JOURNAL OF CLINICAL MICROBIOLOGY, Apr. 2002, p. 1303–1310 Vol. 40, No. 4
0095-1137/02/$04.00+0 DOI: 10.1128/JCM.40.4.1303–1310.2002

Genome Scale Comparison of *Mycobacterium avium* subsp. *paratuberculosis* with *Mycobacterium avium* subsp. *avium* Reveals Potential Diagnostic Sequences

John P. Bannantine,[1]* Emily Baechler,[2,3] Qing Zhang,[2] LingLing Li,[2] and Vivek Kapur[2]

National Animal Disease Center, Agricultural Research Service, U.S. Department of Agriculture, Ames, Iowa,[1] and Biomedical Genomics Center[2] and Department of Medicine,[3] University of Minnesota, Minneapolis, Minnesota

Received 16 November 2001/Returned for modification 6 January 2002/Accepted 27 January 2002

The genetic similarity between *Mycobacterium avium* subsp. *paratuberculosis* and other mycobacterial species has confounded the development of *M. avium* subsp. *paratuberculosis*-specific diagnostic reagents. Random shotgun sequencing of the *M. avium* subsp. *paratuberculosis* genome in our laboratories has shown >98% sequence identity with *Mycobacterium avium* subsp. *avium* in some regions. However, an in silico comparison of the largest annotated *M. avium* subsp. *paratuberculosis* contigs, totaling 2,658,271 bp, with the unfinished *M. avium* subsp. *avium* genome has revealed 27 predicted *M. avium* subsp. *paratuberculosis* coding sequences that do not align with *M. avium* subsp. *avium* sequences. BLASTP analysis of the 27 predicted coding sequences (genes) shows that 24 do not match sequences in public sequence databases, such as GenBank. These novel sequences were examined by PCR amplification with genomic DNA from eight mycobacterial species and ten independent isolates of *M. avium* subsp. *paratuberculosis*. From these analyses, 21 genes were found to be present in all *M. avium* subsp. *paratuberculosis* isolates and absent from all other mycobacterial species tested. One region of the *M. avium* subsp. *paratuberculosis* genome contains a cluster of eight genes, arranged in tandem, that is absent in other mycobacterial species. This region spans 4.4 kb and is separated from other predicted coding regions by 1,408 bp upstream and 1,092 bp downstream. The gene upstream of this eight-gene cluster has strong similarity to mycobacteriophage integrase sequences. The GC content of this 4.4-kb region is 66%, which is similar to the rest of the genome, indicating that this region was not horizontally acquired recently. Southern hybridization analysis confirmed that this gene cluster is present only in *M. avium* subsp. *paratuberculosis*. Collectively, these studies suggest that a genomics approach will help in identifying novel *M. avium* subsp. *paratuberculosis* genes as candidate diagnostic sequences.

Paratuberculosis, or Johne's disease, is a granulomatous enteritis of ruminant animals that may be prevalent in approximately 35% of United States dairy herds (7, 24). Diarrhea, reduced feed intake, weight loss, and eventual death characterize this intestinal disorder in cattle. Based upon prevalence figures and information from animal producers, economic losses for the dairy industry exceed 200 million dollars annually (18). *Mycobacterium avium* subsp. *paratuberculosis* is the etiologic agent of this economically significant disease. This veterinary pathogen has also been implicated as the etiologic agent of Crohn's disease (15), leading researchers to speculate on a potential pathogenic role for this organism in humans.

The control of Johne's disease is severely hampered by inadequate diagnostic tools (26). The prolonged incubation time and presence of subclinical cases permit infected animals to shed large amounts of bacilli in their feces before detection (21). Culture of *M. avium* subsp. *paratuberculosis* from feces has been the most reliable method for identifying infected animals; however, the slow growth of this organism results in a minimum of 6 weeks before culture data are available. Research on the pathogenesis and immunology of *M. avium* subsp. *paratuberculosis* infections of cattle will allow the design of better diagnostic and control procedures. New approaches that yield improved diagnostic tests will enable early detection and removal of subclinically infected animals. This will effectively reduce the incidence of Johne's disease in beef and dairy herds.

With the availability of over 60 published microbial genomes, some of which are in the same genus or even species, the age of comparative genomics has arrived. This approach is particularly useful in the genus *Mycobacterium* due to the number of sequenced species. *M. tuberculosis* and *M. leprae* genomes have been published (5, 6), and projects are under way for *M. bovis*, *M. avium* subsp. *avium*, *M. smegmatis,* and *M. avium* subsp. *paratuberculosis* (3). Comparative mycobacterial genomic approaches have been used to identify small-scale genomic deletions among *M. tuberculosis* isolates (12). Furthermore, large genome rearrangements (2) as well as deleted regions (14) were identified in studies comparing the *M. bovis* BCG vaccine strain with *M. tuberculosis*. Genome-wide comparisons in this genus will lead to an increased understanding of the genes required for pathogenicity as well as highlighting the sequences that make each species distinct.

Our laboratories have been actively engaged in sequencing the genome of *M. avium* subsp. *paratuberculosis* in order to reveal diagnostic sequences and/or antigens as well as to better understand the pathogenesis of Johne's disease. The strong nucleotide identity between *M. avium* subsp. *avium* and *M. avium* subsp. *paratuberculosis* (11, 22) has prevented the devel-

* Corresponding author. Mailing address: National Animal Disease Center, ARS-USDA, 2300 North Dayton Ave., Ames, IA 50010. Phone: (515) 663-7340. Fax: (515) 663-7458. E-mail: jbannant@nadc.ars.usda.gov.

