The complete genome sequence of *Mycobacterium avium* subspecies *paratuberculosis*

Lingling Li*†‡, John P. Bannantine‡§, Qing Zhang*†‡, Alongkorn Amonsin*†‡, Barbara J. May*†, David Alt§, Nilanjana Banerji†¶, Sagarika Kanjilal†‡¶, and Vivek Kapur*†‡‖

*Department of Microbiology, †Biomedical Genomics Center, and ¶Department of Medicine, University of Minnesota, St. Paul, MN 55108; and §National Animal Disease Center, U.S. Department of Agriculture–Agriculture Research Service, Ames, IA 50010

Communicated by Harley W. Moon, Iowa State University, Ames, IA, July 13, 2005 (received for review March 18, 2005)

We describe here the complete genome sequence of a common clone of *Mycobacterium avium* subspecies *paratuberculosis* (*Map*) strain K-10, the causative agent of Johne's disease in cattle and other ruminants. The K-10 genome is a single circular chromosome of 4,829,781 base pairs and encodes 4,350 predicted ORFs, 45 tRNAs, and one rRNA operon. *In silico* analysis identified >3,000 genes with homologs to the human pathogen, *M. tuberculosis* (*Mtb*), and 161 unique genomic regions that encode 39 previously unknown *Map* genes. Analysis of nucleotide substitution rates with *Mtb* homologs suggest overall strong selection for a vast majority of these shared mycobacterial genes, with only 68 ORFs with a synonymous to nonsynonymous substitution ratio of >2. Comparative sequence analysis reveals several noteworthy features of the K-10 genome including: a relative paucity of the PE/PPE family of sequences that are implicated as virulence factors and known to be immunostimulatory during *Mtb* infection; truncation in the EntE domain of a salicyl-AMP ligase (MbtA), the first gene in the mycobactin biosynthesis gene cluster, providing a possible explanation for mycobactin dependence of *Map*; and *Map*-specific sequences that are likely to serve as potential targets for sensitive and specific molecular and immunologic diagnostic tests. Taken together, the availability of the complete genome sequence offers a foundation for the study of the genetic basis for virulence and physiology in *Map* and enables the development of new generations of diagnostic tests for bovine Johne's disease.

genomics | Johne's disease | molecular diagnostics

Mycobacterium avium subspecies *paratuberculosis* (*Map*) is an extremely slow-growing, acid-fast, mycobactin-dependent multispecies pathogen. Infection with this bacterium leads to a chronic granulomatous enteritis in cattle and other wild and domestic ruminants, termed Johne's disease (1). Clinical signs of Johne's disease include diarrhea, weight loss, decreased milk production, and mortality. A recent study estimated that 21% of United States dairy herds are infected, resulting in considerable economic losses to the dairy industry totaling more than $200 million per annum (2, 3). A major concern with Johne's disease is the ease with which the bacterium is spread. Subclinically or clinically infected animals shed *Map* in feces and milk, enabling dissemination to susceptible calves, the environment, and in retail milk (4–6). *Map*-containing milk may be of particular concern because the bacterium has been suggested as a possible cause of Crohn's disease in humans (7).

The detection and diagnosis of *Map*-infected animals poses great difficulties. *Map* culture can require up to 16 weeks or more, and serological tests lack sensitivity because of the seroconversion occurring relatively late during the course of the disease (8). Also, previously developed PCR-based approaches (e.g., IS900) have been shown to lack specificity (9). This result is, in part, due to the high levels of genetic similarity of *Map* with other mycobacteria, in particular, *Mycobacterium avium* (*Mav*) (10). Previous studies from our laboratories and elsewhere show >95% nucleotide sequence similarity between many strains of *Map* and *Mav* (11–13). Therefore, it is widely recognized that the development of rapid, sensitive, and specific assays to identify infected animals is essential to the formulation of rational strategies to control the spread of *Map*.

As a first step toward elucidating the molecular basis of *Map*'s physiology and virulence, and providing a foundation for the development of the next generation of *Map* diagnostic tests and vaccines, we report the complete genome sequence of a common clone of *Map*, strain K-10.

Materials and Methods

Bacterial Strains. We chose to sequence *Map* K-10, a bovine clinical isolate, because it is a virulent, low passage clinical strain that was isolated from a dairy herd in Wisconsin by investigators at the U.S. Department of Agriculture National Animal Disease Center in the mid-1970s. In addition, K-10 is amenable to genetic manipulation by transposon mutagenesis (14). The organism was grown in Middlebrook 7H9 broth supplemented with OADC (Difco), Tween 80, and mycobactin J (Allied Monitor, Fayette, MO).

Sequence Analysis. A random shotgun approach was adopted to sequence the genome of *Map* K-10 (15). To create a small (1.8- to 3.0-kb) insert library, genomic DNA was initially isolated by using a chloroform/cetyltrimethylammonium bromide-based method, as described (16). The DNA was sheared by nebulization (www.genome.ou.edu) and 1.8- to 3.0-kb fragments were cloned into pUC18 for isolation and sequencing. Approximately 30,240 clones and 3,000 PCR fragments were sequenced from both ends via Dye-terminator chemistry on Applied Biosystems 3700 and 3100 sequencing machines. A total of 66,129 sequences were used to generate the final assembly, representing a 6-fold coverage of the genome. Sequence assembly was performed with PHREDPHRAP (http://genome.washington.edu), and the ≈400 gaps that remained at the end of the shotgun phase were closed by primer walking and multiplex random PCR.

Informatics. ORFs were predicted by both ARTEMIS and GLIMMER and verified manually in ARTEMIS. BLASTP analysis of the ORFs was performed at the Computational Biology Center at the

Abbreviations: *Map*, *Mycobacterium avium* subspecies *paratuberculosis*; *Mav*, *Mycobacterium avium*; *Mtb*, *Mycobacterium tuberculosis*.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. AE016958).

