangement of genes in the *atpD* and *rpoB* chromosomal regions was also consistent with other actinomycete genomes. Tandem sequence repeats were found upstream and downstream of the rRNA operon, and downstream of the *groEL* gene. These chromosomal sites and the 16S-23S rRNA intergenic spacer regions were examined in the specimens of 11 patients, and a unique combination of tandem repeat numbers and spacer polymorphisms was found in each patient. These data provide the basis for a more discriminatory typing method for *T. whipplei*.

Key words: *Tropheryma* – Whipple's disease – phylogeny – genome walking – strain typing

Introduction

Whipple's disease is a systemic illness characterized by the presence of monomorphic bacteria and a macrophage infiltrate in affected organs and tissues. Since the first description of this disease in 1907, many attempts have been undertaken to cultivate this bacterium, and many have failed. In 1991–92, the Whipple's disease bacterium was characterized based on its 16S rRNA sequence, using a broad range polymerase chain reaction (PCR) approach and was shown to be a member of the *Actinobacteria* [32, 48]. Co-cultivation of the bacterium in the presence of a human fibroblast cell line was reported in 2000 [30].

Due to the long-standing absence of purified and culturable bacterial cells, little information has been available regarding the phenotypic and genotypic characteristics of the Whipple bacillus, *Tropheryma whipplei*. Many aspects of its natural ecology, routes of transmission, and pathogenicity are unclear. A detailed assessment of its phylogeny based on 16S rDNA sequence analysis revealed an intermediate position between a group of actinomycetes with group B peptidoglycan and the family *Cellulomonadaceae* [26]. The 16S-23S ribosomal intergenic spacer exhibits limited sequence variability, with seven types so far described, and is now the basis for a strain typing scheme [8, 15, 26, 28]. The overall organization of the *T. whipplei* rRNA operon is in general accordance with that of the other actinomycetes, but some features, such as a 23S rRNA insertion sequence and predicted rRNA secondary structures are quite dissimilar to the corresponding features of other actinomycetes [28].

Beyond the rRNA operon, only scant information has been available regarding genes, genetic organization, and predicted gene products. A 620 bp fragment of the *groEL* heat shock protein gene has been determined and employed in a diagnostic PCR test [29]. In addition, a com-

0723-2020/03/26/01-003 $ 15.00/0

plete *rpoB* sequence (beta subunit of DNA-dependent RNA polymerase) was obtained from a laboratory-propagated isolate of *T. whipplei* [6], and used for the same purpose. A phylogenetic analysis of the RpoB protein sequence provided further evidence for a relationship of *T. whipplei* with the actinobacteria.

In order to verify this phylogenetic assignment of *T. whipplei* and enhance strain discrimination capabilities, we targeted conserved genes using degenerate broad-range PCR and combined this with a "genome walking" strategy. Our targets included the genes coding for ATP synthase beta subunit (*atpD*), elongation factor Tu (*tuf*), heat shock protein GroEL (*groEL*), and the beta subunit of DNA-dependent RNA polymerase (*rpoB*), and the RNase P RNA (*rnpB*) gene. The genome walking strategy was also applied to the regions upstream and downstream of the rRNA operon [28]. Despite their conservation, these genes offer reliable phylogenetic information, and the possibility of useful sequence polymorphisms. Examination of these chromosomal regions also provides an early, but limited glimpse at the genome organization of this enigmatic bacterium.

Materials and Methods

Patient specimens

A paraffin-embedded intestinal biopsy specimen from a patient with "classical" intestinal Whipple's disease was used to amplify and assemble all of the chromosomal regions in this study. This specimen had been obtained from a 72-year-old female with a diagnosis established in 1999 based on standard histopathology criteria, and prior to antibiotic treatment. A group of specimens was used to confirm the *T. whipplei* sequence data obtained in this study: 3 intestinal endoscopic biopsy and 3 cerebrospinal fluid (CSF) specimens from patients with Whipple's disease confirmed by both histology and PCR. In addition, a group of 3 intestinal biopsy and 3 CSF specimens from patients who had no evidence of Whipple's disease by histology or PCR was used to establish the specificity of these findings. Ten additional specimens (5 intestinal biopsies and 5 CSF samples) from patients with confirmed Whipple's disease provided bacterial genomic template for analysis of variable sequence sites identified in this study. DNA from the specimens was extracted as described previously [45, 46].

PCR

Several genes that are conserved across a wide range of organisms were selected as targets for consensus PCR: ATP synthase beta subunit (*atpD*), elongation factor Tu (*tuf*), heat shock protein GroEL (*groEL*), beta subunit of RNA polymerase (*rpoB*), and RNase P RNA (*rnpB*). Sequences for each of these genes from a representative set of bacteria were aligned. Degenerate broad-range primers (Table 1) were designed and used in an initial series of PCR experiments. PCR products were cloned into the TA vector system (Invitrogen, Palo Alto, CA, USA), sequenced using standard reagents and equipment (ABI, Foster City, CA, USA), and analyzed using the BLAST software package (http://www.ncbi.nlm.nih.gov/BLAST/). Sequences with similarity to those of other actinomycetes were incorporated into the alignments, and specific internal primers (Table 1) were constructed. The primers were then tested in PCRs with the 6 Whipple's disease and 6 negative control specimens, in order to confirm that the novel sequences were associated with *T. whipplei*.

Genome "walking"

To examine sequence adjacent to regions bounded by the conserved priming sites, specific primers were designed such that their orientation was facing outwards. These primers were initially used in combination with primers designed from more distant conserved sequence regions. This process was repeated until no more adjacent sequence with sufficient conservation could be identified within the aligned data set. As a final step, the newly acquired chromosomal sequence regions were then further extended using restriction site PCR [36]. A second confirmation procedure was then performed to assess the specificity of the sequences at the ends of the chromosomal regions: specific primers facing inwards from the ends of the newly acquired sequence were combined with the previous set of specific outward-facing primers, and tested in PCRs with the 6 Whipple's disease and 6 negative control specimens. The overall PCR strategy is shown in Figure 1. Examination of polymorphic sites in other Whipple's disease specimens was accomplished using additional sets of primers (Table 1).