TABLE 1. Mycobacterial strains used in this study

Isolate[a]	Source[b]	Origin	Additional information
M. avium subsp. *paratuberculosis*			
ATCC 19698	ATCC	Bovine	Isolated from ileum in 1974; type strain
1434	NADC	Ovine	
1045	NADC	Bovine	Isolated from a Holstein lymph node in 1984
1112	NADC	Bovine	Isolated from an Angus lymph node in 1984
1018	NADC	Bovine	Isolated from a Holstein lymph node in 1983
KAY	NADC	Bovine	Isolated from a Holstein ileum in 1993
K-10	NADC	Bovine	Isolated from a Wisconsin dairy herd in 1990
1010	NADC	Bovine	
1113	NADC	Bovine	
M. avium subsp. *avium*			
236	NADC	Bovine	
WP21 CP (9/5/01)	NADC	Avian	Mycobactin J independent, isolated from a wood pigeon
6004 CP (10/16/01)	NADC	Chicken	ATCC 35719; TMC 801
1015	UMN	Deer	
1161	UMN	Avian	
1282	UMN	Human	
1285	UMN	Human	
M. phlei	NADC		
M. smegmatis	NADC		
M. intracellulare	NADC	Porcine	TMC 1472, 35773; *M. avium-M. intracellulare-M. scrofulaceum* complex 6
M. fortuitum	NADC		
M. bovis			
BCG Pasteur (8/11/01)	ATCC		ATCC 35734; TMC 1011
95 1398 (1998–1999)	NADC	Deer	Isolated from a Colorado feedlot
M. tuberculosis TB 14323		Human	

[a] Dates of isolation (month/day/year) are in parentheses.

[b] ATCC, American Type Culture Collection; NADC, National Animal Disease Center; UMN, University of Minnesota.

opment of *M. avium* subsp. *paratuberculosis*-specific DNA sequences or antigens. To date, the only routinely used diagnostic sequence is that of the insertion element IS*900*, which is present in multiple copies in the *M. avium* subsp. *paratuberculosis* genome (9). In this study, we performed a partial genome comparison between the largest annotated contiguous DNA fragments (contigs) of *M. avium* subsp. *paratuberculosis* and the genetically similar *M. avium* subsp. *avium* genome. Sequences present in *M. avium* subsp. *paratuberculosis* but not *M. avium* subsp. *avium* were further analyzed by PCR with genomic DNA from several mycobacterial species. From these analyses, 21 unique *M. avium* subsp. *paratuberculosis* predicted coding sequences were identified. These unique sequences may be used to develop improved diagnostic reagents.

MATERIALS AND METHODS

Mycobacterial strains. Mycobacteria used in this study are listed in Table 1. All mycobacteria were cultured in Middlebrook 7H9 medium with 0.05% Tween 80 and oleic acid-albumin-dextrose-complex (Becton Dickinson Microbiology, Sparks, Md.). Cultures containing *M. avium* subsp. *paratuberculosis* isolates were supplemented with 2 mg of ferric mycobactin J (Allied Monitor Inc., Fayette, Mo.) per liter. All growth flasks were incubated at 37°C without shaking.

Annotation of *M. avium* subsp. *paratuberculosis* contigs greater than 10 kb. The sequencing and assembly strategies used here will be described elsewhere. For these studies, we chose assembled *M. avium* subsp. *paratuberculosis* contig fragments greater than 10 kb. Predicted coding sequences (genes) were identified with ARTEMIS software (http://www.sanger.ac.uk/Software/) and TB-parse, a program used to identify coding sequences in the *M. tuberculosis* genome (5). The results were compared and verified manually in ARTEMIS. A putative ribosome binding site was also evaluated for each coding sequence. The presence of an AG-rich sequence approximately 30 bp upstream of the start codon was scored as a putative ribosome binding site sequence. Similarities were identified by BLASTP analysis by using GenBank and a local database constructed by the Computational Biology Center at the University of Minnesota (http://www.cbc.umn.edu).

Sequence analysis. Sequence alignments of *M. avium* subsp. *paratuberculosis* and *M. avium* subsp. *avium* were compared and visualized with ACT software (http://www.sanger.ac.uk/Software/). *M. avium* subsp. *avium* is being sequenced by The Institute for Genomic Research (TIGR; http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi?organism=m_avium). Sequence alignments used to produce illustrations were made with AssemblyLIGN software (Accelrys, Princeton, N.J.).

DNA hybridization. Genomic DNA was extracted from several species of mycobacteria by a method modified from that described by Whipple et al. (25). One liter of Middlebrook 7H9-cultured mycobacteria was incubated at 37°C until an optical density at 540 between 0.50 and 0.56 was attained. D-Cycloserine was added to the medium at a final concentration of 0.5 mg/ml and incubated for an additional 24 h. Mycobacteria were harvested by centrifugation at 9,950 × *g* for 15 min, and the pellet was resuspended in 11 ml of Qiagen buffer B1 containing 1 mg of Qiagen RNase A per ml. Lipase was added (450,000 U; catalog no. L4384; Sigma, St. Louis, Mo.) to digest mycobacterial cell wall lipids. Following a 2-h incubation at 37°C, 20 mg of lysozyme was added, and incubation proceeded for an additional 3 h at 37°C. Qiagen proteinase K (500 μl; 20 mg/ml) was added and incubated for 1.5 h at 37°C. Qiagen buffer B2 (4 ml) was added, and the slurry was mixed and incubated 16 h at 50°C. The remaining cellular debris was removed by centrifugation at 12,100 × *g* for 20 min. The supernatant was poured over a preequilibrated Qiagen 500/G genomic tip. The loaded column was washed and processed according to the instructions of the manufacturer. *Pst*I-restricted DNA fragments were separated on a 1% agarose gel. DNA-containing gels were depurinated, denatured, and neutralized as described by

TABLE 2. PCR primers used in this study[a]