‡L.L., J.P.B., Q.Z., A.A., S.K., and V.K. have a financial conflict of interest that results from a patent application that has been filed on some DNA sequences that are disclosed and discussed in the manuscript. The patent applications that have been filed are jointly owned by the University of Minnesota and the U.S. Department of Agriculture. As named inventors, these authors may potentially financially benefit from the commercialization of the technology. In addition, some of the technology disclosed in this paper has also been licensed to ANDX, Inc., a University of Minnesota based startup company focusing on the development of molecular diagnostic assays, in which S.K. and V.K. have financial interests and are cofounders.

‖To whom correspondence should be addressed. E-mail: vkapur@umn.edu.

© 2005 by The National Academy of Sciences of the USA

Table 1. Summary of the complete genome of *M. paratuberculosis* K-10 and the comparison with other *Mycobacterium* species

Property	*Map*	*Mav*	*Mtb*	*M. bovis*	*M. leprae*	*M. smegmatis*
Genome size, bp	4,829,781	5,475,738	4,411,532	4,345,492	3,268,203	6,988,209
G+C content, %	69.30	68.99	65.61	65.63	57.79	67.40
Protein coding, %	91.30	NA	90.80	90.59	49.50	92.42
ORFs	4,350	NA	3,959	3,953	1,604	6,897
Gene density, bp per gene	1,112	NA	1,114	1,099	2,037	1,013
Average gene length, bp	1,015	NA	1,012	995	1,011	936
tRNAs	45	45	45	45	45	47
rRNA operon	1	1	1	1	1	2

University of Minnesota (www.cbc.umn.edu), and transfer RNAs were predicted with TRNASCAN-SE (17–20). Comparative genomic analysis was peformed with *Mycobacterium tuberculosis* (*Mtb*), strain H37Rv (GenBank accession no. NC_000962) and *Mav* strain 104 (The Institute for Genomic Research, www.tigr.org) (21, 22).

Results and Discussion

Characteristics of the *Map* Genome. The analysis showed that *Map* K-10 has a single circular sequence of 4,829,781 base pairs, with a G+C content of 69.3% (Table 1 and Fig. 1). The putative origin of replication was identified based on the presence of *dnaA* boxes, characteristic oligomer skew, and G-C skew between the putative genes *dnaA* and *dnaN* (18, 23). The initiation codon for the dnaA gene was chosen as the start point for numbering the genome (Fig. 1). The G+C content is relatively constant throughout the genome. The analysis also revealed only a few genomic regions with lower G+C content corresponding to prophages or coding RNA sequences (Fig. 1).

As in the three other mycobacterial genomes sequenced to date, *Mtb*, *Mycobacterium leprae*, and *Mycobacterium bovis*, a single *rrn* operon (16S-23S-5S) was identified in K-10 along with 50 additional genes coding for functional RNA molecules. The *rrn* operon is located ≈2.75 Mb from the putative *oriC* on the opposite strand. This is ≈1.3 Mb further from the *oriC* than what is described in *Mtb* (22).

Repetitive DNA in *Map*. Approximately 1.5% (or 72.2 kb) of the *Map* genome is comprised of repetitive DNA like insertion sequences, multigene families, and duplicated housekeeping genes. The analysis also identified 17 copies of the previously described insertion sequence *IS900*, seven copies of *IS1311*, and three copies of *ISMav2* in the K-10 genome (Fig. 2). A total of 16 additional *Map* insertion sequence elements were identified in the analysis, totaling 19 different insertion sequences with 58 total copies in the K-10 genome. Although many of these newly

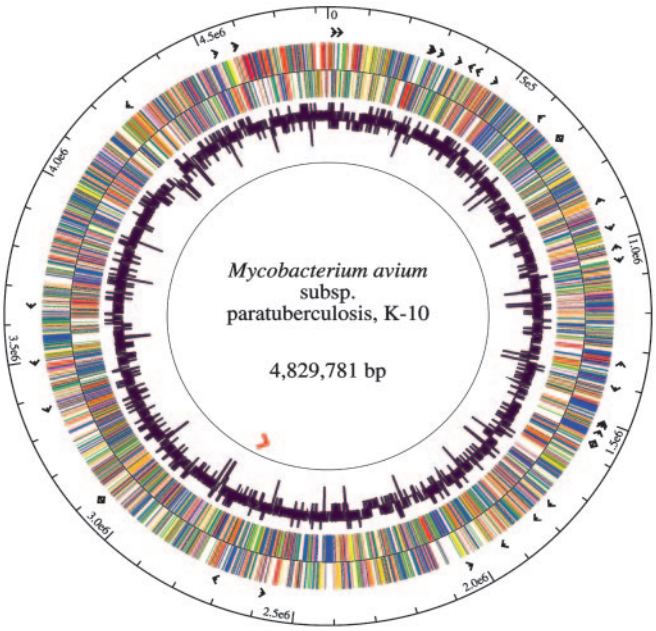

Fig. 1. Circular representation of the *Map* K-10 genome. From inside: red arrow, rRNA operon; dark purple histogram, GC content; multicolored histogram, MAP ORFs coded according to functional classification (small molecule metabolism, blue; macromolecule metabolism, red; cell processes, purple; other processes, yellow; hypotheticals, green; unknowns, orange). The outer colored histogram indicates the same direction of transcription as the origin of replication. The inner colored histogram indicates the opposite direction of transcription as the origin of replication. Black arrows, 45 tRNAs. Outer circle, scale. The figure was generated with GENESCENE software (DNAstar, Madison, WI).

