M. paratuberculosis strain 397
genome optical map contig
5.21 Mb

Eukaryote Microbial Genomics

Bannantine *et al.*

BMC Genomics

RESEARCH ARTICLE

Open Access

Genome sequencing of ovine isolates of *Mycobacterium avium* subspecies *paratuberculosis* offers insights into host association

John P Bannantine[1*], Chia-wei Wu[2], Chungyi Hsu[2], Shiguo Zhou[3], David C Schwartz[3], Darrell O Bayles[1], Michael L Paustian[1], David P Alt[1], Srinand Sreevatsan[4], Vivek Kapur[5] and Adel M Talaat[2,6*]

Abstract

Background: The genome of *Mycobacterium avium* subspecies *paratuberculosis* (*MAP*) is remarkably homogeneous among the genomes of bovine, human and wildlife isolates. However, previous work in our laboratories with the bovine K-10 strain has revealed substantial differences compared to sheep isolates. To systematically characterize all genomic differences that may be associated with the specific hosts, we sequenced the genomes of three U.S. sheep isolates and also obtained an optical map.

Results: Our analysis of one of the isolates, *MAP* S397, revealed a genome 4.8 Mb in size with 4,700 open reading frames (ORFs). Comparative analysis of the *MAP* S397 isolate showed it acquired approximately 10 large sequence regions that are shared with the human *M. avium* subsp. *hominissuis* strain 104 and lost 2 large regions that are present in the bovine strain. In addition, optical mapping defined the presence of 7 large inversions between the bovine and ovine genomes (~ 2.36 Mb). Whole-genome sequencing of 2 additional sheep strains of *MAP* (JTC1074 and JTC7565) further confirmed genomic homogeneity of the sheep isolates despite the presence of polymorphisms on the nucleotide level.

Conclusions: Comparative sequence analysis employed here provided a better understanding of the host association, evolution of members of the *M. avium* complex and could help in deciphering the phenotypic differences observed among sheep and cattle strains of *MAP*. A similar approach based on whole-genome sequencing combined with optical mapping could be employed to examine closely related pathogens. We propose an evolutionary scenario for *M. avium* complex strains based on these genome sequences.

Keywords: *M. paratuberculosis*, Evolution, Johne's disease, Genome, Optical mapping

Background

Mycobacterium avium subspecies *paratuberculosis* (*MAP*) causes Johne's disease in sheep, cattle, goats and other ruminant animals. This disease is chronic in nature with multiple years separating the initial infection from clinical stages of disease [1]. The details of the pathogenic mechanisms occurring during this long incubation period still need further study, but it has been demonstrated that MAP colonizes the small intestine through invasion of both M cells and epithelial cells [2]. The disease is of considerable economic significance to livestock industries, particularly the dairy industry.

Generally, *MAP* is a genetically homogenous subspecies, especially among bovine, human and wildlife isolates [3-5]. However, three lineages of *MAP* have emerged following extensive molecular strain typing and comparative genomic studies-type I and type III strains (ovine) and type II (bovine) strains. The type III strains were originally called intermediate strains and are highly similar genetically, and thus, difficult to distinguish from type I strains. Early on, the type I (*MAP*-S) and type II (*MAP*-C) strains were distinguished based on their molecular fingerprints using IS*1311* polymorphism [6],

* Correspondence: John.Bannantine@ARS.USDA.GOV; atalaat@wisc.edu
[1]National Animal Disease Center, USDA-Agricultural Research Service, Ames, Iowa, USA
[2]The Laboratory of Bacterial Genomics, Department of Pathobiological Sciences, University of Wisconsin-Madison, Madison, Wisconsin, USA
Full list of author information is available at the end of the article

representational difference analysis [7], MLSSR typing [8-10] and *hsp65* sequencing [11]. On the other hand, type III (a sub-lineage of the *MAP*-S strains) was genotyped based on *gyrA* and *gyrB* genes [12].

In addition to these recently published genotypic distinctions between "S" and "C" strains of *MAP*, phenotypic differences have been noted since the middle of the last century [4]. More recently, Motiwala et al. [13] have shown transcriptional changes in human macrophages infected with *MAP*-C, human and bison isolates induce an anti-inflammatory gene expression pattern, while the *MAP*-S isolates showed expression of pro-inflammatory cytokines. Furthermore, some of the ovine strains are pigmented [14]. The ovine and bovine strains likewise are distinct in their growth characteristics. The *MAP*-S strains are more fastidious and slower in their growth rate than the *MAP*-C counterpart. In contrast to *MAP*-C strains, the *MAP*-S strains do not grow readily on Herold's egg yolk media or Middlebrook 7H9 media that is not supplemented with egg yolk [15]. Nutrient limitation will kill *MAP*-S strains but it is only bacteriostatic for *MAP*-C strains [16]. On the transcriptional level, RNA extracted in low iron and heat stressed environments is divergent between *MAP*-S and *MAP*-C strains [17]. Recently, iron storage in low iron conditions was only observed in the *MAP*-C strains but not *MAP*-S strains [18]. Because of these well-documented phenotypic differences, we hypothesized that sequencing of the genomes of ovine isolates and comparing them to other genomes in the MAC group could provide some clues for these host-specific variations.

The *MAP*-C strain K-10 was sequenced in 2005 to obtain a complete genome 4.8 Mb in size [19]. It was subsequently found to possess an inversion due to misalignment that was resolved by optical mapping [20]. Very recently, draft sequences of ten *MAP* isolates have been reported with the presence of two large duplications, especially among human isolates [21]. Finally, another *M. avium* subspecies (strain 104) has also been sequenced but not published as yet. This genome of subspecies *hominissuis* is 5.4 Mb in size and greater than 95% homologous to the *MAP* K-10 genome [3,5,22]. Both of these genomes have served as reference genomes in the current project to assist in assembly, open reading frame (ORF) predictions, and annotation. With the help of next-generation sequencing and optical mapping, we were able to assemble a draft of the standard sheep strain of *MAP* S397 and compare its sequence to other clinical isolates from sheep or the K-10 strain. Interestingly, several inversion regions and single nucleotide polymorphisms distinguished the *MAP*-S strains from their *MAP*-C counterpart. Insights into the evolution of *MAP* strains have been gained through this analysis.

Results

Genome general features

Pyrosequencing indicated that the *MAP* strain S397 has a circular chromosome with at least 4,814,922 bp, a G + C content of 69.31% and contains 4,700 predicted open reading frames (ORFs). The majority of these genes (44.5%) were predicted [23] to encode cytoplasmic proteins (Additional file 1: Table S1) involved in various cellular functions and a minority of extracellular proteins (< 1%). The number of annotated genes in S397 was more than the bovine K-10 strain (Table 1) due to the different annotation methods used on each genome [19]. However, like *MAP* K-10, the S397 genome contains one rRNA operon and 46 tRNA genes representing all 20 amino acids. A detailed comparison between *MAP* strains K-10 and S397 as well as the human, *MAH* 104 is shown in Table 1. The *de novo* assembly of the compiled S397 genome had an average sequencing depth of 24 × in 184 scaffolds (Additional file 2: Table S2). When aligned to the K-10 sequence, over 110 of these scaffolds are separated by a sequence gap of less than 500 bp suggesting the small size of most gaps. Furthermore, when gaps of 3.5 kb or less were ignored, we were able to assemble the whole genome into 3 scaffolds. The two largest sequence gaps are between contig00150c and contig00149c, which is estimated at 30.19 kb and the contig00082-contig00041c gap, which is estimated at 18.87 kb. Additional file 3: Table S3 gives an overview of the ordered scaffolds.

Analysis of the two additional genomes sequenced in this study (JTC1074 and JTC7565) revealed more than 99% identity to the S397 genome sequence (Table 2). A de novo assembly of these genomes sequenced using Illumina platform produced an average sequence depth of 60 ×. As expected, no significant differences were found between the common features of the 3 sequenced sheep isolate genomes. In fact, there were no gene

Table 1 A summary of the genomic features of *M. avium* subspecies isolates from different hosts

	MAP K-10	MAP S397	MAH 104
Origin	Bovine	Ovine	Human
Genome size (bp)	4,829,781	4,814,922	5,475,491
DNA scaffolds	1	184	1
G + C (%)	69.30	69.31	68.99
Protein coding (%)	91.56	91.31	88.66
Total genes	4,415	4,700	5,305
Total protein coding genes (PCG)	4,350	4,642	5,240
PCG without function prediction	3,014	1,085	1,732
PCG connected to KEGG pathways	1,215	1,249	1,303
tRNAs	46	46	46
rRNA operon	1	1	1

Table 2 Reference genome assembly of clinical ovine isolates using simulated *MAP* S397 genome

	JTC1074	JTC7565
Reference organism	MAP S397	MAP S397
Reference length	4,766,015	4,766,015
Consensus length	4,753,502	4,742,737
%Homology[a] to S397	99.74	99.51
%Homology to K-10	98.79	98.56
Average Coverage[b]	59.89	62.77
Standard deviation	32.36	34.65
Non-specific matches read count[c]	1387	1461
Paired read distance distribution	140-360	120-380
No. of SNPs	70	53

[a]Homology% was calculated as: consensus length divided by reference length and then multiplied by 100

[b]Average coverage is the average of all the reads in each area in the consensus sequence

[c]Non-specific match read counts are those reads that can be matched more than one place in the reference genome and such reads were randomly placed in one of the matched spots

differences; hence all three genomes were identically annotated. Similar to other sequenced mycobacterial genomes, *dnaA* was assigned the first locus tag (MAPs_00010).