Gene	Primer 1	Primer 2
10	CGGCGGATCAGCATCTAC	CACCTCATCGTGGCCAGGTT
11	ACCGAACACGAGTGGAGCA	CAGACTCTGACCGACGTCAT
38	GCATTTCGGCTCCCACGGTG	TACGTCGGTTCGGCGCGCAT
48	CTGACACCGGCCTACGAACG	CATCCTCGCCGCGCCAGCAC
49	TGCTCAGGCCAACCCGCGC	TACTGGGCGGCGCACCCGAC
50	GGCATCCGCACCTTCGTCTG	CAATTCGTCGATCGGGCCGA
56	ATGAACACTTCTTCCTCTCTA	CATATCGCGGTGATCCTGAC
57	ATGGCCACCAACGACGACCA	CGCGGCCGTCGGGCCGGCTG
93	TTGCTGCGGGAAGGTTGCC	GAGAACGAGATGTGCGTCAG
134	GCGATGGTCAACGCCACCGG	TACAGCCCCGTGCAGACCGG
135	GCAGGCGTTTGCGTTCTTG	CGAGGTCCGAAATAGCGTAG
159	ATGCGTTTCGCCCTCCCGAC	TCACGCCTTGATTTCGTCCT
217	TGGCCGAACGCGGACTGTTC	TAGGAATCCGCGTCGACGAT
218	CAAGGTTCGTGACGGTATCG	TGACCCCAGCAGGTATGGC
219	CATCTACTGAGCGCCGTTTG	CACGCCGCCACCCCGTCCCG
228	GCAAGGTGGGCTTTGAAG	TGCGTGGGAGGATAAGGC
240	TTGGCACTGGCGTTTATG	ACATCGGGAACACAGGTCTC
241	ATCCTCCGGTTTGGCGGGAA	ACAGAGGTCGATCGGGTCG
250	CAGTCGGCCGGCGAAACGCC	CGCGGCGAAATCGAACGC
251	CACGTGCTGTCCCCATCGGC	CTACGTCTTCGTGACCAAAG
252	TGACCACCGACAACCCCACG	CATGAGGGCTGTCCCTCTCC
253	TTGACCGCGTTGACGGCGTT	CAGCGGTCCGCGCTCTTCGC
254	TGGGCAGCCCGGTGTCCCG	CACGCGCTCCTTTCAGCCTT
255	CAGTCACCCCGCGGCCGGTA	TCTACTGACCCGCAGATCGAA
256	TGGCCGTCAAGGACCAGAAC	CATGACCCTGCCGGCGTCCC
257	TGGCATTGGATCGCGTCGGA	TCAAACCCGGCGAGTTCTTC

[a] Primers are listed 5′ to 3′.

Sambrook et al. (19). DNA was transferred by capillary action (20) to BrightStar-Plus membranes (Ambion, Austin, Tex.), and probes were labeled with [α-^{32}P]dCTP (ICN, Costa Mesa, Calif.) by random priming. Hybridization was performed in an Autoblot hybridization oven (Bellco Biotechnology, Vineland, N.J.) at 45°C for 16 h in ExpressHyb hybridization solution (Clontech, Palo Alto, Calif.). Probed blots were washed sequentially with increasing stringency solutions as described previously (20). Detection was by autoradiography using BioMax MR film (Kodak, Rochester, N.Y.).

PCR amplification. Primers listed in Table 2 were designed from *M. avium* subsp. *paratuberculosis*-specific sequences to amplify mycobacterial genomic DNA. Amplification recipes containing nucleotides, buffer, primers, template, and DNA polymerase were standard except for the addition of 5% dimethyl sulfoxide (Sigma). Amplification conditions included a 5-min denaturation step at 94°C and 35 cycles of 45 s at 94°C, 1 min at 55°C, and 2 min at 72°C. High-fidelity *Pwo* polymerase (Boehringer Ingelheim Pharmaceutical Inc., Ridgefield, Conn.) was used in amplifications to generate probes used in Southern hybridization experiments. All other amplifications used *Taq* DNA polymerase (Roche Molecular Biochemicals, Indianapolis, Ind.). Primers used to amplify the no. 7 sequence for a probe in Southern hybridizations were 5′-ATCAGGC TGACGGGATTGCCC-3′ and 5′-TCAACGAGTGCACGGGAACC-3′.

Nucleotide sequence accession numbers. The nucleotide sequences of all *M. avium* subsp. *paratuberculosis* genes described in this study were deposited in the GenBank/EMBL nucleotide sequence data library under accession numbers AF445420 through AF445446.

RESULTS

Twenty-seven *M. avium* subsp. *paratuberculosis* predicted coding sequences are not present in *M. avium* subsp. *avium*. Our laboratories are sequencing the complete genome of *M. avium* subsp. *paratuberculosis* K-10, a field isolate recovered from a cow with clinical Johne's disease (http://www.cbc.umn.edu/ResearchProjects/AGAC/Mptb/Mptbhome.html). The genome size is estimated to be >5 Mb based on assembled sequence data, and at the time of this analysis (July 2001), 2.65 Mb was contained in contig fragments greater than 10 kb. Contigs that are above 10 kb were annotated with ARTEMIS and represent 48% of the total genome. The average size of the annotated contigs is 25 kb, with one contig over 70 kb. Each gene within the annotated contig set was also checked manually and confirmed by TB-parse. These contigs were aligned with *M. avium* subsp. *avium* sequence data generated at TIGR. TIGR has 612 contigs that total 5,867,714 bp in the 8 July 2001 data set.