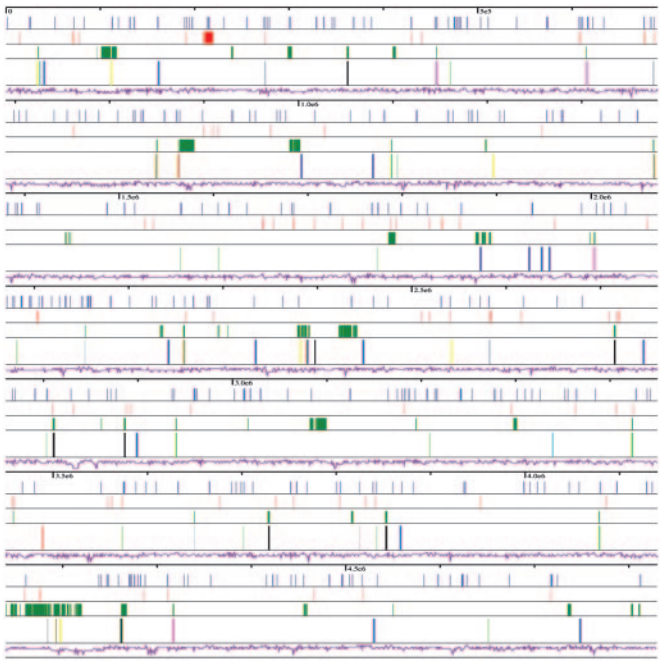

Fig. 2. Linear representation of repeat and unique regions within the *Map* K-10 genome. From the top: Blue histogram, VNTR/DR repeats; red histogram, SSR repeats; green histogram, unique regions; multicolored histogram, insertion sequences (IS900, blue; MAP01, aqua; MAP02, gray; MAP03, pink; MAP04, green; MAP05, purple; MAP06, black, MAP07, orange; MAP08, red; MAP09, light pink; MAP10, light purple; MAP11, beige; MAP12, brown; MAP13,fuschia; MAP14, light aqua; MAP15, light gray; MAP 16, dark green; IS1311, maroon; Mav2, light blue; REP, light green). Purple histogram, GC content. The figure was generated with GENESCENE software (DNAstar, Madison, WI).

AGRICULTURAL SCIENCES

discovered *Map* IS elements are homologs with previously described insertion sequences in *Mtb*, *Mav*, *M. bovis*, and *Mycobacterium marinum*, the analysis also revealed several insertion sequences with no identifiable homologs in other mycobacteria. For example, *IS_MAP02*, present in six copies in the K-10 genome (Fig. 2), has no identified homolog in other mycobacteria and only very low levels of homology (28% identity) with a transposase described in *Legionella pneumophila*. Similarly, *IS_MAP04*, present in four copies in the K-10 genome, has no homologs in other mycobacteria but is similar to insertion sequences found in *Arthrobacter nicotinovorans* and *Streptomyces coelicolor*. These newly discovered IS elements are of particular interest for their use as specific potential diagnostic targets due to their absence in other mycobacteria. In addition, the analysis identified 12 homologs to the REP13E12 family in K-10 (Fig. 2); this is a ≈1,400-bp repeated insertion sequence that was first described in the *Mtb* genome (24, 25).

It is believed that insertion sequences preferentially integrate within intergenic regions so as to avoid the disruption of essential genes (26). Consistent with this hypothesis, the majority of the IS elements found in K-10 appear to be clustered within intergenic regions. For example, MAP0028c and MAP0029c, MAP0849c and MAP0850c, and MAP2155, MAP2156, and MAP2157 are clustered within 5 kb of each other in noncoding regions of the chromosome (Fig. 2). The analysis also shows that insertion sequences in *Map* are absent from the region flanking 32 kb of either side of the *oriC*. A similar observation was made for the *Mtb* genome; however, in the case of *Mtb*, this distance is considerably greater at 600 kb (24). It is thought that there may be detrimental effects to chromosomal replication when insertion sequences are located close to the *oriC*, thereby raising the intriguing possibility that the presence of an insertion sequence, (MAP0028c/IS1311), 32 kb from the *oriC* in *Map*, may contribute to the increased generation interval of *Map* as compared with *Mtb* and other mycobacteria (24).

Protein-Encoding Genes. The K-10 genome contains 4,350 ORFs with lengths ranging from 114 bp (a ribosomal subunit encoding gene) to 19,155 bp (a peptide synthetase), which, in sum, account for 91.5% of the entire genome. A total of 52.5% of the genes are transcribed with the same polarity as that of DNA replication, a fraction that is slightly lower than the 59% observed in *Mtb*. Interestingly, there appears to be a higher bias toward transcription in the same polarity as replication in some other organisms (for e.g., 75% in *Bacillus subtilis*) (26, 27). Although it is tempting to speculate that this bias may, in part, contribute to the slow growth in *Mtb* and the even slower growth in *Map*, it is important to note that only 55% of *Escherichia coli* genes are transcribed in the same polarity as replication, suggesting that gene location in relation to the origin of replication cannot fully explain *Map*'s slow growth in laboratory culture (28).

The analysis showed that a total of 60% of the putative proteins in *Map* had homologs to other microbial proteins with known functions and 25% were homologous to hypothetical proteins (Table 2 and Fig. 4, which are published as supporting information on the PNAS web site). A total of 39 predicted proteins are unique to *Map*, with no identifiable homologs in the current databases. Of the predicted proteins, ≈75% had homology to those identified in *Mtb* (22). Interestingly, the functional redundancy caused by gene duplication that was previously observed in *Mtb* (≈52% of genes are functionally redundant) exists to an even greater extent in *Map* (29). Functional redundancy, based on amino acid homology comparisons, is particularly high among genes involved in lipid metabolism and oxidoreduction; for instance, there are 254 predicted genes functioning as oxidoreductases and oxygenases, compared to 171 in *Mtb*.