The IS elements usually play a role in the genomic diversity among strains of mycobacteria [24] and could act as a good target for molecular diagnostics [25]. Similar to K-10, the S397 genome has all the well-studied insertion sequences (e.g. IS*900*, IS*1311* and IS_*map02*). IS*900* is generally considered a *MAP* specific element that was originally discovered in 1989 [26,27]. A total of 17 copies of IS*900* were found in the S397 genome, which is identical to the K-10 strain. Another element, IS_*map02*, is a *MAP* specific insertion sequence that was discovered by sequencing the K-10 genome. A total of 6 copies of IS_*map02* are present in both S397 and K-10. Likewise, IS*1311* is present 7 times in each genome. No IS elements were found to be unique to one or the other genome.

Organization of the *MAP* S397 genome

Sequence analysis alone was not sufficient to decipher the synteny of the genome. Previously, we used an optical mapping protocol to confirm the organization of the *MAP* K-10 genome [20]. A similar strategy was used to analyze the genome of S397. The raw optical map dataset comprised 2,950 single molecule maps with a total mass of 784.5 Mb, and an average molecule size of 333.6 Kb (Figure 1). After assembly, the compiled optical map contained 905 single molecule optical maps (301.9 Mb; total mass), which covers the genome 58 ×. After a G + C content adjustment by a factor of 0.95, the estimated size of *MAP* S397 optical map is 4.95 Mb, which is slightly higher than the sequence data

suggested. However, if the estimated sequence gaps are added in, the estimated sizes are very similar.

To our surprise, there were 7 inversions that are larger than 22 kb when the S397 genome was compared to the sequenced genome of K-10 compiled by Wynne [20,28]. The total size of these inversions spanned 2.4 Mb of the S397 genome. Individual sizes of those inversions range from 22 to 1,174 kb. As shown in Figure 2B, homologous segments between *MAP* K-10 and S397 are represented by color boxes and to each segment a number was assigned. Detail information of each segment is shown in Table 3. Thirteen out of the 14 segments have at least one IS element on the flanking regions (Figure 2).

Similar to our analysis of inversions discovered in the K-10 strain, we used a PCR-based approach to examine two of the inversion breakpoints in the S397 genome (Figure 3), which are the right end of segment ID #1 and the left end of segment ID#2 (Table 3). As expected, our PCR analysis confirmed the inversion predicted in the genome of K-10 and S397 strains. Because these inversions were readily identified from the optical map and sequence alignment data, we did not attempt to confirm all of the inverted fragments by PCR. Despite these inversions, there is strong synteny between these genomes, underscoring their close relatedness. Both genomes share a number of large-scale clusters of homology where gene order is highly conserved (Additional file 4: Table S4).

Genomic insertions

Further comparative sequence analysis identified several regions that are present in *MAP* S397 and *MAH* 104, but not in *MAP* K-10 (Additional file 5: Table S5). The largest of these is a 9-kb gene cluster encompassing 13 ORFs (MAPs_15940-MAPs_16060). This region was partially identified by representation difference analysis and termed PIG-RDA20 for pigmented strain representational difference analysis-20, as detailed before [7]. It was also mapped to the *MAH* 104 genome by Dohmann and coworkers [7] and was subsequently described by Semret and coworkers as large sequence polymorphism (LSP), LSP[A]4-II [29]. This region contains a copy of the IS*1311* insertion sequence and within the *MAH* 104 genome is flanked by an additional copy of IS*1311*. Another previously described LSP included 9 ORFs (MAPs_46190-MAPs_46270) and totals 6.6 kb. This region was partially identified as the PIG-RDA10 sequence and was mapped to a 16 kb segment of the *MAH* 104 genome [7]. The full sequence was later identified as LSP[A]18 [29], which is equivalent to MAV island 24 [3]. An interesting feature of LSP[A]18 is that it begins and ends with a transcriptional regulator. Eight other LSPs containing 4 or more ORFs not present in K-10 were also observed (Table 4). Overall, a total of 70
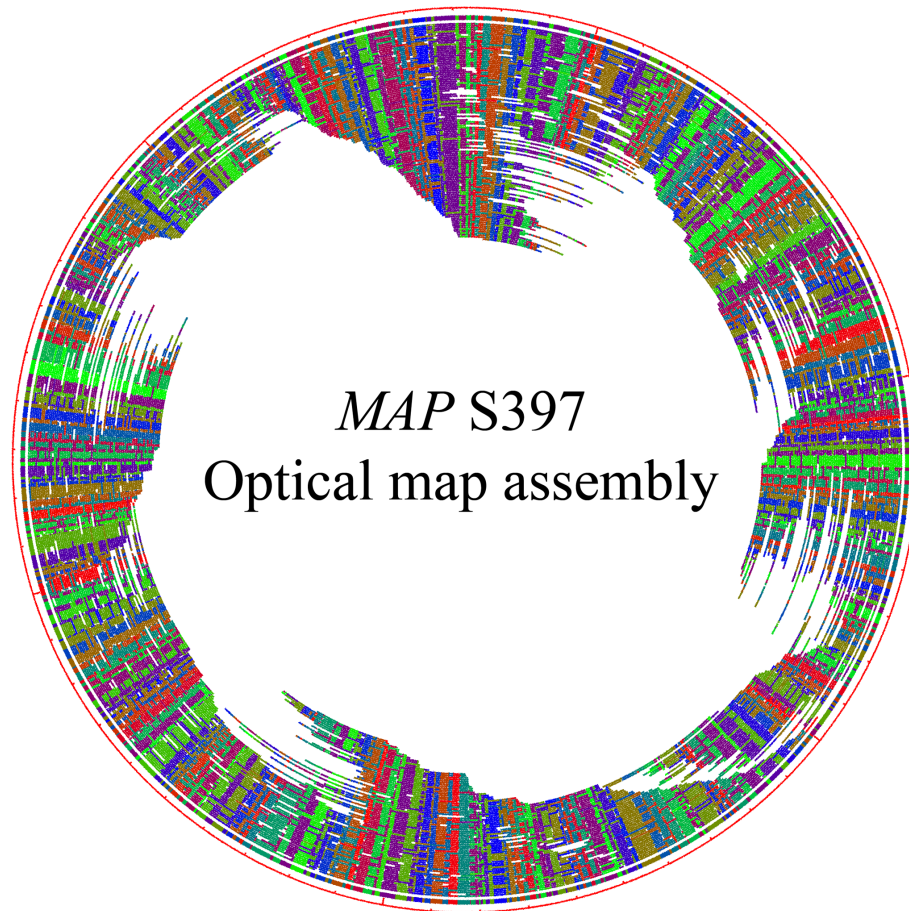
Figure 1 Optical map of the *MAP* S397 genome. A total of 905 optical contigs were assembled into one circular consensus map, which has a 58-fold genome coverage and totaled 4.95 Mb. Optical contigs are represented by arcs of various lengths. Each arc is intersected by radiating lines that represent *Bsi*WI cutting sites, and arbitrary colors represent homologous overlapping fragments.

ORFs were present in *MAP* S397 but absent in the *MAP* K-10 genome (Additional file 5: Table S5).

Several new or only partially described LSPs common to *MAP* S397 and *MAH* 104 strains were also identified. A good example here is the novel LSP found in *MAP* sheep and *MAH* 104 genomes is comprised of 14 ORFs (MAPs_17580 - MAPs_17710), predicted to encode proteins involved in the biosynthesis of glycopeptidolipids [30]. This region in *MAP* S397 revealed the presence of four additional ORFs (*hyp, hlpA, dhgA* and *mtfC*) with homology to glycopeptidolipid biosynthesis genes immediately downstream. The additional 4 ORFs were also not present in the *MAH* 104 sequence. Finally, a putative transcriptional regular labeled as MAPs_44910 is present in *MAP* S397. The protein encoded by this ORF has homology to the GntR-family of transcriptional regulators, which are widely distributed across bacterial species and regulate a variety of cellular processes [31,32].