M. avium subsp. *avium* and *M. avium* subsp. *paratuberculosis* display a high degree of similarity at the nucleotide level as well as local gene order conservation. An analysis of an 11-kb region surrounding the origin of replication for each of these genomes shows 98% nucleotide identity (Q. Zhang, E. Baechler, L. Li, J. P. Bannantine, and V. Kapur, unpublished data). The sequence similarity between orthologs in *M. avium* subsp. *paratuberculosis* and *M. avium* subsp. *avium* was greater than that between *M. avium* subsp. *paratuberculosis* and other mycobacterial species. A more global comparison shows that these strong nucleotide identities are present throughout both genomes. Despite this strong genetic similarity, a total of 27 genes from the annotated *M. avium* subsp. *paratuberculosis* contigs were identified that did not align with the unfinished *M. avium* subsp. *avium* genome by computerized alignments. These unique *M. avium* subsp. *paratuberculosis* sequences are listed in Table 3 along with some sequence characteristics. Of these, three contained weak similarity to proteins in other mycobacterial species or proteins in GenBank (Table 3). This

TABLE 3. *M. avium* subsp. *paratuberculosis* predicted coding sequences not present in *M. avium* subsp. *avium*, as determined in silico

Contig	Gene	Identification in TB-parse[a]	Presence of RBS[b]	No. of amino acids[c]	Best BLAST hit (expect value)[d]
1427	10	Yes	No	322	No similarity
1427	11	Yes	Yes	191	0.99
1467	38	Yes	Yes	173	4.4
1482	48	Yes	Yes	352	3.8
1482	49	Yes	Yes	267	0.95
1482	50	Yes	Yes	293	5.1
1490	56	Yes	Yes	193	No similarity
1490	57	Yes	Yes	103	4.7
1532	93	No	Yes	183	0.009
1553	128	Yes	No	191	**$7e^{-07}$** (hyp. protein)
1556	134	Yes	Yes	143	2.8
1556	135	Yes	Yes	157	0.13
1578	159	Yes	Yes	185	**$2e^{-04}$**
1602	217	Yes	Yes	106	5.1
1602	218	Yes	Yes	835	0.83
1602	219	Yes	Yes	87	1.1
1605	228	No	No	369	0.86
1612	240	Yes	Yes	223	2.8
1612	241	Yes	Yes	123	**$1e^{-05}$**
1614	250	Yes	Yes	199	0.030
1614	251	Yes	No	179	No similarity
1614	252	Yes	Yes	96	No similarity
1614	253	Yes	Yes	74	4.7
1614	254	Yes	Yes	146	9.7
1614	255	Yes	Yes	241	3.6
1614	256	Yes	Yes	141	2.1
1614	257	Yes	Yes	87	0.61

[a] TB-parse is the gene prediction program used to annotate the *M. tuberculosis* genome (6). "Yes," TB-parse identified the listed *M. paratuberculosis* gene; "No," TB-parse did not recognize it as a coding sequence.

[b] Presence or absence of a consensus ribosome binding site (RBS).

[c] Number of amino acids encoded by the predicted coding sequence.

[d] Expect value of the best match in GenBank. The three predicted coding sequences that contain the highest similarity to a sequence in Genbank are in bold. hyp. protein, hypothetical protein in GenBank.

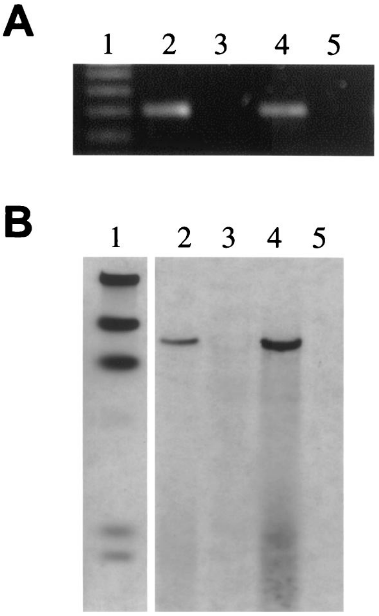

FIG. 1. PCR and Southern hybridization analysis of a unique DNA fragment show that it is conserved among two *M. avium* subsp. *paratuberculosis* isolates tested but not present in *M. avium* subsp. *avium* genomic DNA. (A) PCR amplification of specific products representing unique fragment no. 7 from genomic DNA of two representative strains of *M. avium* subsp. *paratuberculosis* and two representative isolates of *M. avium* subsp. *avium*. Lanes: 1, 100-bp DNA size standards; 2, *M. avium* subsp. *paratuberculosis* K-10; 3, *M. avium* subsp. *avium* TMC801; 4, *M. avium* subsp. *paratuberculosis* ATCC 19698; 5, *M. avium* subsp. *avium* deer isolate. (B) Southern hybridization of *Pst*I-restricted genomic DNA (2 μg each) from these same isolates. A DNA fragment amplified with primers designed from the no. 7 sequence was labeled and used as the probe. The data show the presence of this fragment in *M. avium* subsp. *paratuberculosis* but not *M. avium* subsp. *avium*. The same blot, when stripped and reprobed with the gene encoding the 65-kDa heat shock protein, revealed bands in all four lanes with equal intensities, ruling out the possibility of false-negative hybridizations for the lanes containing *M. avium* subsp. *avium* genomic DNA (not shown). The lanes in panel B are identical to those in panel A except that lane 1 contains λ-*Hin*dIII size standards.

leaves 24 genes with no significant similarity to any known proteins. Since only approximately half of the *M. avium* subsp. *paratuberculosis* genome was used in these analyses, a complete genome analysis may reveal an estimated 50 unique *M. avium* subsp. *paratuberculosis* genes.

Some *M. avium* subsp. *paratuberculosis* sequences that did not align with *M. avium* subsp. *avium,* either in silico or experimentally, contain similarity to other mycobacterial species. One such sequence, designated no. 7, was tested by PCR and Southern hybridization with two *M. avium* subsp. *avium* isolates and two *M. avium* subsp. *paratuberculosis* strains (Fig. 1). An amplified PCR fragment was produced only with *M. avium* subsp. *paratuberculosis* genomic DNA as the template (Fig. 1A). Likewise, DNA hybridization on Southern blots detected only *M. avium* subsp. *paratuberculosis* sequences, not *M. avium* subsp. *avium* (Fig. 1B). However, BLASTP analysis of the no. 7 sequence revealed strong similarity to hypothetical proteins in the *M. tuberculosis* genome. Therefore, caution must be used in determining whether a sequence is truly unique to *M. avium* subsp. *paratuberculosis*. More comprehensive experiments using additional mycobacterial species are necessary before such conclusions can be made.