Sequence analysis

Alignments and phylogenetic analyses were performed using the ARB program package [39]. Clustal W (Genetics Computing Group, Wisconsin) and Iteralign ([1]; http://giotto.stanford.edu/~luciano/iteralign.html) were used for auxiliary protein sequence alignments. A prealigned set of type A RNase P RNA sequences was downloaded from the RNase P RNA database [3]; http://www.mbio.ncsu.edu/RNaseP/home.html) and imported into ARB. Phylogenetic trees were based on the predicted pro-

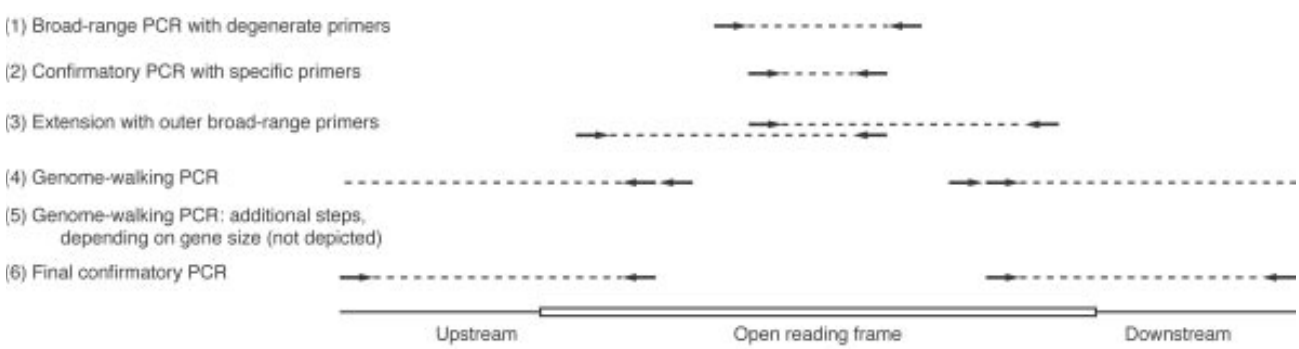

Fig. 1. General scheme of the PCR strategy used to acquire sequences of conserved genes and adjacent regions. The number of steps required to acquire each individual chromosomal region varied, depending on length and the availability of sites for outer degenerate broad-range primers.

Table 1. Key primers used for the acquisition of novel sequences from *T. whipplei* and primers for sequence-based typing of different strains.

Gene, Primer[a]	Sequence	Comments[b]
atpD		
atp508f	5′-ttyggyggygciggigtiggyaarac	Initial broad-range
atp701r	5′-ggctcrtycatytgiccraaiacca	Initial broad-range
atp544f	5′-caggaaatgatacaaagagtg	Specific, initial confirmation
atp653r	5′-gcatcctgcatcycgcgaatc	Specific, initial confirmation
atp79f	5′-ayyggiscigtcgtygacgtigaitt	Secondary broad-range
atp1205r	5′-atrcciargatigcratratrtc	Secondary broad-range
tuf		
eftu46f	5′-ggsaccatyggicayrtygacca	Initial broad-range
eftu671r	5′-ccggtratigwgaagacgtcctc	Initial broad-range
eftu114f	5′-gaggcttccttcgaatacg	Specific, initial confirmation
eftu571r	5′-taaccgcatcgccccacttttg	Specific, initial confirmation
eftu1077r	5′-ggygttgtcrciggcatsaccat	Secondary broad-range
groEL		
whipp-frw1	5′-tgacgggaccacaacatctg	Initial specific [29]
whipp-rev	5′-acatcttcagcaatgataagaagtt	Initial specific [29]
gro64f	5′-ctsgcsgayrcsgtiaaggtsac	Secondary broad-range
gro1492r	5′-gggtsaccttsaccgggtc	Secondary broad-range
gro1390f	5′-ggtaaggtatcttctttacctc	Sequence-based typing
gro+416r	5′-aataccgaaaatatggagggtag	Sequence-based typing
rpoB		
rpob1303f	5′-cagctgwsicarttcatggaccaramcaaycc	Initial broad-range
rpob1844r	5′-gcytgicgytgcatgttigmicccat	Initial broad-range
rpob1497f	5′-tttagcgtgctattccagggtg	Specific, initial confirmation
rpob1679r	5′-agaacacgctcatcaacaaacg	Specific, initial confirmation
rpob470f	5′-cggaattcctyccnhtnatgacnga	Secondary broad-range
rpob2674r	5′-cggaattccccyttrttnccrtgncgnccngc	Secondary broad-range
rpob3221r	5′-cgggatcccgcccanmmytccat	Tertiary broad-range
rnpB		
rnsp59f	5′-giigaggaaagtcciigc	Initial broad-range [9]
rnsp347r	5′-rtaagccggrttctgt	Initial broad-range [9]
rnsp114f	5′-cacccgggataacccgagagctg	Specific, initial confirmation
rnsp308r	5′-taaacgagcagcctaagttccctg	Specific, initial confirmation
rRNA operon upstream region		
tw70r	5′-caaggaccgacaggacgaacc	Sequence-based typing
op-432f	5′-atggaccaataccaaaggaac	Sequence-based typing
rRNA operon downstream region		
tw5670f	5′-cctcaaacccaagcttattcgcc	Sequence-based typing
op+276r	5′-caaaagaatatctataagcac	Sequence-based typing
Restriction-site PCR		
*Nhe*I-RSO	5′-aatacgactcactataggnnnnnnnnnngctagc	Restriction-site oligonucleotide
*Pst*I-RSO	5′-aatacgactcactataggnnnnnnnnnnnctgcag	Restriction-site oligonucleotide
*Msc*I-RSO	5′-aatacgactcactataggnnnnnnnnnnntggcca	Restriction-site oligonucleotide

[a] Numbering of the primers for *atpD*, *tuf*, *groEL*, *rpoB*, and *rnpB* is based on the numbering in the corresponding genes of *Mycobacterium tuberculosis*, as a reference.

[b] Primers were designed in this study, except the 5′ portions of restriction-site oligonucleotides [36] and where other references are given.

Table 2. Overview of the codon usage and total G+C content of the concatenated genes *atpD*, *tuf*, *groEL*, and *rpoB*.

Organism	Total G+C content	Third position G+C content	No. of unused codons (of 64)
T. whipplei	50%	48%	1 (stop TGA)
M. tuberculosis	64%	86%	3
M. leprae	60%	77%	1 (Arg)
Strep. coelicolor	66%	94%	13 (2 stop)
Staph. aureus	38%	21%	7
B. subtilis	46%	38%	4
E. coli	53%	57%	7

tein sequences, and for RNase P RNA, on DNA sequences, and were calculated using the Neighbor-Joining, Maximum Parsimony, and Maximum Likelihood algorithms [7]. Tree topologies were evaluated by bootstrap analysis with 100 re-samplings.