The *Map* genome encodes the complete set of enzymes for many metabolic pathways including glycolysis, the pentose phosphate pathway, the tricarboxylic acid cycle, and the glyoxylate cycle. However, there are genes and putative pathways missing in the *Map* genome that have been described in *Mtb*. For example, *Mtb* encodes the necessary genes for urease production (*ureABC* and *ureDFG*). Bacterial ureases catalyze the hydrolysis of urea to ammonia and carbon dioxide. The ability to acquire nitrogen from urea is important in the colonization of varying environments, including a host (30). Because *Map* lacks this specific pathway (*ureABC* and *ureDFG*), its ability to acquire nitrogen from urea may differ from that of *Mtb* and other urease-producing bacteria.

It is not surprising that the analysis shows a large number ($n \approx 150$) of genes with regulatory functions in the *Map* genome (Table 2 and Fig. 4). This number is greater than what is found in *Mtb* ($n \approx 100$) and is consistent with the ability of *Map* to survive in a wide range of environmental conditions (22). The analysis identified 14 complete two-component systems in *Map* as compared with the 11 described in *Mtb*. This finding is in stark contrast to what is seen in *B. subtilis* and *E. coli*, where >30 copies have been described (27, 28). It is believed that the low number of two-component systems is offset in *Mtb* by the presence of serine/threonine protein kinases (STPKs), which are part of the phosphorelay system (31); this may be true in *Map* as well, because our analysis identified a total of 16 genes with a potential role STPK biosynthesis.

Evolutionary Relationships of *Map*. To understand the level and nature of nucleotide variation between *Map* and *Mtb*, amino acid substitution rates for the aligned gene sequences were calculated based on the algorithm of Nei and Gojobori (32). For genes with critical roles in bacterial survival and fitness, the rate of nonsynonymous substitution (dN) is often lower than the rate of synonymous substitution (dS) because these genes are under functional constraint and the majority of phenotypic changes will be disadvantageous. In comparison to *Mtb*, the majority of *Map* dN:dS ratios fall in the range of 1:10 to 1:4, corresponding to the most highly conserved sequences (Table 3, which is published as supporting information on the PNAS web site). The low average ratio of dN:dS for *Map* compared to *Mtb* indicates that the majority of *Map* genes are under functional constraint (Fig. 5, which is published as supporting information on the PNAS web site). The analysis also identified 63 orthologs with dN/dS ratios >2, indicating that substitutions that result in a change in the primary structure of the proteins are overrepresented at these loci. Of particular interest is a gene (MAP0638) with the highest dN/dS ratio of 7.3 that has homology to an amidophosphoribosyltransferase (ATase). This gene is primarily responsible for regulation and production of purine synthesis and has also been shown to play a role in the rate of DNA and protein synthesis and cell growth (33). It is tempting to speculate that the increased level of substitution in this gene may play a role in a decreased growth rate in *Map* in comparison to *Mtb*, a hypothesis that remains to be tested. Although most of the putative protein products from these highly substituted genes were hypothetical, other genes of interest included several transcriptional regulators and 11 genes that play a putative role as cell envelope proteins (Table 4, which is published as supporting information on the PNAS web site).

Mycobactin Synthesis. One major phenotypic difference between *Map* and other mycobacteria is its inability to produce mycobactin in laboratory culture. Mycobactin is a siderophore that is responsible for the binding or transport of iron into cells. Because of the importance of iron in electron transport and as a key component of various metabolic enzymes, it is essential that bacteria have the ability to acquire iron from various sources. A cluster of 10 genes in *Mtb* (*mbtA–J*) has been shown to be responsible for the production of mycobactin and the

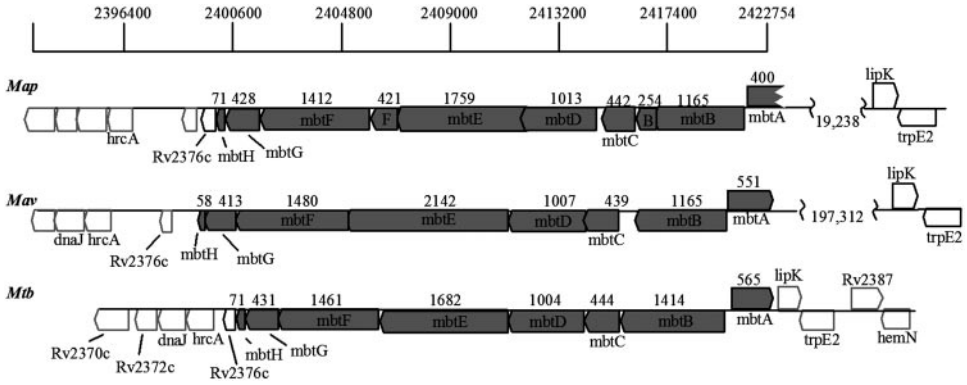

Fig. 3. Homology comparison of the mycobactin gene cluster between *Map* K-10, *Mav* 104, and *Mtb* H37Rv. The gene cluster is shown in gray with amino acid length indicated above each box. Gene names are indicated.