Genomic deletions

A second subset of sequence polymorphism was represented by 32 ORFs that were present in the *MAP* K-10 genome but absent from the genome of *MAP* S397 (Additional file 6: Figure S1). Several of these deletions have already been described earlier. The deletion encompassing MAP1485c-MAP1491 was previously identified by Marsh and coworkers as S strain deletion #1 in an Australian *MAP* sheep isolate [33] and by Semret and coworkers as LSP^{A}20 [29]. An additional larger deletion in the *MAP* S397 included the cluster of ORFs between MAP1728c and MAP1744. This deletion was partially identified by Marsh and coworkers as RDA3 [34], and later fully described as S deletion #2 [33].

A novel deletion comprising the ORFs MAP1432-MAP1438c (partial) was identified in the current study as absent from *MAP* S397. This deletion, termed sΔ-1, was originally discovered by comparative genomic
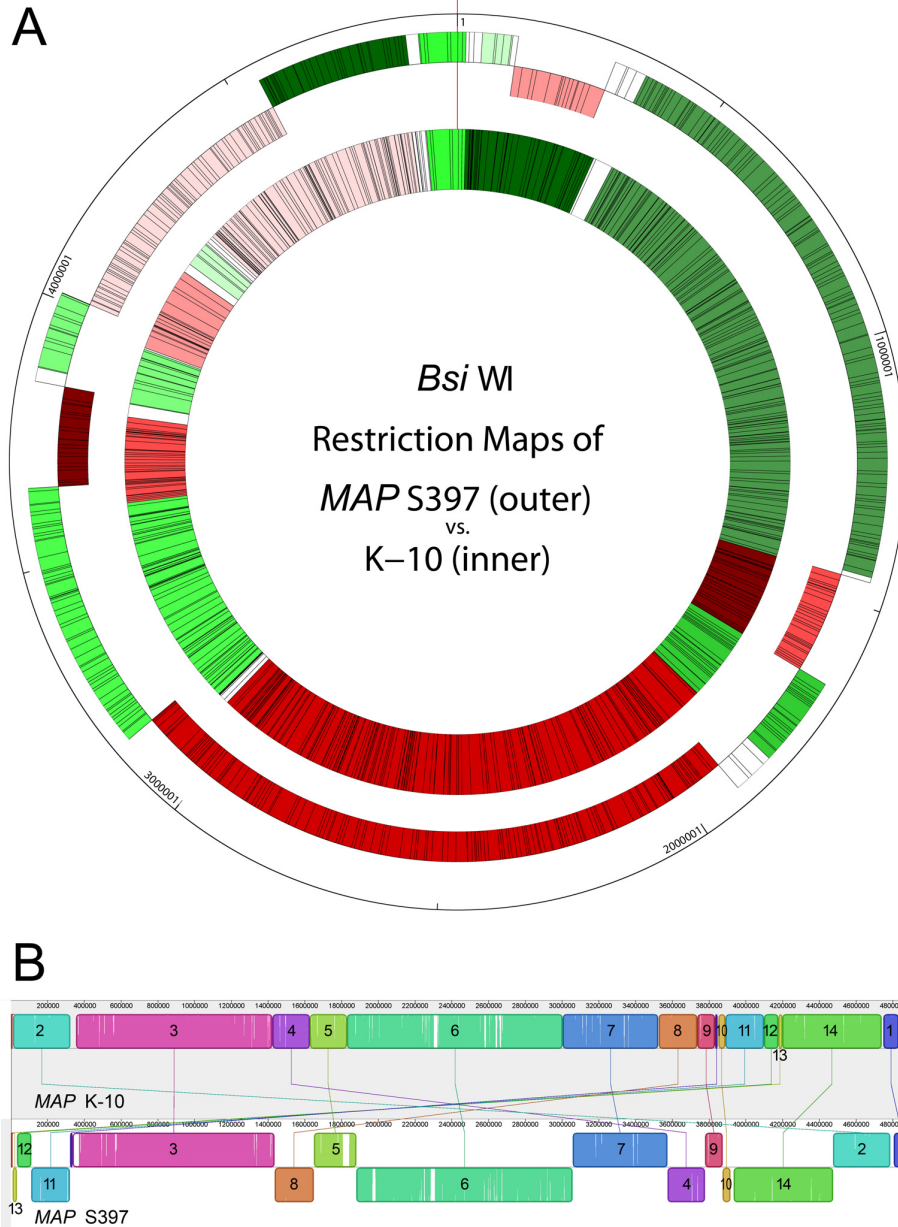
Figure 2 Comparative genome analysis of K-10 and S397 *MAP* strains. **(A)** Comparison of the *Bsi*WI restriction maps between K-10 (inner circle) and S397 (outer circle). Each box represents a restriction fragment. Green boxes are regions in the same direction and red boxes are regions that are inverted between the two genomes. White boxes are fragments that are not aligned. The red thin line at 12 o'clock is the locus of the gene *dnaA*. **(B)** Mauve alignment of all 184 scaffolds of S397 (bottom) with the complete genome of K-10 (top). The colored boxes represent homologous regions present in each genome, which are also connected by lines. Blocks below the centerline of the S397 genome indicate regions with inverse orientation. Regions outside the blocks lack homology between the genomes. Within each block there is a similarity profile of the DNA sequences and the white areas indicate sequences specific to a genome. The scale is in base pairs.

analysis and subsequently confirmed by PCR analysis. This gene cluster is predicted to encode four energy metabolism enzymes as well as a lipase (MAP1438c). MAP1432 encodes a hypothetical protein with homology to the REP13E12, a family of repetitive elements that were originally described in *M. tuberculosis* and have been shown to be targets of phage integration [35]. There is a homolog to MAP1434 that is present in S397 (MAPs_13210). The region around MAPs_13210 is not near the end of a contig and is nearly identical to an

Table 3 Boundaries and flanking ORFs of aligned segments between *MAP* K-10 and *MAP* S397 genomes

|	K-10 coordinates by Wynne [28]		K-10 coordinates by Li [19]		Flanking ORF		Approx. size (Kb)	Alignment between K-10 and S397
ID	Left end	Right end	Left end	Right end	Left end	Right end		
1	4751793	11164	4748984	11164	IS*900*	IS_MAP03 and IS*1311*	92	Forward
2	13889	325818	4197078	3885146	IS_MAP03 and IS*1311*	No known IS element	312	Forward
3	357239	1424396	3853725	2786561	IS_MAP02	No known IS element	1067	Forward
4	1424397	1627687	2786560	2583268	No known IS element	ISmav2	203	Inverted
5	1629051	1828408	2581904	2382587	ISmav2	IS*1311*	199	Forward
6	1829822	3004096	2381173	1206891	IS*1311*	IS*1311*	1174	Inverted
7	3005512	3524059	1205475	686927	IS*1311*	ISmav2	519	Forward
8	3525433	3737882	685553	473103	ISmav2	No known IS element	212	Inverted
9	3737883	3833853	473102	377132	No known IS element	No known IS element	96	Forward
10	3849713	3890501	361272	320484	IS_MAP02	No known IS element	41	Inverted
11	3890502	4098807	320483	112177	No known IS element	IS*1311*	208	Inverted
12	4100223	4176993	110761	33991	IS*1311*	IS*1311*	77	Forward
13	4177670	4199835	33287	11158	IS*1311*	IS*1311*	22	Inverted
14	4201310	unknown[1]	4198493	unknown[1]	IS*1311*	N/A	562	Inverted
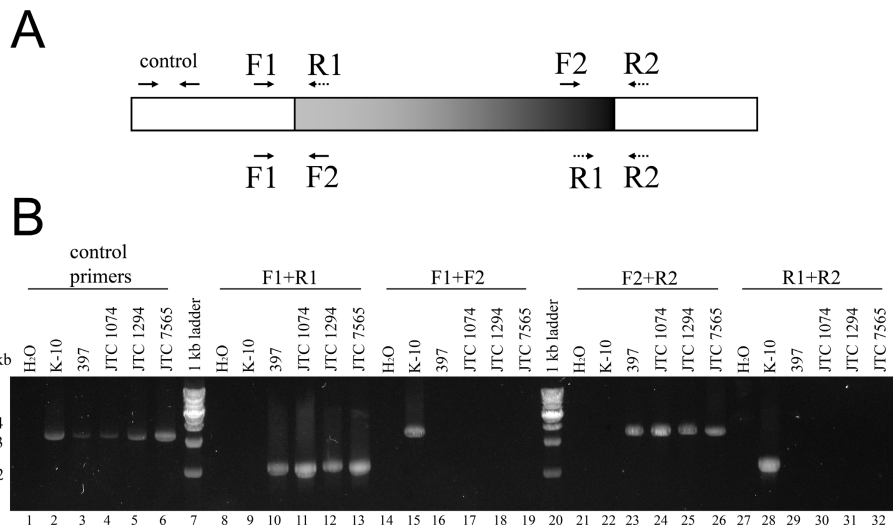
Figure 3 PCR analysis of a 648-kb inverted region [20] between genomes of *MAP* bovine type strain (K-10), and four ovine strains (S397, JTC1074, 1294 and 7565). **(A)** A diagram showing the inverted region (gray-to-black gradient box) and location of primers used in the PCR analysis. All primers were designed according to the published *MAP* K-10 genome sequence [19]. **(B)** PCR results on an ethidium bromide stained agarose gel. Lanes loaded with PCR products amplified with original primer pairs F1 + R1 (lane 8-13) or F2 + R2 (lane 14-19) show no PCR products from the cattle strain (lane 9 and 22) but a 2.1-kb and 3.6-kb band from the sheep strains (lane 10-13 and 23-26), respectively. Lanes with products amplified with switched primer pairs F1 + F2 (lane 14-19) or R1 + R2 (lane 27-32) show a 3.6-kb and 2.3-kb fragment from the cattle strain (lane 15 and 28) but no product from the sheep strains (lane 16-19 and 29-32), respectively. The opposite PCR amplification pattern between the cattle and sheep strains confirmed that this segment is inverted between these 2 genomes.

