Comparative analysis of the genomes of clinical isolates of *Mycobacterium avium* subsp. *hominissuis* regarding virulence-related genes

Brendan Jeffrey,1 Sasha J. Rose,1,2 Kerrigan Gilbert,1 Matthew Lewis1,2 and Luiz E. Bermudez1,2,*

Abstract

**Purpose.** *Mycobacterium avium* subsp. *hominissuis* is a member of the *M. avium* complex, a heterogeneous group of bacteria that cause lung infection in immunocompetent patients or disseminated infection in patients with immunosuppression. The bacteria belonging to this complex have variable virulence, depending on the strain considered, and therefore a representative of the most common clinical phenotype was analysed.

**Methodology.** The genomic sequences of four *M. avium* subsp. *hominissuis* isolates obtained from clinical specimens were completed. Mav101, Mav100 and MavA5 were isolated from the blood of patients with AIDS. MavA5 was disseminated from the lung, while Mav3388 was isolated from the lungs of a patient with chronic lung disease. The sequences were annotated using the published Mav104 genome as a blueprint. Functional and virulence analyses of the sequences were carried out. Mice studies comparing the virulence of the strains were performed.

**Results.** Findings showed that while Mav101 was very similar to Mav104, there were numerous differences between Mav104 and the remaining strains at nucleotide and predicted protein levels. The presence of genes associated with biofilm formation and several known virulence-related genes were sometimes differentially present among the isolates, suggesting overlapping functions by different genetic determinants.

**Conclusions.** The sequences provided important information about *M. avium* heterogeneity and evolution as a pathogen. The limitation is the lack of understanding on possible overlapping functions of genes/proteins.

INTRODUCTION

The *Mycobacterium avium* complex is a ubiquitous group of environmental bacteria that are capable of infecting immunocompromised individuals, as well as patients with chronic lung diseases [1, 2]. *M. avium* usually infects through the respiratory tract in immunocompetent patients [3, 4], whereas it can infect immunocompromised patients through either the respiratory or gastro-intestinal tract [5]. Like the majority of pathogenic mycobacteria, *M. avium* is capable of infecting macrophages, where it can replicate and establish a long-term infection [6]. It is also known that *M. avium* strains are heterogeneous and their virulence can vary substantially [7].

The ever-increasing amount of genomic sequence data generated has expanded the understanding of the diversity seen in pathogenic bacteria. Previous work by Uchiya et al. on the comparative genomic analysis of two *M. avium* strains from pulmonary and disseminated infections clearly demonstrated genetic variation between the different virulent strains [7].

A common route of infection in immune-compromised patients is gastro-intestinal, and it has been shown that *M. avium* actively enters enterocytes in the intestinal mucosa [8, 9]. Virulence determinants have been identified in *M. avium* that are associated with pathogenicity during the initial phases of infection. For example, MAV_4604 and MAV_1406 are associated with biofilm formation, whereas MAV_4891 and MAV_3679 are necessary for invasion of the intestinal mucosa [10]. To be able to infect the respiratory tract, *M. avium*, which is dependent on several genes,
including MAV_3013, forms microaggregates [11, 12]. Once microaggregates are formed, some of the organisms cross the respiratory mucosa, while others establish themselves extracellularly in biofilms [11, 13]. Following the migration across the mucosal barrier, M. avium infects macrophages, where it lives inside vacuoles [14]. A number of M. avium genes have been described which allow the pathogen to survive within the phagocytic cell environment, including MAV_2450, MAV_3321 and MAV_4012 [15]. M. avium does not have the ESX-1 region, which is associated with the pathogenesis of Mycobacterium tuberculosis [16]. M. avium, however, does contain all the other esx regions, including ESX-5, which has been linked with the ability of the bacterium to survive in macrophages [17].

In the current study, we sequenced the genomes of four clinical M. avium isolates with differing in vivo and ex vivo phenotypes. These four strains include Mav101, Mav100, MavA5 and Mav3388. Mav101 is a strain isolated from the blood of an AIDS patient that shows high virulence in mice similar to that of the sequenced clinical isolate Mav104 [18]. Strain Mav100 is also isolated from the blood of an AIDS patient but has a low virulence phenotype in mice [19]. Strain MavA5 is a blood isolate from an infection initiated in the lungs [20], and is associated with robust biofilm production [21]. Strain Mav3388 is a lung isolate from a patient with a chronic lung disease. Therefore, the strains sequenced represent an array of strains isolated from recent lung infection, the blood of AIDS patients, and lung infection in immunocompromised patients. The genome sequences of the four strains, as well as the reference strain Mav104, were compared and the presence of known virulence determinants was analysed.