transport of iron (34). Homologs to the *mbtA–J* cluster were identified in the *Map* genome. However, a direct comparison of the MbtA–J cluster in *Map* with those of *Mav* and *Mtb* show significant differences in primary structure of this region (Fig. 3 and Table 5, which is published as supporting information on the PNAS web site). First, in *Mtb*, *mbtI* (*trpE2*) and *mbtJ* (*lipK*) are adjacent to each other and immediately downstream of the mycobactin biosynthesis gene cluster (Fig. 3) (34). However, in *Map* and *Mav*, there is a 6.6- and 5.7-kb gap, respectively, between the two genes. Furthermore, there is 19.3-kb gap in *Map* and 197.3-kb gap in *Mav* between *mbtJ* and *mbtA* of *mbt* cluster, confirming a preliminary observation made by DNA microarray analysis (35). Because *Mav* can successfully produce mycobactin, the distance in genes is not likely to be the limiting factor that affects the production of mycobactin in *Map*. Second, we identified frame-shift mutations in both *mbtB* and *mbtE* in *Map* when compared with *Mtb*. The *mbtB* frameshift also appears in *Mav*, suggesting that this too is not the limiting factor affecting mycobactin production in *Map* (Fig. 3). In addition, even though the gene is divided, all functional domains identified in *Mtb* appear to be present in the *mbtB* gene product in *Map*, suggesting that its function could still be maintained. A similar frameshift exists in the *mbtE* gene in *Map* as well, but the functional domains of MbtE that are present in *Mtb* are present in *Map* as well. For each of *mbtC*, *mbtD*, and *mbtF*, the gene products include the common domains that are found in corresponding proteins in *Mtb*.

The major difference between *Map*, *Mav*, and *Mtb* in the *mbtA-J* cluster was in the *mbtA* gene. Gene *mbtA* is shorter in *Map*: encoding a 400-aa protein, compared with 565- and 551-residue polypeptides in *Mtb* and *Mav*, respectively. As a result of this truncation, MbtA has only 337 residues in *Map* with homology to the N-terminal of the EntE domain, and lacks >200 residues of the EntE C terminus that are presumably important for protein function. Because MbtA is thought to initiate mycobactin production, the truncation observed in this key gene suggests that the entire cascade leading to mycobactin production may be attenuated in *Map* (34). Therefore, it is tempting to speculate that the truncated EntE domain in the *mbtA* gene product in *Map* might be the limiting factor in mycobactin production, a hypothesis that remains to be formally tested.

Immunological and Virulence-Related Insights from the *Map* Genome. Despite intensive research efforts, there is still little information regarding the molecular basis for *Map* pathogenesis. Hence, we paid particular attention to the identification of genes with a potential role in pathogenesis in the *Map* genome and note several interesting observations. There is a paucity in the number of the PE/PPE family of proteins that are thought to play an important role in mycobacterial infection from both an antigenic as well as an immunologic standpoint. These proteins are acidic and glycine-rich proteins, and are identified by their specific domains (Pro-Pro-Glu and Pro-Glu, respectively) that frequently contain polymorphic repetitive sequences (PGRSs) and multiple copies of major polymorphic tandem repeats, respectively (22, 36–39). These proteins are thought to be expressed on the cell surface and provide the antigenic variation that elicits varying immunological responses in *Mtb* depending on the type of PE/PPE protein expressed on the cell surface (22, 40). Genome-scale comparisons between two isolates of *Mtb* show that regions of the genome encoding PE/PPE proteins have a higher single-base substitution frequency, further supporting the hypothesis that they are recognized by the immune system and hence subject to positive Darwinian selection (40). Although these families of proteins comprise 10% of the *Mtb* genome, there were only six PE homologs and 36 PPE homologs in *Map* (comprising 1% of the genome) compared to 38 and 68, respectively, in *Mtb* (22). Within the PE family, there is no intact PE-PGRS subfamily of proteins identified in the K-10 genome, although this subfamily has been identified in other mycobacteria including *M. bovis* and *M. marinum* (41). Interestingly, this subfamily of proteins is also absent in *Mav* and *M. leprae* (42). Although the exact significance of this observation is unknown, it may suggest a more limited, less variable, and different immune response toward *Map* as compared with *Mtb*. This observation also leads to the tempting speculation that antimicrobial agents and vaccines directed against these major virulence factors may be more likely to be effective against *Map* as compared with *Mtb*.

Pathogens often express proteins that alter the effects of the host's immune response so as to evade destruction. Because mycobacteria are facultative pathogens and are assumed to selectively express specific genes to allow for survival inside the host macrophage, much attention has been directed toward the characterization of virulence genes that are important for the entry and persistence in the host (43). One such gene, the mammalian cell entry (*mce*) gene, has been identified in *Mtb* (44) and was shown to enhance *E. coli*'s ability to survive in macrophages (44). Four copies of the *mce* gene are present in *Mtb* (22). The analysis of the complete sequence of the *Map* genome revealed eight homologs of the *mce* gene. The gene has also been identified in different mycobacterial species, including *Mav*, *M. bovis*, and *Mycobacterium smegmatis*. The wide distribution of the *mce* operons in pathogenic and nonpathogenic mycobacteria implies that the mere presence of these genes does not endow a bacterium with the ability to cause disease. However, the role of this operon in virulence may be determined by its expression under specific conditions (43, 45, 46).

Mycobacteria were originally classified as such by the presence of mycolic acids (47, 48). Not only do mycobacteria produce this type of lipid, but these organisms are also known for their ability to produce and use a vast array of other lipophilic molecules (47). Importantly, these diverse structures that are located primarily on the cell wall are thought to play a role in pathogenesis in many mycobacterial species by their ability to allow entry into host cells or suppress or evade host immune defense mechanisms (49, 50). Increased survival of mycobacteria may also be enabled, in part, by their ability to preferentially use fatty acids instead of carbohydrates for basic metabolic needs (51–53). Our analysis shows that there are ≈80 more genes in *Map* ($n \approx 266$) that are predicted to be involved in lipid metabolism than there are in *Mtb*. Although this difference in number of lipid metabolism and biosynthesis related genes is due primarily to genetic redundancy in *Map*, there are some noteworthy differences. For example, *Map* contains a gene (MAP3194) encoding hydroxymethylglutaryl–CoA lyase, an enzyme that is found in other bacteria as well as in humans, and catalyzes the last step of ketogenesis and leucine catabolism (54). The enzyme may play a role in fatty acid biosynthesis by altering what is produced and distributed to the cell membrane (54). This difference between *Map* and *Mtb* indicates there may be variation in lipid metabolism and biosynthesis that may play a role in antigens present/absent on the cell surface, thus, affecting host immune defense mechanisms.