Table 4 Large sequences present in the three sheep strain genomes but absent in *MAP* K-10.

LSP		Location				Gene Content
Name	Synonym	Start	Stop	Size	Contig	ORFs
LSP[s]1	LSP[A]4-II (RD20)	34,308	43,318	9.01	00036c	MAPs_15940-16060
LSP[s]2	LSP[A]18 (RD10)	116,723	123,369	6.64	00127c	MAPs_46190-46270
LSP[s]3		15,581	19,356	3.78	00035c	MAPs_14620-14660
LSP[s]4		124,533	128,162	3.63	00127c	MAPs_46290-46320
LSP[s]5		87,461	90,928	3.47	00126c	MAPs_17580-17610
LSP[s]6		4,736	7,731	2.99	00161	MAPs_40470-40500
LSP[s]7		2	2,889	2.88	00133c	MAPs_17640-17670
LSP[s]8		47,323	49,712	2.39	00199c	MAPs_02730-02760
LSP[s]9		70,659	72,500	1.84	00033	MAPs_23120-23150
LSP[s]10		5,879	7,457	1.58	00100	MAPs_42460-42490

LSP Large sequence polymorphism as identified before [29]
Size is in kilobases

inverted stretch in K-10, thus leading to the conclusion that MAPs_13210 is only a homolog of MAP1434, but that the gene itself is not present in the S397 genome.

Interestingly, MAP2656 was initially identified as absent via microarray analysis [5] but sequencing of *MAP* S397 identified a homologue with 100% identity (MAPs_10401 & MAPs_10402). Likewise, MAP2325 was identified as being absent from Australian sheep isolates of *MAP* [33]. This ORF was not identified as missing from *MAP* S397 as sequencing confirmed the presence of an ORF (MAPs_34380) with 100% identity to MAP2325. These discrepancies may represent a geographic difference between *MAP* isolates recovered from sheep in Australia and the United States or it may be an error from the microarray experiment. These were the only observed differences between the microarray and sequence data. Overall, genomic alignments indicated the presence of a significant number of insertions and deletions between ovine and bovine strains of *MAP* that are suggested to be associated with their respective host.

Evolutionary analysis of the *MAP* S397 genome
Genomic insertions and deletions have been previously used to determine evolutionary relationships among MAC strains [36]. With the genome sequence of these ovine isolates of *MAP*, we can now add comprehensive SNP and inversion data to strengthen evolutionary hypotheses. Earlier genotyping of the *MAP* S397 utilizing SNP of *recF*, *gyrA* and *gyrB* genes indicated that this strain belong to the *MAP* type III, a sublineage of the *MAP-S* cluster of isolates [37]. To examine the evolutionary history of *MAP*, we analyzed the genome sequence of S397 compared to other clinical isolates circulating in sheep as well as the standard cattle strain, K-10. Our first level of analysis included the alignment of the S397 genome to that of the JTC1074 and JTC7565.

This alignment resulted in identical genome organization of all three ovine isolates, as expected. Additionally, we examined the relationship among S397 (ovine origin) with both K-10 (bovine origin) and *MAH* 104 (human origin). Such analysis identified several events of inversions and potential insertions/deletions between genomes belonging to the ovine isolates and other isolates of bovine and human origins (Figure 4). The optical map of S397 confirmed these inversions as well. Moreover, when the draft genome sequence of *M. intracellulare* was added to the comparison, the whole contig00148 (accession number GenBank: ABIN01000141) aligns to the region spanning the right breakpoint (Figure 4) of *MAH* 104 and *MAP* ovine strains, an indication of a conserved genome synteny among *M. intracellulare*, *MAH* and *MAP* sheep strains, but distinct from *MAP* bovine strains.

In the second level of analysis on the nucleotide level, a core of 42 single nucleotide polymorphisms (SNPs) were present in both JTC isolates compared to S397. In addition, a very small number of unique SNPs in JTC1074 (N = 22) and JTC7565 (N = 11) were not present in any other genome in this study. Collectively, this small level of polymorphism indicates the clonal nature of ovine isolates, which contrasts sharply with the 4,438 SNPs between the ovine S397 and the bovine K-10 strains (Figure 5A). Additionally, when analyzing genome-wide SNPs, it appears that *MAP* S397 and K-10 split off recently from the *hominissuis* progenitor strain (Figure 5B). A similar result is obtained when SNPs are restricted to coding sequences (Figure 5C).

Discussion
Comparative genomic hybridizations using DNA microarrays have revealed large sequence polymorphisms (LSPs) between *MAP-S* and *MAP-C* strains [36,41]. Two large deletions of an Australian sheep isolate were found
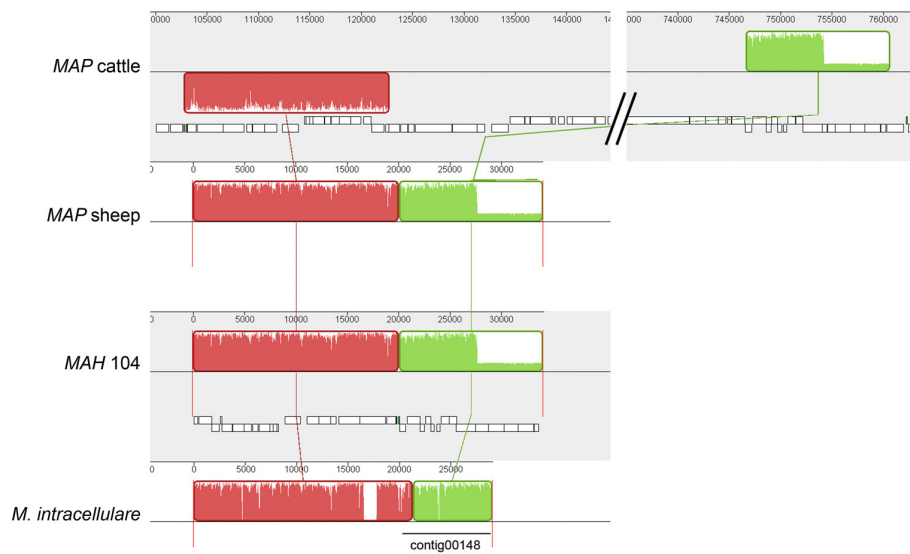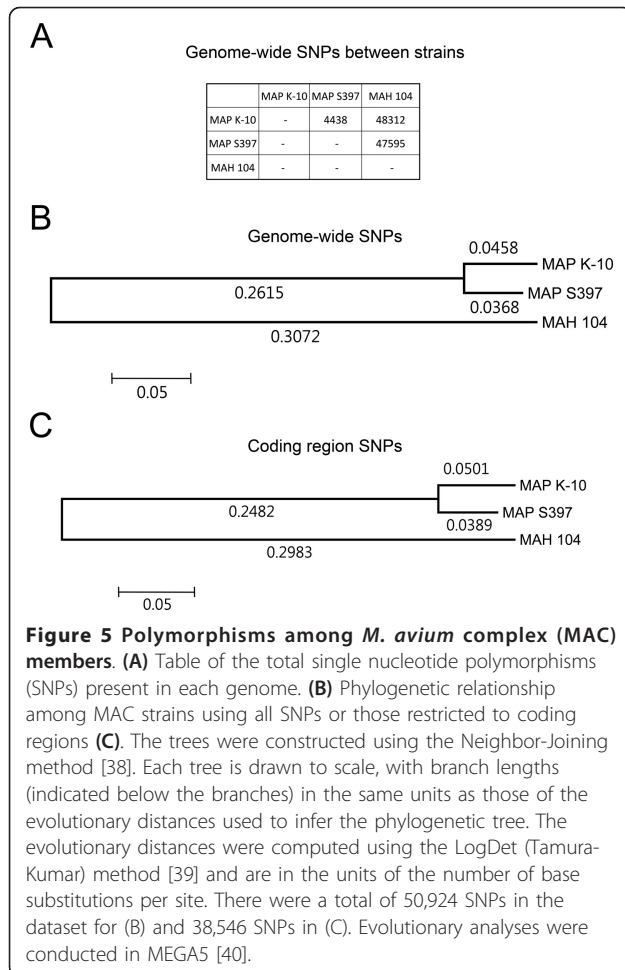
Figure 4 Genomic alignment of inversion breakpoints among members of the *M. avium* complex. Regions spanning the right breakpoint are depicted. The junction between the red and green boxes shown in the *MAP* sheep panel represents the breakpoint. Note that the breakpoint is within contig00148 of *M. intracellulare*. The alignment shows that the genome synteny among *MAP* sheep (S397), *MAH* 104 and *M. intracellulare* is conserved.