PCR analysis. PCR amplification was performed on several mycobacterial species, strains, and isolates to experimentally determine the specificity for 26 of the 27 sequences (Table 4). Gene 128 was not included in these analyses because it had the lowest expect value (highest similarity to a sequence in GenBank) of the 27 sequences by BLASTP analysis (Table 3). These data show that primers designed from all 26 *M. avium* subsp. *paratuberculosis* K-10 genes could produce an amplified product in all 10 *M. avium* subsp. *paratuberculosis* strains or isolates tested. In addition, despite an absence of any homologous sequences in public databases, PCR products of the correct size were obtained for five genes by using templates from other mycobacterial species. Following this analysis, a core group of 21 genes that are present only in *M. avium* subsp. *paratuberculosis* remained (Table 4).

Sequence analysis of an *M. avium* subsp. *paratuberculosis*-specific eight-gene cluster. Table 3 lists eight genes present on contig fragment 1614. These eight genes are arranged in tandem, span a total of 4.4 kb at the end of the 1614 contig (Fig. 2), and are present only in *M. avium* subsp. *paratuberculosis* (Table 4). Located 1,408 bp upstream of gene 250 is an integrase gene that contains similarity to other mycobacteriophage integrases. As larger contiguous fragments were assembled from the gap closure phase of the *M. avium* subsp. *paratuberculosis* genome project, a search to define the ends of the 4.4-kb sequence not present in *M. avium* subsp. *avium* was performed. This 4.4-kb segment containing genes 250 to 257, herein termed no. 481, is located at the end of the 46-kb contig 1614 and it was found to align with the 94-kb contig 1398 present in a more recent contig assembly data set (Fig. 2). The no. 481 sequence aligned near the center of the 94-kb contig essentially at 35 to 45 kb. A trimmed portion of the 1398 contig is shown in the alignment in Fig 2. The results of this analysis further extended the region of no. 481 sequence to 9.4 kb, none of which aligns with the *M. avium* subsp. *avium* sequence in silico.

A TBLASTX analysis was performed on the 9.4-kb sequence (designated contig 1398-trimmed in Fig. 2). The results of these analyses revealed that, while no sequences aligned with *M. avium* subsp. *avium*, the ends of contig 1398-trimmed align with sequences in *M. tuberculosis* (Table 5). The open reading frames designated by a question mark in Table 5 are present on contig 1398, which has not yet been annotated. This again leaves a core sequence of eight open reading frames, comprising the no. 481 sequence, that are present only in *M. avium* subsp. *paratuberculosis*. This core sequence is flanked by 1,408 bp of noncoding sequence downstream and 1092-bp of noncoding sequence upstream (Fig. 2). Therefore, this novel core sequence is well separated from other predicted open reading frames.

Southern hybridization analysis shows that the no. 481 sequence is specific to *M. avium* subsp. *paratuberculosis*. To confirm experimentally that no. 481 is present only in *M. avium* subsp. *paratuberculosis*, three arbitrarily chosen genes of the no. 481 sequence (251, 253, and 255) were radiolabeled and used as probes in DNA hybridization with several mycobacterial species, including *M. fortuitum*, *M. bovis*, *M. intracellulare*, *M. avium* subsp. *avium*, and *M. avium* subsp. *paratuberculosis*

TABLE 4. PCR analysis of *M. paratuberculosis* predicted coding sequences

Strain	Amplification of gene[a]														
	56	57	159	217	218	228	240	250	251	252	253	254	255	256	257
M. avium subsp. *paratuberculosis*															
ATCC 19698	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1434	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1045	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1112	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1018	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Kay	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
K-10	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1010	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
1113	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
M. avium subsp. *avium*															
236	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
WP21	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
TMC 801	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
1015	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
1161	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
1282	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
1285	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
M. phlei	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
M. smegmatis	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
M. intracellulare	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
M. fortuitum	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
M. bovis															
BCG	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
95–1398	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
M. tuberculosis	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−

Strain	10	11	38	48	49	50	93	134	135	219	241
M. avium subsp. *paratuberculosis*											
ATCC 19698	+	+	+	+	+	+	+	+	+	+	+
1434	+	+	+	+	+	+	+	+	+	+	+
1045	+	+	+	+	+	+	+	+	+	+	+
1112	+	+	+	+	+	+	+	+	+	+	+
1018	+	+	+	+	+	+	+	+	+	+	+
Kay	+	+	+	+	+	+	+	+	+	+	+
K-10	+	+	+	+	+	+	+	+	+	+	+
1010	+	+	+	+	+	+	+	+	+	+	+
1113	+	+	+	+	+	+	+	+	+	+	+
M. avium subsp. *avium*											
236	−	−	−	−	−	−	+	+	−	−	−
WP21	−	−	−	+	+	+	+	+	−	−	−
TMC 801	−	−	−	+	+	+	+	+	−	−	−
1015	−	−	−	+	+	+	+	+	−	−	−
1161	−	−	−	+	+	+	+	+	−	−	−
1282	−	−	−	−	−	−	+	+	−	−	−
1285	−	−	−	−	−	−	+	+	−	−	−
M. phlei	−	−	−	−	−	−	−	+	−	−	−
M. smegmatis	−	−	−	−	−	−	−	−	−	−	−
M. intracellulare	−	−	−	+	+	+	+	−	−	−	−
M. fortuitum	−	−	−	−	−	−	−	−	−	−	−
M. bovis											
BCG	−	−	−	−	+	−	−	−	−	−	−
95–1398	−	−	−	−	+	−	−	−	−	−	−
M. tuberculosis	−	−	−	−	+	−	−	−	−	−	−

[a] +, an amplification product of the correct size was detected by ethidium bromide staining; −, no amplification product was detected.