Sequences at the NCBI microbial genomes website (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html) and a set of *Streptomyces coelicolor* cosmids [31] were used for comparisons of *T. whipplei* gene arrangements with those of other microbial genomes. Sequences were tested for open reading frames (ORFs) using the ORF finder at the NCBI website, the Glimmer program (TIGR, The Institute for Genomic Research, Rockville, MD), and the Artemis annotation tool (Sanger Institute, Cambridge, UK).

The following additional sequence analysis resources were used: for codon usage and nucleotide composition, Codonfrequency and Composition (GCG package); for protein family assignment, Pfam (http://pfam.wustl.edu); for transmembrane protein prediction, TMpred (http://www.ch.embnet.org/software/TMPRED_form.html); for signal peptide prediction, SignalP (http://www.cbs.dtu.dk/services/SignalP); to search for blocks of conserved sequences, Block Searcher ([12]; http://blocks.fhcrc.org/blocks/blocks_search.html), to search for tRNAs, tRNAscan SE (http://www.genetics.wustl.edu/eddy/tRNAscan-SE); and to search for repeat regions, Tandem Repeats Finder (http://c3.biomath.mssm.edu/trf.html).

Nucleotide sequence data deposition

The sequences of the chromosomal regions determined in this work have been deposited in the GenBank/EMBL databases under accession nos. AF483648-AF483654. The secondary structure of RNase P RNA was determined by James W. Brown (North Carolina State University) and was deposited at the RNase P RNA website (http://www.mbio.ncsu.edu/RNaseP/home.html).

Results and Discussion

Assembled sequences

Multiple locus genome sequence analysis has significantly enhanced our understanding of the evolution and functional capabilities of fastidious microorganisms. *T. whipplei* is a pre-eminent example of a microorganism that has resisted characterization because of its recalcitrance to cultivation. Consensus PCR and genome-walking techniques provide an opportunity for genetic characterization of *T. whipplei* directly from a relevant infected site, i.e. the diseased human intestinal lamina propria. The recent published reports of *T. whipplei* laboratory cultivation describe bacterial generation times as long as 18 days. Our approach for genetic characterization avoids mutations that might arise during prolonged laboratory propagation and become fixed as a result of laboratory adaptation.

DNA sequences from each of 7 unlinked chromosomal loci were assembled based on PCR products amplified from an intestinal biopsy specimen of a patient with Whipple's disease. These chromosomal regions encompass the ATP synthase (*atpD*) gene (assembly size 2392 bp), elongation factor Tu (*tuf*) gene (assembly size 1865 bp), heat shock protein GroEL (*groEL*) gene (assembly size 2551 bp), DNA dependent RNA polymerase (*rpoB*) gene (assembly size 3815 bp), RNase P RNA (*rnpB*) gene (assembly size 568 bp), and the regions upstream (802 bp) and downstream (1200 bp) of the rRNA operon. The newly-determined sequences comprise 13,193 bp with a G+C content of 49% (individual regions: 46.1%, 50.0%, 49.2%, 50.2%, 51.8%, 47.1%, 49.7%, respectively), which is lower than the estimated G+C content of a cultivated isolate (59%) using HPLC [22], and in keeping with a figure of 46.3% from a *T. whipplei* full genome sequencing project (J. Parkhill, unpublished data; http://www.sanger.ac.uk/Projects/T_whippelii/). tRNA genes were not identified. The 16S-23S rRNA intergenic spacer sequence from our index patient was type 2 [15, 28].

To assess whether the amplified and assembled sequences were specific to Whipple's disease tissues and hence, to *T. whipplei*, PCRs were performed with primers designed from specific sequences at the center of the assembled regions and primers from the ends (Table 1, Fig. 1). All PCRs with these specific primer pairs yielded products of the expected size from all 6 Whipple's disease tissues and from none of the six non-Whipple's disease tissues.

T. whipplei codon usage and chromosomal G+C content was compared to a group of well-studied actinobacteria and other bacteria using a sequence data set from *Mycobacterium tuberculosis*, *Mycobacterium leprae*, *S. coelicolor*, *Staphylococcus aureus*, *Bacillus subtilis*, and *Escherichia coli*. The data set consisted of concatenated sequences for *atpD*, *tuf*, *groEL*, and *rpoB* (7716–8244 bp in the different species). The results are shown in Table 2. While the third-position G+C content generally reflected overall G+C content, the small number (i.e., one) of unused codons in the *T. whipplei* sequence set is unusual. The evenness of codon usage (or degree of bias) in an organism may be correlated with its growth rate [19]. The relative level of expression of a gene is also thought to be

a determinant of the degree of codon bias displayed by that gene. All four genes in our analysis are among the top 20 predicted most highly expressed genes based on codon usage, as determined in an analysis of the genomes of four fast-growing bacteria [19]. Thus, the relatively unbiased use of codons in these *T. whipplei* genes supports the observation from *in vitro* cultivation experiments that this is a very slowly growing organism [30].

ATPase gene region

The ATPase chromosomal region (2392 bp) contains an ORF that is predicted to encode an F_1F_0 type ATP synthase beta subunit (*atpD* gene) with 474 predicted amino acids (aa) and a calculated molecular mass of 52 kDa. The AtpD putative protein sequence displays typical conserved amino acid signatures for this family of proteins: Walker A (P-loop), GGAGVGKTV; Walker B, LLFID; and the DELSEED sequence ([35, 47]; Karlheinz Altendorf, University of Osnabrück, Germany, personal communication). Upstream of *atpD*, there is a 485 bp (160 aa) ORF that is predicted to encode the C-terminal portion of the ATP synthase gamma subunit (*atpG*), and downstream of *atpD*, there is a complete ORF (285 bp, 94 aa) encoding an ATP synthase epsilon subunit (*atpC*).

In bacteria which have F_1F_0 type ATP synthases, the genes for the subunits of this complex are usually arranged in an operon and in the order *atpIBEFHAGDC* [13]. In accordance with this pattern, the characterized genes from the *atp* operon of *T. whipplei* are organized in the order *atpGDC*. The putative epsilon subunit (AtpC) of *T. whipplei* (94aa) is significantly shorter than the epsilon subunits of related actinobacteria, *M. tuberculosis* (121 aa), *M. leprae* (121 aa), *S. coelicolor* (124 aa), as well as many other bacteria for which this subunit has been characterized. A Block search [12] identified only 1 of 2 conserved AtpD sequence motif blocks. An epsilon subunit (86 aa) of similarly unusual size is found in *Caulobacter crescentus*, and in like fashion contains only 1 of 2 sequence blocks. The biological significance of the apparently truncated *T. whipplei* and *C. crescentus* AtpC proteins remains unclear.