by genomic hybridization to the *MAP* K-10 array [33]. One deletion encompassed 8 ORFs (MAP1485c-MAP1491) and a second deletion encompassed 17 ORFs extending from MAP1728c to MAP1744. These deletions relative to the bovine strains were later observed in U.S. ovine *MAP* isolates [5,13]. Construction of a *MAP* array containing *MAH* sequences revealed LSPs in the ovine strains that were missing in the bovine K-10 strain [5,42]. These documented differences formed the basis for whole-genome sequencing of a sheep isolate to enable comprehensive description of all genetic differences from *MAP-S* and *MAP-C* strains. We took advantage of next-generation sequencing technology combined with optical mapping [20] to decipher the complete genome of *MAP* isolates from sheep flocks raised in the USA. Our analysis confirmed earlier polymorphisms among *MAP*-S and *MAP*-C strains and revealed novel regions of difference. Surprisingly, both genome sequencing and optical mapping showed remarkable differences between *MAP*-S and *MAP*-C strains despite the overall similarity in the clinical signs of Johne's disease in sheep and cattle. Recently, a study using a large number of *MAP* isolates provided an example of such a genomic polymorphism including 2 large regions of duplication, termed vGI-17 (containing 63 ORFs) and vGI-18 (containing 109 ORFs), observed in most *MAP*-C strains but not *MAP*-S isolates [21]. Both of these duplications were also missing in our sequenced *MAP*-S genomes as determined by PCR

amplification using outward facing primers reported by Wynne et al. (data not shown).

There are 70 genes present in all three ovine isolates that are absent from the K-10 strain, an indication for *MAP* adaptation to specific hosts (in this case sheep). Analysis of additional ovine and bovine isolates is needed to strengthen any linkage between these genes with host association. Within this subset, we identified a surprising number of genes annotated as hypothetical proteins (N = 30). Six transcriptional regulators were also present among these genes with the remaining genes showing weak homology to sequences in the GenBank database. We hypothesize that these genes could be responsible for the observed phenotypic differences between ovine and bovine strains and warrant future studies to address this hypothesis.

Based on extensive genomic rearrangements between *MAP* bovine and ovine strains, we were able to provide a possible evolutionary scenario for members of the MAC group. A genomic region spanning the inversion of *MAP* bovine strains, *MAP* ovine strains and *MAH* 104 are shown in Figure 4. To diverge into these three subspecies, the common ancestor appears to have undergone two independent genomic inversion events (Figure 6A). Specifically, it would take one inversion event to diverge between *MAH* 104 and *MAP* sheep strains followed by a second inversion event between *MAP* sheep strains and the *MAP* cattle strain (Figure 6A). Therefore, assuming that one strain diverges into

Figure 5 Polymorphisms among *M. avium* complex (MAC) members. **(A)** Table of the total single nucleotide polymorphisms (SNPs) present in each genome. **(B)** Phylogenetic relationship among MAC strains using all SNPs or those restricted to coding regions **(C)**. The trees were constructed using the Neighbor-Joining method [38]. Each tree is drawn to scale, with branch lengths (indicated below the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the LogDet (Tamura-Kumar) method [39] and are in the units of the number of base substitutions per site. There were a total of 50,924 SNPs in the dataset for (B) and 38,546 SNPs in (C). Evolutionary analyses were conducted in MEGA5 [40].

Figure 6 An evolutionary scenario for members of the *M. avium* complex. **(A)** Depicted is a two-step inversion process as one possible scenario explaining how *MAH* evolved into *MAP* K-10 through *MAP* S397. To examine evolutionary relationship among the MAC, genome alignment around the inversion segment is depicted with Mauve version 2.3.1 [44]. Divergence between *MAH* 104 and *MAP* sheep strains or between *MAP* sheep strains and cattle strains would take only one inversion, whereas divergence between *MAH* 104 and *MAP* cattle strains would need two independent inversion events. **(B)** Our proposed model for evolution of the *M. avium* complex.

another strain by taking the shortest evolutionary path, it would be least likely that *MAH* directly evolved from *MAP* cattle strains or vice versa. This strongly suggests that *MAP* sheep strains are the intermediate taxon of the three. Data from Behr and coworkers suggest *MAH* 104 is the ancestor strain [36]. Moreover, when the genome of *M. intracellulare* is added to the comparison, the genome synteny was conserved among *M. intracellulare*, *MAH* and *MAP* sheep strains, but not in *MAP* cattle strains. Thus, it is possible that the common ancestor of the MAC must resemble either *MAH* 104 or *M. intracellulare*, and *MAP* bovine strains are the latest diverged strains among them with *MAP* S397 as an intermediary strain (Figure 6B). This model partially agrees with a hypothesis that suggests *MAH* differentiated into two lineages, *MAP* ovine and bovine strains, by delineating chronological genomic insertion/deletion events without considering other genomic rearrangement events [36]. Of the 70 genes in S397 that are absent in K-10, 57 are present in *MAH* 104 and only 13 are absent from *MAH* 104. Further genotyping of the S397 clustered this isolate with the group of *MAP*-S
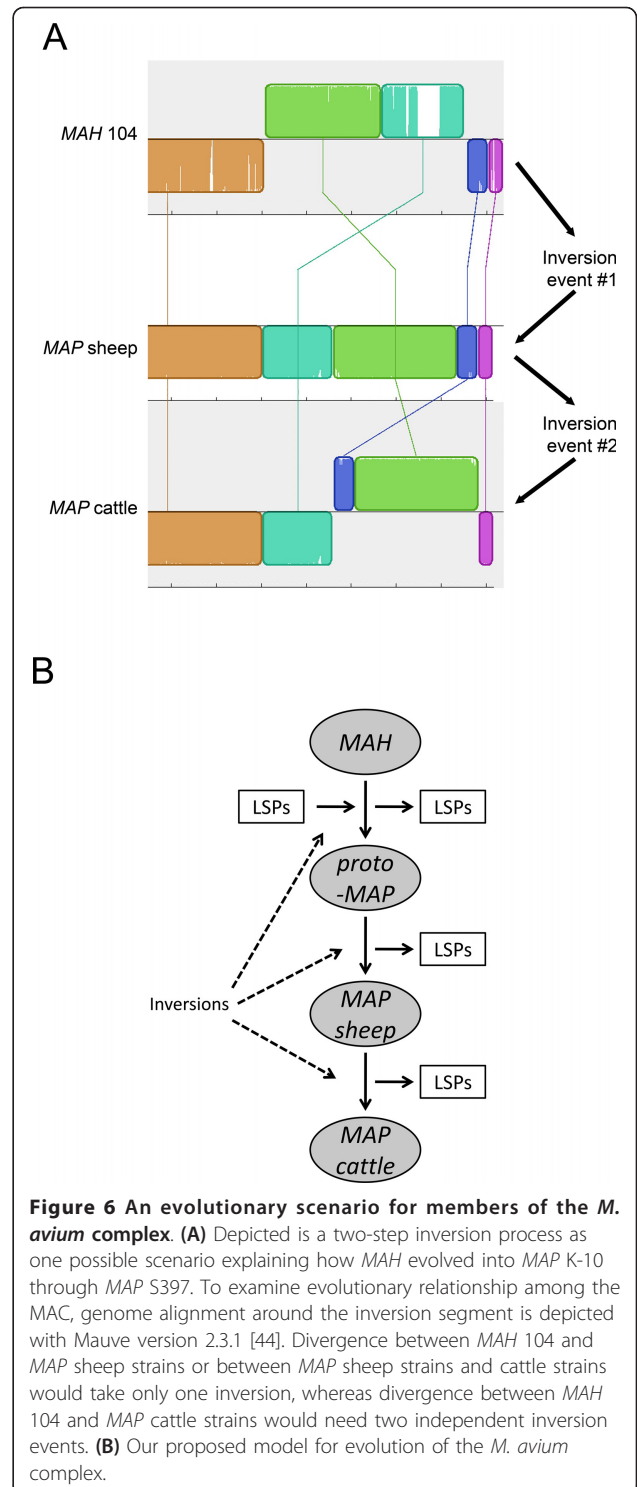
type III [37], a sub-lineage of the sheep strains. However, we prefer to maintain the *MAP*-S designation since the type III genotype was based on 3 SNPs present in a subgroup of sheep isolates with no distinctive clinical or pathological features. Finally, a recent study analyzing the sequence polymorphisms of IS*1311* among the