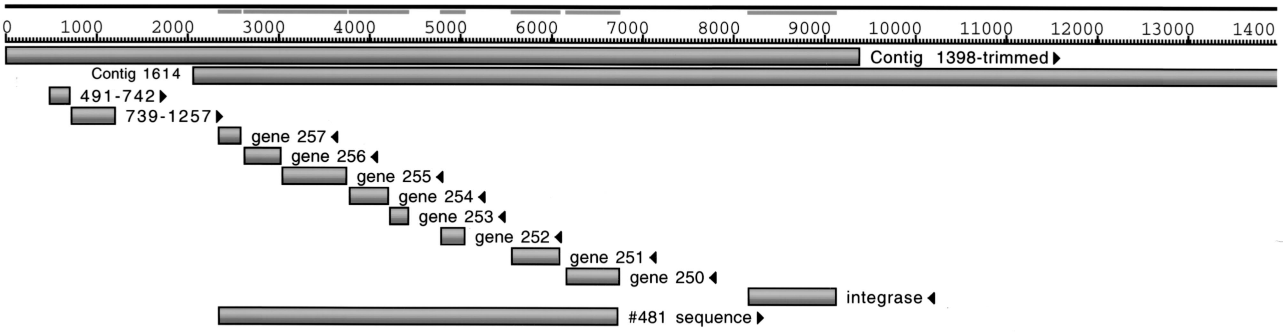

FIG. 2. Sequence alignment schematic showing positions of predicted coding sequences relative to assembled contig fragments. Alignments of contig 1614 and a trimmed fragment of the 94-kb contig 1398 are shown along with each predicted coding sequence listed in Table 5. Coding sequences labeled by start and stop coordinates are present on a contig that has not yet been annotated. Note that the core region of genes 250 to 257 is well separated from neighboring coding regions. The integrase gene upstream of gene 250 is also designated gene 249.

(Fig. 3). Only an *M. avium* subsp. *paratuberculosis* fragment greater than 9.5 kb was detected by each of the three gene probes.

DISCUSSION

A major research effort in the study of *M. avium* subsp. *paratuberculosis* has been directed at unraveling the complexities surrounding diagnosis of infected animals. However, no DNA sequence besides the IS*900* element has been routinely used to detect the presence of *M. avium* subsp. *paratuberculosis* (9, 16, 17). IS*900* is a repeated sequence in *M. avium* subsp. *paratuberculosis*, present in 14 copies in strain K10 (V. Kapur, L. Li, Q. Zhang, and J. P. Bannantine, unpublished data). The results of this study reveal an initial list of 27 sequences that are likely specific to *M. avium* subsp. *paratuberculosis*, as determined by an in silico comparison with *M. avium* subsp. *avium*. Subsequent analysis by PCR amplification has trimmed this list down to 21 *M. avium* subsp. *paratuberculosis*-specific sequences. Nearly one half of the genome was analyzed; therefore, the list reported here will likely expand when the genome sequence is completed. These novel sequences provide investigators with a list of potential diagnostic candidate sequences that can be applied in a multiplex PCR format to better diagnose Johne's disease in cattle.

A surprising finding revealed by this comparative genomic approach was the presence of *M. avium* subsp. *paratuberculosis* sequences that contain similarity to *M. tuberculosis* but are absent in *M. avium* subsp. *avium*. This was observed for portions of contig 1398-trimmed and the no. 7 sequence, which is not listed in Table 3. Because *M. avium* subsp. *avium* is most closely related to *M. avium* subsp. *paratuberculosis*, it is the genome of choice for initial screening of novel *M. avium* subsp. *paratuberculosis* sequences. However, each sequence must be subsequently evaluated for specificity experimentally with a complete panel of *Mycobacterium* sp. DNA before specificity is concluded.

One of the annotated contigs (1614) contained 8 of the 27 *M. avium* subsp. *paratuberculosis* sequences not present in *M. avium* subsp. *avium*. This region of eight genes (no. 481 sequence) was examined further, as it seems likely to be cotranscribed from one or a small number of promoters, although this was not experimentally shown. The function of this novel gene cluster is not known, although it is possible that the no. 481 sequence may represent a cryptic prophage or prophage remnant. The presence of an upstream coding sequence with strong identity to mycobacteriophage integrases supports this hypothesis. The discovery of a novel mycobacteriophage that is selectively present in certain mycobacterial species is not unprecedented. The prophage phiRv1 is present in *M. tuberculosis* and some strains of *M. bovis* but is missing from all *M. bovis* BCG genomes (1, 2, 14). Conversely, other mycobacteriophages do not have the genomic structure reported for the no. 481 sequence. For example, the genomes of D29 and L5 mycobacteriophages are much larger at 50 kb, with the integrase present near the middle of the bacteriophage genome next to an *att*P attachment site (8). Furthermore, the integrase gene is separated from the rest of the no. 481 sequence by 1.4 kb, and further upstream of the putative integrase gene on the 1614 contig, another 2.3 kb of sequence separates the next predicted coding sequence. This situation is different from the high coding density seen in bacteriophages (8). Finally, sequencing of the *M. tuberculosis* genome shows that there are several segments containing phage-related genes (3). One of these appears to be small and contains part of a phage-like integrase gene and a putative excisionase gene but no other functions that are obviously phage related (5).