A putative ORF downstream of *atpC* in the *T. whipplei* genome displayed little similarity with ORFs at the same location in the genomes of other actinobacteria, although all showed evidence for a signal sequence and a transmembrane region. The Iteralign program [1] recognized a highly similar region in the sequences of *T. whipplei* (ALYIIGLFAF), *S. coelicolor* and *S. lividans* (both AVVVIGLFVF), starting at residue 15 of all three proteins. This suggests that these predicted products encoded by these ORFs may be analogs, but it remains unclear if their gene products are functionally related to the ATP synthases.

Elongation factor-Tu gene region

The elongation factor Tu chromosomal region (1865 bp) contains an ORF predicted to encode a typical Tuf protein (397 aa, 43.5 kDa). This Tuf sequence contains a conserved P-loop motif: IGHVDHGKTT [35]. Within the 151 bp upstream and 520 bp downstream of *tuf*, no ORF was found that yielded a hit with an E-value better than 10^{-1} in Blast searches.

Bacterial *tuf* genes are commonly located within an *rpsL* operon, containing the genes for the ribosomal proteins S12 (*rpsL*) and S7 (*rpsG*), elongation factor G (*fus*), and elongation factor Tu (*tuf*), in that order. Many proteobacteria (e.g., *E. coli*) have two almost identical *tuf* alleles, both expressed at high levels, of which only *tufA* is located in the *rpsL* operon [16, 38]. Most low G+C Gram-positive bacteria and actinobacteria have one *tuf* allele, which is located downstream of *fus* in the *rpsL* operon. For comparison, *fus-tuf* intergenic regions range from 109–362 bp in *M. tuberculosis*, *M. leprae*, *S. coelicolor*, and *C. glutamicum*. Two streptomycetes are exceptions to the single *tuf* allele pattern within the actinobacteria; *Streptomyces ramocissimus* has three *tuf* alleles (*tuf1*, *tuf*2, *tuf3*), and *S. coelicolor* has two (*tuf1*, *tuf3*). The *tuf2* and *tuf3* alleles are not expressed at detectable levels, nor located in the *rpsL* operon, and *tuf3* is quite dissimilar (63–65% aa identities) to the other alleles [42, 44]. However, apart from the streptomycetes, no other organism within the *Actinobacteria* has been found so far with more than one *tuf* allele. A preliminary analysis of the *T. whipplei* genome reveals only one *tuf* gene; the rest of the *rpsL* operon is found at an unlinked locus (http://www.sanger.ac.uk/Projects/T_whippelii/).

GroEL gene region

The GroEL chromosomal region (2551 bp) contains an ORF predicted to encode a typical GroEL heat shock protein (540 aa, 57 kDa). Within the 512 bp upstream and 416 bp downstream of *groEL*, no ORF was found that yielded a Blast search hit with an E-value better than 10^{-2}; nevertheless, one potential ORF (336 bp) was strongly predicted by the Glimmer program downstream of *groEL* (score 99 of 99). There were two copies of a 21 bp sequence repeat (TTTATGTCATTCTCTTGCAGA) within this potential downstream ORF. CIRCE regulatory elements are commonly found upstream of *groES* or *groEL* genes in many bacteria [10], but none was not found in the *T. whipplei* chromosomal region.

Bacterial *groEL* genes are often located in *groESL* operons, where they are located downstream of *groES*. Those actinobacteria for which information is available, have two *groEL* genes, one of which (*groEL1*) is located downstream of *groES* (*C. diphtheriae*, *M. tuberculosis*, *M. leprae*, *S. coelicolor*: 12–139 bp downstream), while the other (*groEL2*) is located elsewhere on the genome [4, 5, 31, 34]. No *groES* gene was identified within 512 bp upstream of the *T. whipplei groEL*. Furthermore, the *T. whipplei* predicted GroEL protein sequence contains a C-terminal MDF amino acid sequence. Similar glycine-methionine-rich motifs are found in actinobacteria GroEL2 proteins, whereas actinobacteria GroEL1 proteins contain histidine-rich C-terminal sequences [2, 34]. These data suggest that this *T. whipplei groEL* gene is a *groEL2* allele. Preliminary analysis of the complete *T. whipplei*

genome sequence indicates that this is the only *T. whipplei* groEL allele and confirms that it is not physically linked to the *groES* locus (unpublished data; http://www.sanger.ac.uk/Projects/T_whippelii/).

RpoB gene region

The RpoB chromosomal region (3815 bp) contains an ORF predicted to encode a typical beta subunit of a DNA-dependent RNA polymerase (1157 aa, 128 kDa). The sequence of this contig is almost identical to a previously-determined *rpoB* sequence from a cultivated *T. whipplei* isolate (Drancourt et al., 2001). However, in the currently reported analysis a GTG start codon was chosen (178 bp downstream of the previously proposed start), based on an inspection of actinobacteria RpoB alignments. Within 27 bp of the end of the *rpoB* ORF there is the beginning of a predicted gene encoding the beta' subunit of RNA polymerase (*rpoC*). The RNA polymerase genes *rpoB* and *rpoC* occur in a tandem arrangement in the majority of bacterial genomes, as is the case with the *T. whipplei* genome.