MAC also supports the hypothesis that *MAP* ovine strains are the intermediary taxa between *MAH* and *MAP* bovine strains [43].

Conclusions

Genome sequencing of *MAP-S* strains have revealed extensive genome inversions and previously characterized deletions when compared to the K-10 strain. Furthermore, there appears to be a high degree of homology within US *MAP-S* strains as suggested by the remarkably low number of SNPs present in the three isolates sequenced. Evolutionary analysis based on whole genome sequencing suggests MAH is the progenitor strain, followed by MAP-S, followed by MAP-C strains.

Overall, Next-generation sequencing combined with optical mapping provided us with a high resolution tool to decipher the evolution of important pathogenic mycobacteria. Comparative sequence analysis of the *MAP* isolates from sheep has improved our understanding of the evolutionary history of members of MAC and provided the foundation for novel insights into the pathogenesis of this important pathogen. Similar approaches can be used to examine other closely related pathogens.

Methods

MAP ovine isolates

Isolates were cultured in Middlebrook 7H9 broth (BD Biosciences, San Jose, CA) media supplemented with 10% OADC (2% glucose, 5% bovine serum albumin factor V, and 0.85% NaCl), 0.05% Tween 80 and 2 μg/ml of Mycobactin J at 37°C [45]. The *MAP* ovine S397 strain was obtained from a Suffolk breed in Iowa. It was isolated from the distal ileum at necropsy in 2004. The other 2 sheep isolates of *MAP* (JTC1074 and JTC7565) were isolated from the intestine of infected sheep in Texas and obtained from the Johne's Testing Center at the University of Wisconsin-Madison. All isolates were genotyped using the IS*1311* restriction endonuclease, which yielded the 2-band pattern typical of ovine strains [6].

Genome sequencing

Genomic DNA was extracted as described in detail previously [3,46]. For the S397 strain, the DNA (1-5 μg) was sequenced using Roche 454 pyrosequencing (GS20 and FLX) at the National Animal Disease Center. A whole-genomic shotgun sequencing library was prepared according to Roche protocols. The library was used with the appropriate emulsion based PCR kits to produce sufficient beads for sequencing using the Roche Standard Chemistry GS-LR 70 sequencing kit. For the JTC1074 and JTC7565, the purified genomic DNA (~5 μg) of each strain was sent to Genomic Resource Center at the University of Maryland for Illumina whole

genome sequencing (Multiplexing Sample Preparation oligonucleotide Kit) as outline before [47]. The adapters and indexing oligonucleotides were purchased from Illumina (5 Paired End Cluster Generation Kits-v4). The CLC Genomic Workbench software (version 4.0.3) was used to perform reference and de novo assembly on all sequenced genomes.

Genome annotation

The S397 sequence was annotated using the Integrated Microbial Genomes Expert Review (IMG-ER) pipeline [48]. The sequences of the JTC isolates were annotated based on S397. Genes were each designated by the locus tag "MAPs" to distinguish it as a *MAP* sheep strain gene. This locus tag is followed by a five digit unique identifier, which incrementally increases by ten (i.e. MAPs_45660... MAPs_45670... MAPs_45680...). With this numbering configuration, additional genes can easily be added as they are discovered or when remaining gaps are closed.

Genome comparison

The genome data for *MAP* K-10 (accession no. GenBank: NC_002944.2) and *M. avium* subsp. *hominissuis* (*MAH*) strain 104 (GenBank: NC_008595.1) were used in alignments in the Artemis and Artemis Comparison Tool (ACT) programs or Mauve 2.3.1 [49]. BLASTP analysis was used for similarity searches and protein sequence analysis. In addition, Mauve algorithm was used to align two or more genomes [50]. For detecting single nucleotide polymorphisms (SNPs) among sheep isolates, the CLC Genomic Workbench was used. The coverage range setting for each strain was at 10-55 reads, and the frequency of the mutation was at least in 50% of the reads.

Optical mapping

Shotgun optical mapping, as previously described [20,51-55], was used to construct a physical restriction map for the S397 genome. Genomic DNA, in agarose inserts [56], was electroeluted into a solution containing a lambda DNA sizing standard (30 pg/μl), and then were mounted on cleaned, derivitized glass surfaces using a microfluidic device [57] followed by polymerization of a thin layer of polyacrylamide (3.3% containing 0.02% Triton X-100). Mounted DNA was digested with 20 units of *Bsi*WI (NEB, Ipswich, MA) for 1 to 2 hrs at 37°C. Fluorochrome-stained DNA fragments were imaged by fluorescence microscopy with a 63 × objective lens (Carl Zeiss, Thornwood, NY) and a high-resolution digital camera (Princeton Instruments, Trenton, NJ). Images were acquired and processed using "ChannelCollect" and "Pathfinder" -custom software [57] that converts captured images into map data sets. Bayesian inference and an efficient dynamic programming algorithm were also being used to fine-tune the parameters including standard

deviation, digestion rate, false cut, and false match probability etc. [54,58,59]. The final circular optical map contig was built using an iterative assembly process [60] including rounds of pair-wise alignment (single molecule maps vs. seed maps; provisional assemblies) and assembly [52,54]. Due to the high G + C content of *MAP*, which skews fragment sizing by integrated fluorescence intensity measurement, the final maps were globally scaled (0.95) to correct this problem [20,61]. A laboratory software implementation of an optical map alignment algorithm [62] was used to align between optical fragments generated from *MAP* S397 and the *in silico* restriction maps of *MAP* K-10, which provided a whole-genome rearrangement comparison between the two genomes. This restriction framework was used to generate a temporary rearranged genome as the reference sequence to guide the assembly of *MAP* S397 *de novo* contigs with the function "move contigs" in Mauve 2.3.1 [49].

PCR amplification of inversion breakpoints and deletions

PCR reactions were performed in 25-μl reaction mixture containing 1 M betaine (Sigma-Aldrich, St. Louis, MO), 50 mM potassium glutamate (Sigma-Aldrich), 10 mM Tris-HCl pH 8.8, 0.1% Triton X-100, 2 mM magnesium chloride, 0.2 mM dNTPs, 0.5 μM each primer, 0.5 U of GoTaq® Flexi DNA Polymerase (Promega, Madison, WI) and 25 ng of genomic DNA. The amplification thermocycle started with an initial step of 94°C for 5 minutes followed by 5 cycles of 94°C for 30 s, 62°C for 30 s with 1°C decrease for each cycle and 72°C for 3.5 min, and followed by 30 cycles of 94°C for 30 s, 57°C for 30 s and 72°C for 3.5 min. PCR primers used for examining the breakpoints included control F: AAGCATCACCTGCATGAGC, control R: CGGGAATTTATCCGTTTCAG, F1: GGGATCGATCTTGACCACAT, R1: GTGCCTGGACTCGATTTTGT, F2: AAGAGGTCGGAGGTTCGAGT and R2: CGGTGAGAGATTTCGTCACA. Primers used to demonstrate the S397 sΔ-1 deletion included F18: CGTCTTCCCCGTCGTCGTTC, B24: CGATGAGAGTCCGTGCGTGG, F15: CGGCGGGCGGTCAGGGTTTG, B17: GCAGGTTGGGGGTTCGGCTTG, F7: GGTGGTCGGCGTCCTCGTAG, B9: CGTCGTCACAGCGAAAACGG, F3: CCACCCGCCTCACACCACTC, B4: AGGACGCCGACCACCAAACG. Conditions for the amplifications are essentially as described immediately above except that Advantage GC Genomic LA PCR Polymerase kit (Clontech) was used for each reaction.