Map lacks one of the largest operons in *Mtb* involved in the production of phenolpthiocerol, a polyketide. In the *Mtb* genome, this operon (*ppsABCDE*) is immediately upstream of another gene cluster (*mas*), which encodes an enzyme responsible for the synthesis of mycocerosic acid. In combination, these two products form the abundant cell-wall-associated molecule phthiocerol dimycocerosate (DIM) (22). Importantly, it has been found that *Mtb* isolates lacking this cell wall lipid are attenuated in virulence (55–57). *Map* contains no homologs to either the *ppsABCDE* or the *mas* gene homologs. Instead, the analysis identified 35 other genes with a possible role in polyketide synthesis (including 12 involved in mycobactin biosynthesis; Table 2). Included among these are chalcone synthase-like genes. These polyketides are found primarily in plants, but a four-gene cluster (*pks7-10*) has been identified in *Mtb* with high similarity to a chalcone synthase-like genes (22). Similarly, a four-gene cluster was identified in *Map* (MAP1369, MAP1370, MAP1371, and MAP1372). These polyketides are believed to be involved in the production of DIM in *Mtb* as well as in its virulence (57). Mutations in *pks10* showed attenuation of the virulent *Mtb* H37Rv isolate upon infection of alveolar macrophages (57). Therefore, although *Map* lacks the two major gene clusters required for the production of the polyketide DIM in *Mtb*, the analysis identified other genes that may play a role in DIM and polyketide synthesis in *Map*. These results suggest that there are likely to be considerable differences in the presence or expression of various lipids on the cell surface of *Map* that may in turn have a major influence on the virulence and host specificity of this bacterium. In addition, the relative lack of functional redundancy in this pathway in *Map* suggests that the inactivation of the putative DIM polyketide biosynthesis pathway related genes in *Map* may be a promising approach to the development of attenuated or vaccine strains in this organism.

***Map* Diagnostics and Strain Differentiation.** The identification of unique sequences within the *Map* genome has already provided a foundation on which to design and implement better diagnostic assays for *Map* detection. Our analysis has identified ≈161 unique sequence regions in the *Map* genome, with the longest region being 15.9 kb in length (Fig. 1). More importantly, our preliminary studies show that these unique sequences have considerable potential for the development of more specific and sensitive diagnostic assays for detection of *Map* infection with both molecular and immunoassay based approaches (58–61). Our more recent studies, enabled by the identification of the unique regions in the *Map* genome, have resulted in the development of highly sensitive real-time PCR-based approaches for the sensitive and specific detection of *Map* directly from bovine feces (N.B., L.L., A.A., J.P.B., V.K., and S.K., unpublished data). Studies are also underway to heterologously express unique Map genes as well as the genes encoding cell surface proteins for construction of a partial protein array to evaluate the humoral and cell-mediated immunostimulatory capabilities of these recently discovered unique proteins. Therefore, the combination of genomic information, molecular tools, and immunological assays will provide key insights to the host immune response to *Map* infection. Overall, the elucidation of all of the unique sequences as well as those that may be associated with the cell surface of *Map* provides a strong foundation on which to develop the next generation of specific and sensitive diagnostic assays for *Map*.

Short sequence repeats (SSRs) or variable number tandem repeat (VNTR) sequences have been used as markers for differentiation and subtyping strains of several bacterial species including *Mtb*, *Yersinia pestis*, and *Bacillus anthracis* (62–65). SSRs consist of simple homopolymeric tracts of single nucleotide (mononucleotide repeat) or multimeric tracts (of homogeneous or heterogeneous repeat), such as di- or trinucleotide repeats, which can be identified as VNTRs in the genome of the organism (66). The variability of the repeats is believed to be caused by slipped-strand mispairing, the genetic instability of polynucleotide tracts, especially poly(G-T), and DNA recombination between homologous repeat sequences. In preliminary bioinformatic analyses, we had identified 185 mono-, di-, and trinucleotide repeat sequences dispersed throughout the *Map* genome, of which 78 were perfect repeats (67). Comparative nucleotide sequencing of the 78 loci in six *Map* isolates from different host species and geographic locations identified a subset of 11 polymorphic SSRs with an average of 3.2 alleles per locus (67), and has provided the foundation for the development of highly discriminatory and powerful multilocus SSR (MLSSR)-based typing approach for strain differentiation among isolates of *Map* (68). In the current investigation, we identified an additional 362 sequences representing either direct or indirect sequence repeats of length distribution of 6–74 bp with a repeat number ranging from 2 to 16, and with a mutual homology of 67–100% (Fig. 2). Based on our recent success using MLSSR for *Map* strain differentiation and in understanding the molecular epidemiology of the organism, it is likely that these repeat elements in the *Map* genome will provide additional strain differentiation capabilities and may enable rapid and facile discrimination of epidemiologically and geographically distinct strains of isolates of *Map* by using nonsequencing based approaches as well (69).

It is noteworthy that the availability of the full complement of all of the genes in *Map* provides an opportunity to perform a complete metabolic reconstruction to enable a better understanding of the natural physiology of this organism and the metabolic requirements for growth, and thereby enable faster growth of *Map* in laboratory culture.