RNaseP RNA gene region

A complete gene (*rnpB*, 351 bp) for the *T. whipplei* RNase P RNA was recovered and analyzed. The predicted secondary structure of the RNA molecule was that of a typical type A RNase P RNA, but without an additional bulge in helix L15 that is present in most other actinobacteria [9]. The predicted *T. whipplei* RNase P RNA structure is available at the RNase P RNA website (http://www.mbio.ncsu.edu/RNaseP/home.html).

rRNA operon region

Using restriction site PCR (Sarkar et al., 1993), 455 bp of additional sequence was characterized upstream of the previously-sequenced rRNA operon [28]. Two copies of a 14-bp imperfect tandem repeat (AACTGWTACTGAGT) were identified in this region, but no predicted gene. As a point of comparison, the genomes of *M. tuberculosis* and *M. leprae* contain a *murA* gene (UDP-*N*-acetylglucosamine enolpyruvyl transferase) within 296 and 308 bp upstream of their single-copy 16S rRNA gene, while *S. coelicolor*, with 6 rRNA operons, has a different complement of genes upstream of its rRNA operons. Downstream of the *T. whipplei* rRNA operon, 539 bp of new sequence was acquired, thereby extending chromosomal characterization 1033 bp beyond the end of the *T. whipplei* 5S rDNA. The absence of sequence variability or ambiguity in these amplified products and extended chromosomal regions flanking the *T. whipplei* rRNA operon is in keeping with preliminary findings from the *T. whipplei* genome project that there is only one copy of an rRNA operon in this genome (http://www.sanger.ac.uk/Projects/T_whippelii/). Two different repeat regions were identified within the downstream region, one consisting of 11 repeated Gs, and the other consisting of 6 tandem copies of a 9 bp repeat (GTTCTAGTA). These variable number tandem repeat loci provide the basis for a new potentially useful *T. whipplei* strain typing scheme (see below).

T. whipplei phylogeny

All of the molecules characterized in this study (AtpD, Tuf, GroEL, RpoB, and *rnpB*) have previously been used as phylogenetic markers [9, 21, 23, 33, 43, 49]. The use of non-rRNA markers has been proposed for independent assessment of 16S rRNA-based relationships [24, 25], and may provide differing levels of taxonomic resolution and differing perspectives on organismal history [25, 33, 49]. The results of separate analyses using each of the 5 genetic loci are shown in Fig. 2. A tree based on a comparative analysis of 16S rRNA sequences from a similar set of organisms is provided for comparison. The ATPase sequences of the *Chlamydia* and *Thermus/Deinococcus* phyla, and the RNase P RNA sequences of low G+C Gram-positives could not be included because they represent different (non-homologous) molecule types (V-type ATPases, type B RNAase P RNAs).

In all of these analyses the class *Actinobacteria* was monophyletic. *T. whipplei* was placed within this class in all cases, and by each of the phylogenetic algorithms. These findings confirm previous assignments based on 16S rRNA [26, 32, 48], 5S and 23S rRNA [28], and RpoB [6] analyses. Closest relatives of *T. whipplei* in the different trees were *Nocardioides jenseni* (*rnpB*), *Propionibacterium acnes* (GroEL), *Micrococcus luteus* (Tuf), and *S. coelicolor* (AtpD and RpoB); however, in contrast with 16S rDNA-based analysis [26], a fine resolution picture of relationships within the *Actinobacteria* is not possible with these alternative genetic loci, due to the much smaller number of available sequences. The branching orders at the level of different bacterial phyla varied between analyses and treeing algorithms. This is in accordance with the conclusions of previous assessments of bacterial phylogeny [17, 23, 37]. RNase P RNA is a smaller and much more variable molecule than the other markers, but it has been used to resolve differences within narrow taxonomic ranges, such as within LL-2,6-diaminopimelic acid-containing actinomycetes [49] and *Chlamydia* spp. [14]. As expected from its large size, the most stable tree topology was achieved with RpoB.

The GroEL tree (Fig. 2C) revealed different groupings for the known GroEL1 and GroEL2 proteins of mycobacteria, corynebacteria, and streptomycetes; the *T. whipplei* sequence was clearly more similar to the GroEL2, than the GroEL1, proteins of other actinobacteria. To check for a possible sequence chimera, the two halves of *T. whipplei* GroEL sequence were analyzed separately; both were more similar to GroEL2 sequences. Different phylogenetic groupings of GroEL1 and GroEL2 sequences for the actinobacteria have been noted previously [18, 43]. GroEL1 proteins appear to be subject to fewer selective or functional constraints and evolve faster than GroEL2 proteins. This is reflected by the larger evolutionary distances within the GroEL1 group in Figure 2C. Assuming a *groEL* gene duplication event during the evo-