Nucleotide sequence accession number

This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AFIF00000000.

Additional material

Additional file 1: Table S1. PSORT analysis of MAP S397 genes to determine their localization.

Additional file 2: Table S2. A list of 184 genomic DNA scaffolds of MAP S397 genome.

Additional file 3: Table S3. Synteny of annotated genes between MAP K-10 and MAP S397 genomes.

Additional file 4: Table S4. MAP S397 genes that are absent in MAP K-10.

Additional file 5: Table S5. MAP K-10 genes that are absent in MAP S397. All supplemental tables are in Excel format.

Additional file 6: Figure S1. An overview alignment of MAP S397 scaffold assembly and BsiWI restriction fragments.

Author details
[1]National Animal Disease Center, USDA-Agricultural Research Service, Ames, Iowa, USA. [2]The Laboratory of Bacterial Genomics, Department of Pathobiological Sciences, University of Wisconsin-Madison, Madison, Wisconsin, USA. [3]Laboratory for Molecular and Computational Genomics, Department of Chemistry and Laboratory of Genetics, UW Biotechnology Center, University of Wisconsin-Madison, Madison, Wisconsin, USA. [4]Department of Veterinary Population Medicine and Department of Veterinary and Biomedical Sciences, University of Minnesota, St. Paul, Minnesota, USA. [5]Department of Veterinary and Biomedical Sciences and Huck Institutes of the Life Sciences, Penn State University, University Park, Pennsylvania, USA. [6]Department of Food Hygiene, Cairo University, Cairo, Egypt.

Authors' contributions
JPB, MLP and DPA sequenced the S397 genome. JPB and DOB annotated the S397 genome. JPB and CW integrated and analyzed data from the S397 contigs and optical mapping, performed genomic comparison, PCR confirmation as well as SNP, inversion and evolutionary analyses. CH, CW and AMT sequenced and analyzed genomes of isolates JTC1074 and JTC7565. SZ and DCS collected, assembled and aligned S397 optical map. JPB, CW and AMT wrote the manuscript. JPB, SS, VK and AMT assisted with the analysis. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

References
1. Wu CW, Livesey M, Schmoller SK, Manning EJ, Steinberg H, Davis WC, *et al*: Invasion and persistence of *Mycobacterium avium* subsp. *paratuberculosis* during early stages of Johne's disease in calves. *Infect Immun* 2007, **75**:2110-2119.
2. Bermudez LE, Petrofsky M, Sommer S, Barletta RG: Peyer's patch-deficient mice demonstrate that *Mycobacterium avium* subsp. *paratuberculosis* translocates across the mucosal barrier via both M cells and enterocytes but has inefficient dissemination. *Infect Immun* 2010, **78**:3570-3577.
3. Wu CW, Glasner J, Collins M, Naser S, Talaat AM: Whole-genome plasticity among *Mycobacterium avium* subspecies: insights from comparative genomic hybridizations. *J Bacteriol* 2006, **188**:711-723.

4. Taylor AW: **Varieties of *Mycobacterium johnei* isolated from sheep.** *J Pathol Bacteriol* 1951, **63**:333-336.

5. Paustian ML, Zhu X, Sreevatsan S, Robbe-Austerman S, Kapur V, Bannantine JP: **Comparative genomic analysis of *Mycobacterium avium* subspecies obtained from multiple host species.** *BMC Genomics* 2008, **9**:135-149.

6. Marsh I, Whittington R, Cousins D: **PCR-restriction endonuclease analysis for identification and strain typing of Mycobacterium avium subsp. paratuberculosis and Mycobacterium avium subsp. avium based on polymorphisms in IS1311.** *Mol Cell Probes* 1999, **13**:115-126.

7. Dohmann K, Strommenger B, Stevenson K, de JL, Stratmann J, Kapur V, *et al*: **Characterization of genetic differences between *Mycobacterium avium* subsp. *paratuberculosis* type I and type II isolates.** *J Clin Microbiol* 2003, **41**:5215-5223.

8. Amonsin A, Li LL, Zhang Q, Bannantine JP, Motiwala AS, Sreevatsan S, *et al*: **Multilocus short sequence repeat sequencing approach for differentiating among *Mycobacterium avium* subsp. *paratuberculosis* strains.** *J Clin Microbiol* 2004, **42**:1694-1702.

9. Sevilla I, Li L, Amonsin A, Garrido JM, Geijo MV, Kapur V, *et al*: **Comparative analysis of *Mycobacterium avium* subsp. *paratuberculosis* isolates from cattle, sheep and goats by short sequence repeat and pulsed-field gel electrophoresis typing.** *BMC Microbiol* 2008, **8**:204.

10. Thibault VC, Grayon M, Boschiroli ML, Willery E, Allix-Beguec C, Stevenson K, *et al*: **Combined multilocus short sequence repeat and mycobacterial interspersed repetitive unit- variable-number tandem repeat typing of *Mycobacterium avium* subsp. *paratuberculosis* isolates.** *J Clin Microbiol* 2008, **46**:4091-4094.

11. Turenne CY, Semret M, Cousins DV, Collins DM, Behr MA: **Sequencing of *hsp6* distinguishes among subsets of the *Mycobacterium aviu* complex.** *J Clin Microbiol* 2006, **44**:433-440.

12. Castellanos E, Juan Ld, Domínguez L, Aranaz A: **Progress in molecular typing of Mycobacterium avium subspecies paratuberculosis.** *Res Vet Sci* 2011, doi:10.1016/j.rvsc.2011.05.017.

13. Motiwala AS, Janagama HK, Paustian ML, Zhu X, Bannantine JP, Kapur V, *et al*: **Comparative transcriptional analysis of human macrophages exposed to animal and human isolates of *Mycobacterium avius* subspecies *paratuberculosis* with diverse genotypes.** *Infect Immun* 2006, **74**:6046-6056.

14. Stevenson K, Hughes VM, de JL, Inglis NF, Wright F, Sharp JM: **Molecular characterization of pigmented and nonpigmented isolates of Mycobacterium avium subsp. paratuberculosis.** *J Clin Microbiol* 2002, **40**:1798-1804.

15. Whittington RJ, Marsh IB, Saunders V, Grant IR, Juste R, Sevilla IA, *et al*: **Culture Phenotypes of Genomically and Geographically Diverse *Mycobacterium avium* subsp. *paratuberculosis* Isolates from Different Hosts.** *J Clin Microbiol* 2011, **49**:1822-1830.

16. Gumber S, Taylor DL, Marsh IB, Whittington RJ: **Growth pattern and partial proteome of Mycobacterium avium subsp. paratuberculosis during the stress response to hypoxia and nutrient starvation.** *Vet Microbiol* 2009, **133**:344-357.

17. Gumber S, Whittington RJ: **Analysis of the growth pattern, survival and proteome of Mycobacterium avium subsp. paratuberculosis following exposure to heat.** *Vet Microbiol* 2009, **136**:82-90.

18. Janagama HK, Senthilkumar TM, Bannantine JP, Rodriguez GM, Smith I, Paustian ML, *et al*: **Identification and functional characterization of the iron-dependent regulator (*IdeR*) of *Mycobacterium avium* subsp. *paratuberculosis*.** *Microbiology* 2009, **155**:3683-3690.

19. Li L, Bannantine JP, Zhang Q, Amonsin A, May BJ, Alt D, *et al*: **The complete genome sequence of *Mycobacterium avium* subspecies *paratuberculosis*.** *Proc Natl Acad Sci USA* 2005, **102**:12344-12349.

20. Wu CW, Schramm TM, Zhou S, Schwartz DC, Talaat AM: **Optical mapping of the *Mycobacterium avium* subspecies *paratuberculosis* genome.** *BMC Genomics* 2009, **10**:25.

21. Wynne JW, Bull TJ, Seemann T, Bulach DM, Wagner J, Kirkwood CD, *et al*: **Exploring the zoonotic potential of *Mycobacterium avium* subspecies *paratuberculosis* through comparative genomics.** *PLoS One* 2011, **6**: e22171.

22. Bannantine JP, Zhang Q, Li LL, Kapur V: **Genomic homogeneity between Mycobacterium avium subsp. avium and Mycobacterium avium subsp. paratuberculosis belies their divergent growth rates.** *BMC Microbiol* 2003, **3**:10.

23. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, *et al*: **PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.** *Bioinformatics* 2010, **26**:1608-1615.

24. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, *et al*: **Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination.** *Proc Natl Acad Sci USA* 1997, **94**:9869-9874.