Concluding Comments. In summary, the complete genome sequencing and comparative genomics analyses described herein have provided key insights and a strong foundation for future investigations on the genetics, evolution, natural physiology, and virulence of this important animal pathogen. Furthermore, the results of our studies provide the foundation for the development of the next generation of diagnostic and strain differentiation approaches, and provide a framework for the application of genomics and proteomics based approaches for the development of vaccines to prevent and control Johne's disease in domestic livestock.

Funding for this project was provided, in part, by grants from the U.S. Department of Agriculture Cooperative State Research, Education, and Extension Service National Research Initiative competitive grants program as well as the Agricultural Research Service (to V.K. and J.P.B.).

1. Harris, J. E. & Lammerding, A. M. (2001) *J. Food Prot.* **64,** 2103–2110.
2. Chi, J., VanLeeuwen, J. A., Weersink, A. & Keefe, G. P. (2002) *Prev. Vet. Med.* **55,** 137–153.
3. Ott, S. L., Wells, S. J. & Wagner, B. A. (1999) *Prev. Vet. Med.* **40,** 179–192.
4. Grant, I. R., Ball, H. J. & Rowe, M. T. (2002) *Appl. Environ. Microbiol.* **68,** 2428–2435.
5. Streeter, R. N., Hoffsis, G. F., Bech-Nielsen, S., Shulaw, W. P. & Rings, D. M. (1995) *Am. J. Vet. Res.* **56,** 1322–1324.
6. Sweeney, R. W. (1996) *Vet. Clin. North Am. Food Anim. Pract.* **12,** 305–312.
7. Greenstein, R. J. (2003) *Lancet Infect. Dis.* **3,** 507–514.
8. Clarke, C. J. (1997) *Vet. J.* **153,** 245–247.
9. Cousins, D. V., Whittington, R., Marsh, I., Masters, A., Evans, R. J. & Kluver, P. (1999) *Mol. Cell Probes* **13,** 431–442.
10. Colgrove, G. S., Thoen, C. O., Blackburn, B. O. & Murphy, C. D. (1989) *Vet. Microbiol.* **19,** 183–187.
11. Bannantine, J. P., Zhang, Q., Li, L. L. & Kapur, V. (2003) *BMC Microbiol.* **3,** 10.
12. Hurley, S. S., Splitter, G. A. & Welch, R. A. (1989) *J. Clin. Microbiol.* **27,** 1582–1587.
13. Saxegaard, F. & Baess, I. (1988) *Apmis* **96,** 37–42.
14. Foley-Thomas, E. M., Whipple, D. L., Bermudez, L. E. & Barletta, R. G. (1995) *Microbiology* **141,** 1173–1181.
15. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., *et al.* (1995) *Science* **269,** 496–512.
16. Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. & Struhl, K. (1999) in *Current Protocols in Molecular Biology* (Wiley, New York), Vol. 1, pp. 2.4.1–2.4.5.
17. Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. & Barrell, B. (2000) *Bioinformatics* **16,** 944–945.
18. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999) *Nucleic Acids Res.* **27,** 4636–4641.
19. Aggarwal, G. & Ramaswamy, R. (2002) *J. Biosci.* **27,** 7–14.
20. Lowe, T. M. & Eddy, S. R. (1997) *Nucleic Acids Res.* **25,** 955–964.
21. Cole, S. T. & Barrell, B. G. (1998) *Novartis Found. Symp.* **217,** 160–172; discussion, 172–177.
22. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., III, *et al.* (1998) *Nature* **393,** 537–544.
23. Lobry, J. R. (1996) *Mol. Biol. Evol.* **13,** 660–665.
24. Gordon, S. V., Heym, B., Parkhill, J., Barrell, B. & Cole, S. T. (1999) *Microbiology* **145,** 881–892.
25. Lee, T. Y., Lee, T. J., Belisle, J. T., Brennan, P. J. & Kim, S. K. (1997) *Tuber. Lung Dis.* **78,** 13–19.
26. Perret, X., Viprey, V., Freiberg, C. & Broughton, W. J. (1997) *J. Bacteriol.* **179,** 7488–7496.
27. Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., *et al.* (1997) *Nature* **390,** 249–256.
28. Blattner, F. R., Plunkett, G., III, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., *et al.* (1997) *Science* **277,** 1453–1474.
29. Tekaia, F., Gordon, S. V., Garnier, T., Brosch, R., Barrell, B. G. & Cole, S. T. (1999) *Tuber. Lung Dis.* **79,** 329–342.
30. Burne, R. A. & Chen, Y. Y. (2000) *Microbes Infect.* **2,** 533–542.
31. Av-Gay, Y. & Everett, M. (2000) *Trends Microbiol.* **8,** 238–244.
32. Nei, M. & Gojobori, T. (1986) *Mol. Biol. Evol.* **3,** 418–426.
33. Yamaoka, T., Yano, M., Kondo, M., Sasaki, H., Hino, S., Katashima, R., Moritani, M. & Itakura, M. (2001) *J. Biol. Chem.* **276,** 21285–21291.
34. Quadri, L. E., Sello, J., Keating, T. A., Weinreb, P. H. & Walsh, C. T. (1998) *Chem. Biol.* **5,** 631–645.