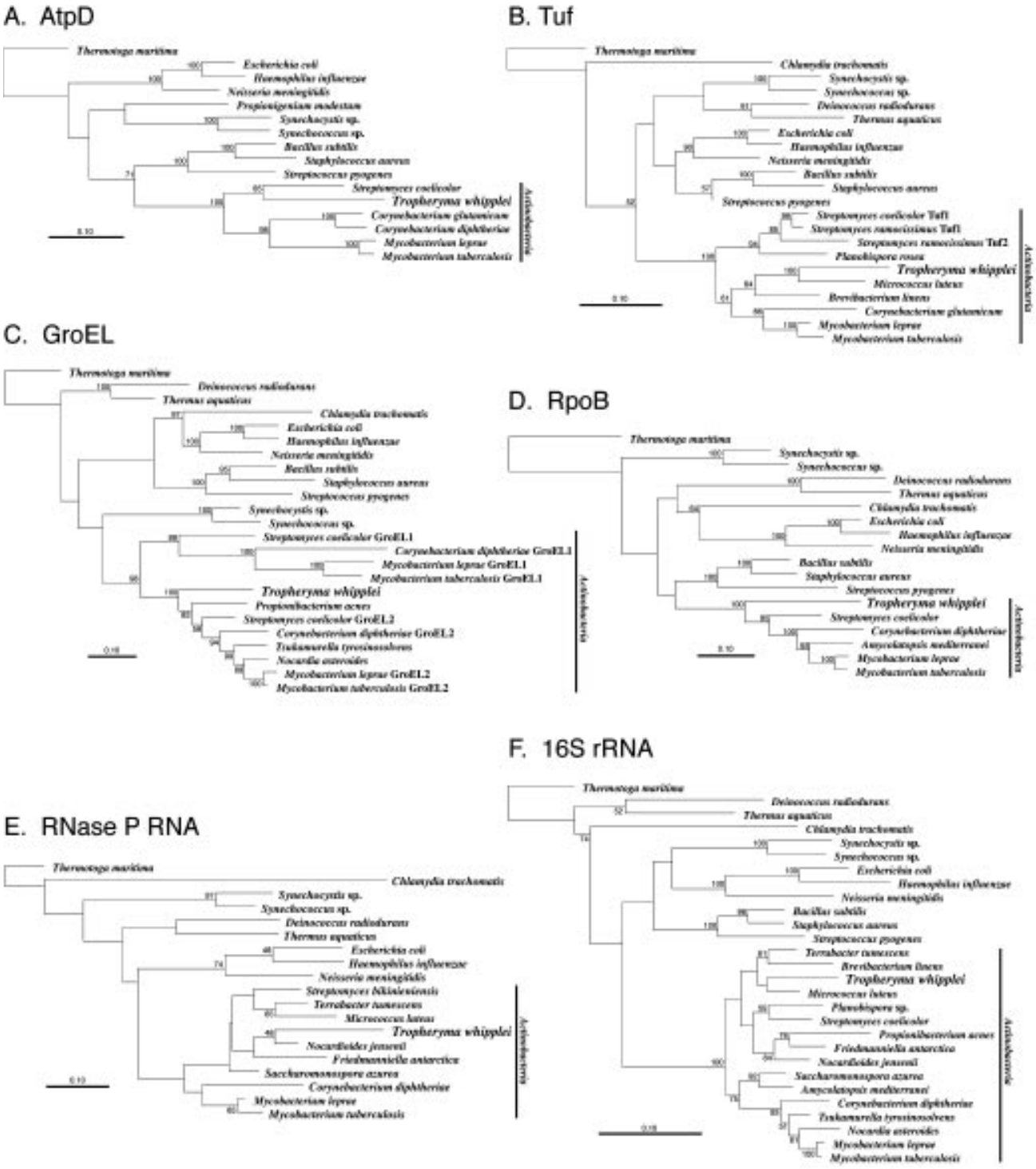

Fig. 2. Phylogenetic analyses. (A) ATP synthase beta subunit protein; (B) Elongation factor Tu protein; (C) GroEL protein; (D) RpoB protein; (E) RNase P RNA; (F) 16S rRNA. The trees in this figure were calculated using Maximum Likelihood algorithms and evaluated by bootstrap analysis with 100 re-samplings. The scale bar indicates 0.1 amino acid or nucleotide substitutions per position, and bootstrap values ≥50% are indicated. The phylum *Actinobacteria* is monophyletic in each tree and is marked by a vertical bar.

lution of the *Actinobacteria*, the absence of the *groES*-linked *groEL1* allele in the *T. whipplei* genome suggests subsequent gene loss during the evolutionary history of this organism.

Variable numbers of tandem repeats (VNTRs)

A total of 4 repeat-sequence motifs were found in the 7 *T. whipplei* chromosomal regions, including one homopolymeric G-tract and three tandem repeat sequences. The tandem repeats were located in the GroEL chromosomal region (21-mer: TTTATGTCATTCTCTTGCAGA), and upstream and downstream of the rRNA operon (upstream, 14-mer: AACTGWTACTGAGT; downstream, 9-mer: GTTCTAGTA). In order to explore their suitability for strain typing, these regions were amplified and sequenced from tissue samples of 10 additional Whipple's disease patients. The results are provided in Table 3, together with the results of 16S-23S rRNA intergenic spacer typing. Within this specimen/strain collection, there were 1 or 2 copies at the GroEL VNTR locus, 1–3 in the rRNA operon upstream VNTR locus, and 3–8 at the rRNA operon downstream VNTR locus. Each *T. whipplei* strain had a unique combination of rRNA spacer type and VNTR locus repeat numbers. Although not examined in detail, variation in the poly-G-tract was also observed; the biopsy of the index patient in this study had 11 Gs, and previously examined biopsies had 8 and 10 Gs, respectively [28].

Sequence repeats are prone to slipped-strand mispairing during DNA replication, which leads to frequent insertion and deletion of repeat units. This phenomenon leads to the emergence of strain-specific differences, and also provides a mechanism for genetic adaptation to different environments [11, 40, 41], such as might be required if *T. whipplei* transits from an external environmental niche to the human intestinal tract [27]. Repeat-sequence switching is a common mechanism of bacterial phase variation, and can affect genes relevant to antigenicity or pathogenicity [11]. VNTR-based typing has proven useful in the analysis of *Bacillus anthracis* population structure [20]. Multiple-locus VNTR analysis has high discriminatory power and has revealed distinct geographic clustering of *B. anthracis* strains, despite the relatively invariant nature of the *B. anthracis* genome. Our results indicate high discriminatory power of VNTR analysis for *T. whipplei*. Further studies will be needed to determine the usefulness of a VNTR typing system for *T. whipplei*; questions arise as to which of the VNTR loci are stable in distinct strains of *T. whipplei*, and which are subject to variation during the course of the disease or during propagation of the organism *in vitro*.

Conclusions

The sequences determined in this study, comprising a total of 13,193 bp, were extracted directly from a relevant naturally-infected site, using a single paraffin-embedded intestinal biopsy specimen in which *T. whipplei* was particularly abundant. This genomic information will soon be greatly extended with the completion and analysis of a full genome sequence (http://www.sanger.ac.uk/Projects/T_whippelii/) from a different strain. Among the findings from our data, all phylogenetic analyses using five non-rRNA data sets confirmed the position of *T. whipplei* within the phylum, *Actinobacteria*. The organization of the *T. whipplei atpD*, *groEL*, and *rpoB* chromosomal regions was consistent with other actinobacteria genomes. The discovery of VNTR loci may provide the basis for a new *T. whipplei* strain typing scheme with high discriminatory power. This method awaits further evaluation as a source of new insight into the pathogenesis, epidemiology and transmission of Whipple's disease.

Acknowledgements

This work was supported by a grant from the Deutsche Forschungsgemeinschaft (Ma 1663/3-1), a Dean's Fellowship from Stanford University to M.M., NIH Digestive Disease

Table 3. Results of rRNA intergenic spacer typing and VNTRs in the GroEL and rRNA operon upstream and downstream chromosomal regions for different *T. whipplei* strains.

Patient	Sample	rRNA spacer type	GroEL contig[c]	rRNA operon upstream[c]	rRNA operon downstream[c]
1[a]	IB[b]	2	2×	2×	6×
2	IB	1	1×	1×	4×
3	IB	1	1×	2×	4×
4	CSF	1	1×	3×	5×
5	IB	1	1×	3×	7×
6	IB	2	1×	2×	6×
7	CSF	2	1×	2×	7×
8	CSF	2	1×	2×	8×
9	CSF	2	1×	3×	5×
10	CSF	2	2×	2×	3×
11	IB	2	2×	2×	8×

[a] This is the index patient whose sample was used to assemble the chromosomal region sequences in this study.
[b] IB, intestinal biopsy.
[c] The number of repeats is given.