25. Nodieva A, Jansone I, Broka L, Pole I, Skenders G, Baumanis V: **Recent nosocomial transmission and genotypes of multidrug-resistant *Mycobacterium tuberculosis*.** *Int J Tuberc Lung Dis* 2010, **14**:427-433.

26. Collins DM, Gabric DM, de Lisle GW: **Identification of a repetitive DNA sequence specific to *Mycobacterium paratuberculosis*.** *FEMS Microbiol Lett* 1989, **60**:175-178.

27. Green EP, Tizard ML, Moss MT, Thompson J, Winterbourne DJ, McFadden JJ, *et al*: **Sequence and characteristics of IS*900*, an insertion element identified in a human Crohn's disease isolate of *Mycobacterium paratuberculosis*.** *Nucleic Acids Res* 1989, **17**:9063-9073.

28. Wynne JW, Seemann T, Bulach D, Coutts SA, Talaat AM, Michalski WP: **Re-sequencing the *Mycobacterium avium* subsp. *paratuberculosis* K10 genome: improved annotation and revised genome sequence.** *J Bacteriol* 2010, **192**:6319-6320.

29. Semret M, Turenne CY, de HP, Collins DM, Behr MA: **Differentiating host-associated variants of *Mycobacterium avium* by PCR for detection of large sequence polymorphisms.** *J Clin Microbiol* 2006, **44**:881-887.

30. Eckstein TM, Belisle JT, Inamine JM: **Proposed pathway for the biosynthesis of serovar-specific glycopeptidolipids in *Mycobacterium avium* serovar 2.** *Microbiology* 2003, **149**:2797-2807.

31. Haydon DJ, Guest JR: **A new family of bacterial regulatory proteins.** *FEMS Microbiol Lett* 1991, **63**:291-295.

32. Vindal V, Suma K, Ranjan A: **GntR family of regulators in *Mycobacterium smegmatis*: a sequence and structure based characterization.** *BMC Genomics* 2007, **8**:289.

33. Marsh IB, Bannantine JP, Paustian ML, Tizard ML, Kapur V, Whittington RJ: **Genomic comparison of *Mycobacterium avium* subsp. *paratuberculosis* sheep and cattle strains by microarray hybridization.** *J Bacteriol* 2006, **188**:2290-2293.

34. Marsh IB, Whittington RJ: **Deletion of an *mmp* gene and multiple associated genes from the genome of the S strain of *Mycobacterium avium* subsp. *paratuberculosis* identified by representational difference analysis and *in silic* analysis.** *Mol Cell Probes* 2005, **19**:371-384.

35. Gordon SV, Brosch R, Billault A, Garnier T, Eiglmeier K, Cole ST: **Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays.** *Mol Microbiol* 1999, **32**:643-655.

36. Alexander DC, Turenne CY, Behr MA: **Insertion and deletion events that define the pathogen *Mycobacterium avium* subsp. *paratuberculosis*.** *J Bacteriol* 2009, **191**:1018-1025.

37. Ghosh P, Hsu C-Y, Alyamani E, Shehata MM, Al-Dubaib MA, Al-Naeem A, *et al*: **Genome-wide Analysis of the Emerging Infection with *Mycobacterium avium* subspecies *paratuberculosis* in the Arabian Camels (*Camelus dromedarius*).** *PLoS ONE* 2012.

38. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**:406-425.

39. Tamura K, Kumar S: **Evolutionary distance estimation under heterogeneous substitution pattern among lineages.** *Mol Biol Evol* 2002, **19**:1727-1736.

40. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods.** *Mol Biol Evol* 2011, **28**:2731-2739.

41. Semret M, Alexander DC, Turenne CY, de HP, Overduin P, van Soolingen D, *et al*: **Genomic polymorphisms for *Mycobacterium avium* subsp. *paratuberculosis* diagnostics.** *J Clin Microbiol* 2005, **43**:3704-3712.

42. Castellanos E, Aranaz A, Gould KA, Linedale R, Stevenson K, Alvarez J, *et al*: **Discovery of Stable and Variable Differences in the *Mycobacterium avium* subsp *paratuberculosi* Type I, II, and III Genomes by Pan-Genome Microarray Analysis.** *Appl Environ Microbiol* 2009, **75**:676-686.

43. Sohal JS, Singh SV, Singh PK, Singh AV: **On the evolution of 'Indian Bison type' strains of *Mycobacterium avium* subspecies *paratuberculosis*.** *Microbiol Res* 2010, **165**:163-171.

44. Darling AC, Mau B, Blattner FR, Perna NT: **Mauve: multiple alignment of conserved genomic sequence with rearrangements.** *Genome Res* 2004, **14**:1394-1403.

45. Wu CW, Schmoller SK, Shin SJ, Talaat AM: **Defining the stressome of** *Mycobacterium avium* subsp. *paratuberculosis* **in vitro and in naturally infected cows.** *J Bacteriol* 2007, **189**:7877-7886.

46. Bannantine JP, Baechler E, Zhang Q, Li L, Kapur V: **Genome scale comparison of** *Mycobacterium avium* subsp. *paratuberculosis* **with** *Mycobacterium avium* subsp. *avium* **reveals potential diagnostic sequences.** *J Clin Microbiol* 2002, **40**:1303-1310.

47. Hegedus Z, Zakrzewska A, Agoston VC, Ordas A, Racz P, Mink M, *et al*: **Deep sequencing of the zebrafish transcriptome response to mycobacterium infection.** *Mol Immunol* 2009, **46**:2918-2930.

48. Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC: **IMG ER: a system for microbial genome annotation expert review and curation.** *Bioinformatics* 2009, **25**:2271-2278.

49. Darling AE, Mau B, Perna NT: **progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement.** *PLoS One* 2010, **5**: e11147.

50. Perna NT, Mayhew GF, Posfai G, Elliott S, Donnenberg MS, Kaper JB, *et al*: **Molecular evolution of a pathogenicity island from enterohemorrhagic** *Escherichia coli* **O157:H7.** *Infect Immun* 1998, **66**:3810-3817.

51. Lin J, Qi R, Aston C, Jing J, Anantharaman TS, Mishra B, *et al*: **Whole-genome shotgun optical mapping of** *Deinococcus radiodurans.* *Science* 1999, **285**:1558-1562.

52. Zhou S, Bechner MC, Place M, Churas CP, Pape L, Leong SA, *et al*: **Validation of rice genome sequence by optical mapping.** *BMC Genomics* 2007, **8**:278-295.

53. Zhou S, Deng W, Anantharaman TS, Lim A, Dimalanta ET, Wang J, *et al*: **A whole-genome shotgun optical map of** *Yersinia pestis* **strain KIM.** *Appl Environ Microbiol* 2002, **68**:6321-6331.

54. Zhou S, Kile A, Kvikstad E, Bechner M, Severin J, Forrest D, *et al*: **Shotgun optical mapping of the entire** *Leishmania major* **Friedlin genome.** *Mol Biochem Parasitol* 2004, **138**:97-106.

55. Zhou S, Wei F, Nguyen J, Bechner M, Potamousis K, Goldstein S, *et al*: **A single molecule scaffold for the maize genome.** *PLoS Genet* 2009, **5**: e1000711.

56. Schwartz DC, Cantor CR: **Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis.** *Cell* 1984, **37**:67-75.

57. Dimalanta ET, Lim A, Runnheim R, Lamers C, Churas C, Forrest DK, *et al*: **A microfluidic system for large DNA molecule arrays.** *Anal Chem* 2004, **76**:5293-5301.

58. Lim A, Dimalanta ET, Potamousis KD, Yen G, Apodoca J, Tao C, *et al*: **Shotgun optical maps of the whole** *Escherichia coli* **O157:H7 genome.** *Genome Res* 2001, **11**:1584-1593.

59. Zhou S, Kvikstad E, Kile A, Severin J, Forrest D, Runnheim R, *et al*: **Whole-genome shotgun optical mapping of** *Rhodobacter sphaeroides* **strain 2.4.1 and its use for whole-genome shotgun sequence assembly.** *Genome Res* 2003, **13**:2142-2151.

60. Teague B, Waterman MS, Goldstein S, Potamousis K, Zhou S, Reslewic S, *et al*: **High-resolution genome structure by single molecule analysis.** *Proc Natl Acad Sci USA* 2010, **107**:10848-10853.

61. Lai Z, Jing J, Aston C, Clarke V, Apodaca J, Dimalanta ET, *et al*: **A shotgun optical map of the entire** *Plasmodium falciparum* **genome.** *Nat Genet* 1999, **23**:309-313.

62. Valouev A, Schwartz DC, Zhou S, Waterman MS: **An algorithm for assembly of ordered restriction maps from single DNA molecules.** *Proc Natl Acad Sci USA* 2006, **103**:15770-15775.