Horizontally transferred DNA segments that may corre-

TABLE 5. Tera-BLASTX data of nucleic acid database using predicted coding sequences in contig 1398-trimmed

ORF[a]	Position[b]	No. of amino acids	Best BLAST hit	e value
?	491–742	83	*M. tuberculosis* Rv2517c	2e-20
?	739–1257	172	*M. tuberculosis* Rv2516c	9e-55
257	2349–2627 c	92	No similarity	
256	2624–3049 c	141	No similarity	
255	3051–3776 c	241	No similarity	
254	3782–4222 c	146	No similarity	
253	4222–4446 c	74	No similarity	
252	4789–5079 c	96	No similarity	
251	5570–6109 c	179	No similarity	
250	6164–6763 c	199	No similarity	
249	8171–9151 c	326	Integrase (*M. tuberculosis* CDC1551)	9e-77

[a] ORF, open reading frame. ?, ORF has not yet been annotated.
[b] c, coding sequence is on the complementary DNA strand.

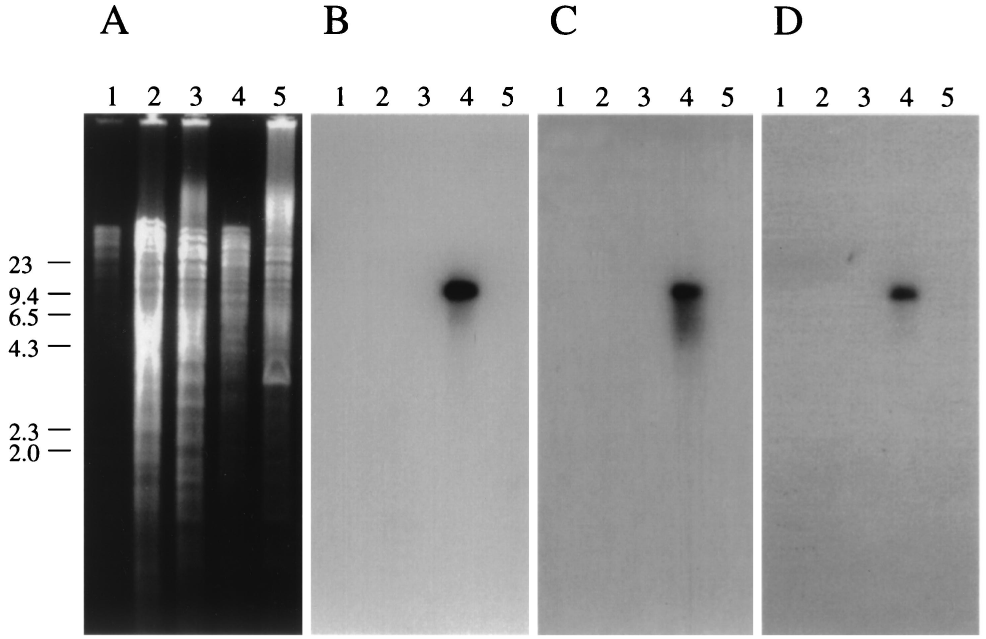

FIG. 3. DNA hybridization of various mycobacterial species shows that genes 251, 253, and 255 are present only in *M. avium* subsp. *paratuberculosis*. The 1% agarose gel in panel A containing *Pst*I-restricted DNA fragments was transferred to a nylon membrane and sequentially probed with genes 255 (B), 253 (C), and 251 (D). The nylon filter was stripped between hybridizations. An *M. avium* subsp. *paratuberculosis* DNA fragment greater than 9.4 kb was detected in each hybridization. λ-*Hin*dIII size standards are indicated on the left. Lanes: 1, *M. fortuitum*; 2, *M. bovis* BCG; 3, *M. intracellulare*; 4, *M. avium* subsp. *paratuberculosis* K-10; 5, *M. avium* subsp. *avium* TMC801.

spond to pathogenicity islands can often be identified by differences in their GC content (10, 13). The 9.0-kb GS element in *M. avium* subsp. *paratuberculosis* (4, 23), for example, has an average GC content of 57.1% (4), significantly lower than the 69.31% average for the *M. avium* subsp. *paratuberculosis* genome. Although the source of this low-GC island was never identified, the element is bounded by short inverted repeats, further suggesting its acquisition by horizontal transfer. The fact that the no. 481 sequence is adjacent to a putative integrase may suggest that the no. 481 sequence is part of a horizontally acquired element. However, the GC content of the no. 481 region (66%) is similar to that of the rest of the genome (69.31%), which argues against a recent horizontally transferred element from a species with different GC content. In *M. tuberculosis,* the average GC content is 65.6%, although some areas show dramatic differences in GC content. Regions that were unusually GC rich or poor were found to correspond to the novel PE-PGRS gene family (5) or to genes encoding polyketide synthases or transmembrane proteins (5).

One of the primary goals set when our laboratories undertook the sequencing of the *M. avium* subsp. *paratuberculosis* genome was to identify novel sequences with potential diagnostic utility. This communication represents our initial efforts to achieve this goal. Heterologous expression of these genes is in progress. The resulting purified proteins will then be used in studies to determine if they are recognized by sera from cattle with Johne's disease. These findings may have significant application in the development of new diagnostic tests to identify cattle infected with Johne's disease.

ACKNOWLEDGMENTS

We thank individuals at TIGR for their work on sequencing the *M. avium* subsp. *avium* genome. We thank Chad Reinke and Janis Hansen for technical assistance.

This work was funded by USDA-NRI grant 00-02215 to V.K. and J.P.B. Portions of this work were also supported by the Agricultural Research Service.