Research Center grant DK56339 (M.M., D.A.R.), and by grants from the Donald E. and Delia B. Baxter Foundation and the Wellcome Trust/Beowulf Genomics (ME019238) to D.A.R. We thank A. von Herbay (University of Heidelberg, Germany) for providing patient specimens and histopathological information, C. A. Cummings (Stanford University) for advice regarding sequence analysis, J. W. Brown (North Carolina State University) for determining the secondary structure of the RNase P RNA as well as providing valuable information regarding the RNase P RNA database, and L. Brocchieri (Stanford University) for advice regarding protein sequence alignments.

References

1. Brocchieri, L., Karlin, S.: A symmetric-iterated multiple alignment of protein sequences. J. Mol. Biol. 276, 249–264 (1998).
2. Brocchieri, L., Karlin, S.: Conservation among HSP60 sequences in relation to structure, function, and evolution. Protein Sci. 9, 476–486 (2000).
3. Brown, J. W.: The Ribonuclease P Database. Nucleic Acids Res. 27, 314 (1999).
4. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., 3rd, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Barrell, B. G., et al.: Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393, 537–544 (1998).
5. Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honore, N., Garnier, T., Churcher, C., Harris, D., Mungall, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R. M., Devlin, K., Duthoy, S., Feltwell, T., Fraser, A., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Lacroix, C., Maclean, J., Moule, S., Murphy, L., Oliver, K., Quail, M. A., Rajandream, M. A., Rutherford, K. M., Rutter, S., Seeger, K., Simon, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Taylor, K., Whitehead, S., Woodward, J. R., Barrell, B. G.: Massive gene decay in the leprosy bacillus. Nature 409, 1007–1011 (2001).
6. Drancourt, M., Carlioz, A., Raoult, D.: *rpoB* sequence analysis of cultured *Tropheryma whippelii*. J. Clin. Microbiol. 39, 2425–2430 (2001).
7. Felsenstein, J.: PHYLIP (Phylogeny Inference Package) version 3.5c. Department of Genetics, University of Washington, Seattle, Wash. 1993.
8. Geissdörfer, W., Wittmann, I., Rollinghoff, M., Schoerner, C., Bogdan, C.: Detection of a new 16S-23S rRNA spacer sequence variant (type 7) of *Tropheryma whippelii* in a patient with prosthetic aortic valve endocarditis. Eur. J. Clin. Microbiol. Infect. Dis. 20, 762–763 (2001).
9. Haas, E. S., Banta, A. B., Harris, J. K., Pace, N. R., Brown, J. W.: Structure and evolution of ribonuclease P RNA in Gram-positive bacteria. Nucleic Acids Res. 24, 4775–4782 (1996).
10. Hecker, M., Schumann, W., Völker, U.: Heat-shock and general stress response in *Bacillus subtilis*. Mol. Microbiol. 19, 417–428 (1996).
11. Henderson, I. R., Owen, P., Nataro, J. P.: Molecular switches – the ON and OFF of bacterial phase variation. Mol. Microbiol. 33, 919–932 (1999).
12. Henikoff, S., Henikoff, J. G.: Protein family classification based on searching a database of blocks. Genomics 19, 97–107 (1994).
13. Hensel, M., Lill, H., Schmid, R., Deckers-Hebestreit, G., Altendorf, K.: The ATP synthase (F1F0) of *Streptomyces lividans*: sequencing of the atp operon and phylogenetic considerations with subunit beta. Gene 152, 11–17 (1995).
14. Herrmann, B., Winqvist, O., Mattsson, J. G., Kirsebom, L. A.: Differentiation of *Chlamydia* spp. by sequence determination and restriction endonuclease cleavage of RNase P RNA genes. J. Clin. Microbiol. 34, 1897–1902 (1996).
15. Hinrikson, H. P., Dutly, F., Nair, S., Altwegg, M.: Detection of three different types of '*Tropheryma whippelii*' directly from clinical specimens by sequencing, single-strand conformation polymorphism (SSCP) analysis and type-specific PCR of their 16S-23S ribosomal intergenic spacer region. Int. J. Syst. Bacteriol. 49, 1701–1706 (1999).
16. Hoogvliet, G., van Wezel, G. P., Kraal, B.: Evidence that a single EF-Ts suffices for the recycling of multiple and divergent EF-Tu species in *Streptomyces coelicolor* A3(2) and *Streptomyces ramocissimus*. Microbiology 145, 2293–2301 (1999).
17. Hugenholtz, P., Goebel, B. M., Pace, N. R.: Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J. Bacteriol. 180, 4765–4774 (1998).
18. Hughes, A. L.: Contrasting evolutionary rates in the duplicate chaperonin genes of *Mycobacterium tuberculosis* and *M. leprae*. Mol. Biol. Evol. 10, 1343–1359 (1993).
19. Karlin, S., Mrazek, J., Campbell, A., Kaiser, D.: Characterizations of highly expressed genes of four fast-growing bacteria. J. Bacteriol. 183, 5025–5040 (2001).
20. Keim, P., Price, L. B., Klevytska, A. M., Smith, K. L., Schupp, J. M., Okinaka, R., Jackson, P. J., Hugh-Jones, M. E.: Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. J. Bacteriol. 182, 2928–2936 (2000).
21. Klenk, H. P., Zillig, W.: DNA-dependent RNA polymerase subunit B as a tool for phylogenetic reconstructions: branching topology of the archaeal domain. J. Mol. Evol. 38, 420–432 (1994).
22. La Scola, B., Fenollar, F., Fournier, P. E., Altwegg, M., Mallet, M. N., Raoult, D.: Description of *Tropheryma whipplei* gen. nov., sp. nov., the Whipple's disease bacillus. Int. J. Syst. Evol. Microbiol. 51, 1471–1479 (2001).
23. Ludwig, W., Neumaier, J., Klugbauer, N., Brockmann, E., Roller, C., Jilg, S., Reetz, K., Schachtner, I., Ludvigsen, A., Bachleitner, M., Fischer, U., Schleifer, K. H.: Phylogenetic relationships of Bacteria based on comparative sequence analysis of elongation factor Tu and ATP-synthase beta-subunit genes. Antonie Van Leeuwenhoek 64, 285–305 (1993).
24. Ludwig, W., Schleifer, K. H.: Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. FEMS Microbiol. Rev. 15, 155–173 (1994).
25. Ludwig, W., Strunk, O., Klugbauer, S., Klugbauer, N., Weizenegger, M., Neumaier, J., Bachleitner, M., Schleifer, K. H.: Bacterial phylogeny based on comparative sequence analysis. Electrophoresis 19, 554–568 (1998).
26. Maiwald, M., Ditton, H. J., von Herbay, A., Rainey, F. A., Stackebrandt, E.: Reassessment of the phylogenetic position of the bacterium associated with Whipple's disease and determination of the 16S-23S ribosomal intergenic spacer sequence. Int. J. Syst. Bacteriol. 46, 1078–1082 (1996).
27. Maiwald, M., Schuhmacher, F., Ditton, H. J., von Herbay, A.: Environmental occurrence of the Whipple's disease bacterium (*Tropheryma whippelii*). Appl. Environ. Microbiol. 64, 760–762 (1998).
28. Maiwald, M., von Herbay, A., Lepp, P. W., Relman, D. A.: Organization, structure, and variability of the rRNA operon of the Whipple's disease bacterium (*Tropheryma whippelii*). J. Bacteriol. 182, 3292–3297 (2000).