35. Semret, M., Zhai, G., Mostowy, S., Cleto, C., Alexander, D., Cangelosi, G., Cousins, D., Collins, D. M., van Soolingen, D. & Behr, M. A. (2004) *J. Bacteriol.* **186,** 6332–6334.
36. Ramakrishnan, L., Federspiel, N. A. & Falkow, S. (2000) *Science* **288,** 1436–1439.
37. Skeiky, Y. A., Ovendale, P. J., Jen, S., Alderson, M. R., Dillon, D. C., Smith, S., Wilson, C. B., Orme, I. M., Reed, S. G. & Campos-Neto, A. (2000) *J. Immunol.* **165,** 7140–7149.
38. Hermans, P. W., van Soolingen, D. & van Embden, J. D. (1992) *J. Bacteriol.* **174,** 4157–4165.
39. Poulet, S. & Cole, S. T. (1995) *Arch. Microbiol.* **163,** 87–95.
40. Fleischmann, R. D., Alland, D., Eisen, J. A., Carpenter, L., White, O., Peterson, J., DeBoy, R., Dodson, R., Gwinn, M., Haft, D., *et al.* (2002) *J. Bacteriol.* **184,** 5479–5490.
41. Banu, S., Honore, N., Saint-Joanis, B., Philpott, D., Prevost, M. C. & Cole, S. T. (2002) *Mol. Microbiol.* **44,** 9–19.
42. Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honore, N., Garnier, T., Churcher, C., Harris, D., *et al.* (2001) *Nature* **409,** 1007–1011.
43. Haile, Y., Bjune, G. & Wiker, H. G. (2002) *Microbiology* **148,** 3881–3886.
44. Arruda, S., Bomfim, G., Knights, R., Huima-Byron, T. & Riley, L. W. (1993) *Science* **261,** 1454–1457.
45. Kumar, A., Bose, M. & Brahmachari, V. (2003) *Infect. Immun.* **71,** 6083–6087.
46. Kumar, A., Chandolia, A., Chaudhry, U., Brahmachari, V. & Bose, M. (2005) *FEMS Immunol. Med. Microbiol.* **43,** 185–195.
47. Besra, G. S., Sievert, T., Lee, R. E., Slayden, R. A., Brennan, P. J. & Takayama, K. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 12735–12739.
48. Minnikin, D. E. & Goodfellow, M. (1980) *Soc. Appl. Bacteriol. Symp. Ser.* **8,** 189–256.
49. Rhoades, E. R. & Ullrich, H. J. (2000) *Immunol. Cell Biol.* **78,** 301–310.
50. Schorey, J. S., Carroll, M. C. & Brown, E. J. (1997) *Science* **277,** 1091–1093.
51. Bloch, H. & Segal, W. (1956) *J. Bacteriol.* **72,** 132–141.
52. Segal, W. & Bloch, H. (1957) *Am. Rev. Tuberc.* **75,** 495–500.
53. Smith, I. (2003) *Clin. Microbiol. Rev.* **16,** 463–496.
54. Ashmarina, L. I., Rusnak, N., Miziorko, H. M. & Mitchell, G. A. (1994) *J. Biol. Chem.* **269,** 31929–31932.
55. Cox, J. S., Chen, B., McNeil, M. & Jacobs, W. R., Jr. (1999) *Nature* **402,** 79–83.
56. Sirakova, T. D., Dubey, V. S., Cynamon, M. H. & Kolattukudy, P. E. (2003) *J. Bacteriol.* **185,** 2999–3008.
57. Sirakova, T. D., Thirumala, A. K., Dubey, V. S., Sprecher, H. & Kolattukudy, P. E. (2001) *J. Biol. Chem.* **276,** 16833–16839.
58. Bannantine, J. P., Hansen, J. K., Paustian, M. L., Amonsin, A., Li, L. L., Stabel, J. R. & Kapur, V. (2004) *J. Clin. Microbiol.* **42,** 106–114.
59. Bannantine, J. P., Huntley, J. F., Miltner, E., Stabel, J. R. & Bermudez, L. E. (2003) *Microbiology* **149,** 2061–2069.
60. Bannantine, J. P., Baechler, E., Zhang, Q., Li, L. & Kapur, V. (2002) *J. Clin. Microbiol.* **40,** 1303–1310.
61. Paustian, M. L., Amonsin, A., Kapur, V. & Bannantine, J. P. (2004) *J. Clin. Microbiol.* **42,** 2675–2681.
62. Gascoyne-Binzi DM, B. R., Frothingham R, Robinson G, Collyns TA, Gelletlie R, Hawkey PM. *J. Clin. Microbiol.* **39,** 69–74.
63. Adair, D. M., Worsham, P. L., Hill, K. K., Klevytska, A. M., Jackson, P. J., Friedlander, A. M. & Keim, P. (2000) *J. Clin. Microbiol.* **38,** 1516–1519.
64. Kim, S. G., Shin, S. J., Jacobson, R. H., Miller, L. J., Harpending, P. R., Stehman, S. M., Rossiter, C. A. & Lein, D. A. (2002) *J. Vet. Diagn. Invest.* **14,** 126–131.
65. Kim, W., Hong, Y. P., Yoo, J. H., Lee, W. B., Choi, C. S. & Chung, S. I. (2002) *FEMS Microbiol. Lett.* **207,** 21–27.
66. Wiid, I. J., Werely, C., Beyers, N., Donald, P. & van Helden, P. D. (1994) *J. Clin. Microbiol.* **32,** 1318–1321.
67. Amonsin, A., Li, L. L., Zhang, Q., Bannantine, J. P., Motiwala, A. S., Sreevatsan, S. & Kapur, V. (2004) *J. Clin. Microbiol.* **42,** 1694–1702.
68. Ghadiali, A. H., Strother, M., Naser, S. A., Manning, E. J. & Sreevatsan, S. (2004) *J. Clin. Microbiol.* **42,** 5345–5348.
69. Motiwala, A. S., Amonsin, A., Strother, M., Manning, E. J., Kapur, V. & Sreevatsan, S. (2004) *J. Clin. Microbiol.* **42,** 1703–1712.