REFERENCES

1. **Behr, M. A., M. A. Wilson, W. P. Gill, H. Salamon, G. K. Schoolnik, S. Rane, and P. M. Small.** 1999. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science **284:**1520–1523.
2. **Brosch, R., S. V. Gordon, C. Buchrieser, A. S. Pym, T. Garnier, and S. T. Cole.** 2000. Comparative genomics uncovers large tandem chromosomal duplications in *Mycobacterium bovis* BCG Pasteur. Yeast **17:**111–123.
3. **Brosch, R., A. S. Pym, S. V. Gordon, and S. T. Cole.** 2001. The evolution of mycobacterial pathogenicity: clues from comparative genomics. Trends Microbiol. **9:**452–458.
4. **Bull, T. J., J. M. Sheridan, H. Martin, N. Sumar, M. Tizard, and J. Hermon-Taylor.** 2000. Further studies on the GS element. A novel mycobacterial insertion sequence (IS1612), inserted into an acetylase gene (mpa) in *Mycobacterium avium* subsp. *silvaticum* but not in *Mycobacterium avium* subsp. *paratuberculosis*. Vet. Microbiol. **77:**453–463.
5. **Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry III, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, B. G. Barrell, et al.** 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature **393:**537–544.
6. **Cole, S. T., K. Eiglmeier, J. Parkhill, K. D. James, N. R. Thomson, P. R. Wheeler, N. Honore, T. Garnier, C. Churcher, D. Harris, K. Mungall, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. M. Davies, K. Devlin, S. Duthoy, T. Feltwell, A. Fraser, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, C. Lacroix, J. Maclean, S. Moule, L. Murphy, K. Oliver, M. A. Quail, M. A. Rajandream, K. M. Rutherford, S. Rutter, K. Seeger, S. Simon, M. Simmonds, J. Skelton, R. Squares, S. Squares, K. Stevens, K. Taylor, S. Whitehead, J. R. Woodward, and B. G. Barrell.** 2001. Massive gene decay in the leprosy bacillus. Nature **409:**1007–1011.
7. **Collins, M. T., D. C. Sockett, W. J. Goodger, T. A. Conrad, C. B. Thomas,**

and D. J. Carr. 1994. Herd prevalence and geographic distribution of, and risk factors for, bovine paratuberculosis in Wisconsin. J. Am. Vet. Med. Assoc. **204:**636–641.

8. **Ford, M. E., G. J. Sarkis, A. E. Belanger, R. W. Hendrix, and G. F. Hatfull.** 1998. Genome structure of mycobacteriophage D29: implications for phage evolution. J. Mol. Biol. **279:**143–164.
9. **Green, E. P., M. L. Tizard, M. T. Moss, J. Thompson, D. J. Winterbourne, J. J. McFadden, and J. Hermon-Taylor.** 1989. Sequence and characteristics of IS900, an insertion element identified in a human Crohn's disease isolate of *Mycobacterium paratuberculosis*. Nucleic Acids Res. **17:**9063–9073.
10. **Hurtado, A., and F. Rodriguez-Valera.** 1999. Accessory DNA in the genomes of representatives of the *Escherichia coli* reference collection. J. Bacteriol. **181:**2548–2554.
11. **Ji, Y. E., M. J. Colston, and R. A. Cox.** 1994. Nucleotide sequence and secondary structures of precursor 16S rRNA of slow-growing mycobacteria. Microbiology **140:**123–132.
12. **Kato-Maeda, M., J. T. Rhee, T. R. Gingeras, H. Salamon, J. Drenkow, N. Smittipat, and P. M. Small.** 2001. Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Res. **11:**547–554.
13. **Lio, P., and M. Vannucci.** 2000. Finding pathogenicity islands and gene transfer events in genome data. Bioinformatics **16:**932–940.
14. **Mahairas, G. G., P. J. Sabo, M. J. Hickey, D. C. Singh, and C. K. Stover.** 1996. Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis*. J. Bacteriol. **178:**1274–1282.
15. **McFadden, J. J., P. D. Butcher, R. Chiodini, and J. Hermon-Taylor.** 1987. Crohn's disease-isolated mycobacteria are identical to *Mycobacterium paratuberculosis*, as determined by DNA probes that distinguish between mycobacterial species. J. Clin. Microbiol. **25:**796–801.
16. **Millar, D., J. Ford, J. Sanderson, S. Withey, M. Tizard, T. Doran, and J. Hermon-Taylor.** 1996. IS900 PCR to detect *Mycobacterium paratuberculosis* in retail supplies of whole pasteurized cows' milk in England and Wales. Appl. Environ. Microbiol. **62:**3446–3452.
17. **Moss, M. T., E. P. Green, M. L. Tizard, Z. P. Malik, and J. Hermon-Taylor.** 1991. Specific detection of *Mycobacterium paratuberculosis* by DNA hybridisation with a fragment of the insertion element IS900. Gut **32:**395–398.
18. **Ott, S. L., S. J. Wells, and B. A. Wagner.** 1999. Herd-level economic losses associated with Johne's disease on US dairy operations. Prev. Vet. Med. **40:**179–192.
19. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
20. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98:**503–517.
21. **Stabel, J. R.** 1998. Johne's disease: a hidden threat. J. Dairy Sci. **81:**283–288.
22. **Stahl, D. A., and J. W. Urbance.** 1990. The division between fast- and slow-growing species corresponds to natural relationships among the mycobacteria. J. Bacteriol. **172:**116–124.
23. **Tizard, M., T. Bull, D. Millar, T. Doran, H. Martin, N. Sumar, J. Ford, and J. Hermon-Taylor.** 1998. A low G+C content genetic island in *Mycobacterium avium* subsp. *paratuberculosis* and *M. avium* subsp. *silvaticum* with homologous genes in *Mycobacterium tuberculosis*. Microbiology **144:**3413–3423.
24. **Wells, S. J., S. L. Ott, and A. H. Seitzinger.** 1998. Key health issues for dairy cattle–-new and old. J. Dairy Sci. **81:**3029–3035.
25. **Whipple, D. L., R. B. Le Febvre, R. E. Andrews, Jr., and A. B. Thiermann.** 1987. Isolation and analysis of restriction endonuclease digestive patterns of chromosomal DNA from *Mycobacterium paratuberculosis* and other *Mycobacterium* species. J. Clin. Microbiol. **25:**1511–1515.
26. **Whitlock, R. H., and C. Buergelt.** 1996. Preclinical and clinical manifestations of paratuberculosis (including pathology). Vet. Clin. N. Am. Food Anim. Pract. **12:**345–356.