29. Morgenegg, S., Dutly, F., Altwegg, M.: Cloning and sequencing of a part of the heat shock protein 65 gene (hsp65) of "*Tropheryma whippelii*" and its use for detection of "*T. whippelii*" in clinical specimens by PCR. J. Clin. Microbiol. 38, 2248–2253 (2000).
30. Raoult, D., Birg, M. L., La Scola, B., Fournier, P. E., Enea, M., Lepidi, H., Roux, V., Piette, J. C., Vandenesch, F., Vital-Durand, D., Marrie, T. J.: Cultivation of the bacillus of Whipple's disease. N. Engl. J. Med. 342, 620–625 (2000).
31. Redenbach, M., Kieser, H. M., Denapaite, D., Eichner, A., Cullum, J., Kinashi, H., Hopwood, D. A.: A set of ordered cosmids and a detailed genetic and physical map for the 8 Mb *Streptomyces coelicolor* A3(2) chromosome. Mol. Microbiol. 21, 77–96 (1996).
32. Relman, D. A., Schmidt, T. M., Macdermott, R. P., Falkow, S.: Identification of the uncultured bacillus of Whipple's Disease. N. Engl. J. Med. 327, 293–301 (1992).
33. Renesto, P., Gautheret, D., Drancourt, M., Raoult, D.: Determination of the *rpoB* gene sequences of *Bartonella henselae* and *Bartonella quintana* for phylogenic analysis. Res. Microbiol. 151, 831–836 (2000).
34. Rinke de Wit, T. F., Bekelie, S., Osland, A., Miko, T. L., Hermans, P. W., van Soolingen, D., Drijfhout, J. W., Schoningh, R., Janson, A. A., Thole, J. E.: Mycobacteria contain two *groEL* genes: the second *Mycobacterium leprae* groEL gene is arranged in an operon with *groES*. Mol. Microbiol. 6, 1995–2007 (1992).
35. Saraste, M., Sibbald, P. R., Wittinghofer, A.: The P-loop – a common motif in ATP- and GTP-binding proteins. Trends Biochem. Sci. 15, 430–434 (1990).
36. Sarkar, G., Turner, R. T., Bolander, M. E.: Restriction-site PCR: a direct method of unknown sequence retrieval adjacent to a known locus by using universal primers. PCR Methods Appl. 2, 318–322 (1993).
37. Schmidt, T. M., DeLong, E. F., Pace, N. R.: Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J. Bacteriol. 173, 4371–4378 (1991).
38. Sela, S., Yogev, D., Razin, S., Bercovier, H.: Duplication of the *tuf* gene: a new insight into the phylogeny of eubacteria. J. Bacteriol. 171, 581–584 (1989).
39. Strunk, O., Gross, O., Reichel, B., May, M., Hermann, S., Stuckmann, N., Nonhoff, B., Ginhart, T., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., Schleifer, K.-H., Ludwig, W.: ARB: a software environment for sequence data. Department of Microbiology, Technische Universität München, Munich, Germany. 1999.
40. van Belkum, A., Scherer, S., van Alphen, L., Verbrugh, H.: Short-sequence DNA repeats in prokaryotic genomes. Microbiol. Mol. Biol. Rev. 62, 275–293 (1998).
41. van Belkum, A.: The role of short sequence repeats in epidemiologic typing. Curr. Opin. Microbiol. 2, 306–311 (1999).
42. van Wezel, G. P., Woudt, L. P., Vervenne, R., Verdurmen, M. L., Vijgenboom, E., Bosch, L.: Cloning and sequencing of the tuf genes of *Streptomyces coelicolor* A3(2). Biochim. Biophys. Acta 1219, 543–547 (1994).
43. Viale, A. M., Arakaki, A. K., Soncini, F. C., Ferreyra, R. G.: Evolutionary relationships among eubacterial groups as inferred from GroEL (chaperonin) sequence comparisons. Int. J. Syst. Bacteriol. 44, 527–533 (1994).
44. Vijgenboom, E., Woudt, L. P., Heinstra, P. W., Rietveld, K., van Haarlem, J., van Wezel, G. P., Shochat, S., Bosch, L.: Three *tuf*-like genes in the kirromycin producer *Streptomyces ramocissimus*. Microbiology 140, 983–998 (1994).
45. von Herbay, A., Ditton, H. J., Maiwald, M.: Diagnostic application of a polymerase chain reaction assay for the Whipple's disease bacterium to intestinal biopsies. Gastroenterology 110, 1735–1743 (1996).
46. von Herbay, A., Ditton, H. J., Schuhmacher, F., Maiwald, M.: Whipple's disease: Staging and monitoring by cytology and polymerase chain reaction analysis of cerebrospinal fluid. Gastroenterology 113, 434–441 (1997).
47. Walker, J. E., Saraste, M., Runswick, M. J., Gay, N. J.: Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J. 1, 945–951 (1982).
48. Wilson, K. H., Blitchington, R., Frothingham, R., Wilson, J. A. P.: Phylogeny of the Whipple's disease-associated bacterium. Lancet 338, 474–475 (1991).
49. Yoon, J. H., Park, Y. H.: Comparative sequence analyses of the ribonuclease P (RNase P) RNA genes from LL-2,6-diaminopimelic acid-containing actinomycetes. Int. J. Syst. Evol. Microbiol. 50, 2021–2029 (2000).

Corresponding author:
David A. Relman, VA Palo Alto Health Care System 154T, 3801 Miranda Avenue, Palo Alto, CA 94304, USA
Tel.: ++1 (650) 852-3308; Fax: ++1 (650) 852-3291;
e-mail: relman@stanford.edu