**METHODS**

**Mycobacterial strains and DNA preparation**

M. avium strains 100, 101, 3388 and A5 were grown in Middlebrook 7H9 media supplemented with 10 % oleic acid/albumin/dextrose/catalase (OADC) enrichment (Hardy Diagnostics, Santa Maria, CA, USA) at 37 °C for seven to nine days prior to genomic DNA extraction. Mycobacterial genomic DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany), following the pretreatment for Gram-positive bacteria protocol. Modifications to the protocol included incubating the bacterial pellet with 100 ng µl⁻¹ of lysozyme (Sigma, St. Louis, MO, USA) at 37 °C overnight instead of the enzymatic buffer, and performing two elutions to recover the maximum amount of genomic DNA. Vacuum centrifugation was used to concentrate the DNA samples prior to library preparation for sequencing.

**Mouse infection**

C57BL/6 mice (eight to ten weeks old) were obtained from Jackson laboratories for challenge studies. Fifteen mice per time point were infected through the caudal vein with approximately 2–3×10⁶ of bacteria in a volume of 100 µl. Spleens were obtained from mice at one week and four weeks post-infection, and homogenized in 5 ml of 7H9 broth (Difco) containing 20 % glycerol. Homogenized samples were then serially diluted, plated onto 7H10 agar plates, incubated at 37 °C, and colony-forming units were counted after 15 days.

**Genome sequencing and annotation**

Extracted genomic DNA was processed for sequencing using commercial Illumina DNA preparation kits (Illumina, San Diego, CA, USA) by following the manufacturer protocols. A paired-end 80 cycle sequencing reaction was performed on the Illumina Genome Analyzer IIX platform at the Center for Genome Research and Biocomputing core labs (Oregon State University, Corvallis, OR, USA). All reads were initially processed to remove the final 26 base pairs and any reads containing ambiguous bases (N's). Illumina-derived draft genomes were first assembled using the reference-guided assembly program MAQ [22] using the previously sequenced M. avium 104 genome [17] as the reference strain. De novo assembly was performed using Velvet [23] to complete regions that could not be resolved using MAQ, and a final draft genome sequence of several large contigs for each strain was generated. Draft genomes were automatically annotated using the PGAP annotation pipeline at NCBI, www.ncbi.nlm.nih.govgenome/annotation_prok/ [24].

**Comparative genomic analysis**

Large-scale insertions and deletions between strains were detected using the Mauve multiple genome alignment tool with the progressive Mauve alignment algorithm [25]. Orthologous protein sequences between pairs of sequenced strains were determined using the InParanoid algorithm with the blosum45 substitution matrix for prokaryotes [26]. The MultiParanoid program [27] was then used to automatically cluster orthologues from all sequenced strains.

Functional analysis of genes found in the core genome, as well as novel genes found in each sequenced strain, were performed using the Blast2Go suite of functional analysis tools [28]. Briefly, all protein sequences from the sequenced strains were aligned against the ‘nr’ blast database using the blastp program, with an E-value cut-off of 1E-6. Blast output was imported into the Blast2Go suite and sequences were then mapped and annotated following the program instructions, with the minimal alignment length cut-off set to 33 % query coverage. Gene ontology (GO) term distribution data for molecular function, biological process and cellular component were then extracted from the annotated sequences.

Analysis of possible virulent gene composition in each of the sequenced strains was carried out using ad hoc Python scripts, where protein annotation of interest was analysed in both the functional analysis of all proteins using the Blast2Go program, as well as parsing the original Blast output files for terms of interest. In the case of the proline/proline/glutamic (ppe) gene analysis, all candidate protein...
sequences were manually analysed to confirm the presence of the ppe motif in the N-terminal domain of the protein.

**Genome sequence accession numbers**

The draft genomes of the four sequenced *M. avium* strains have been deposited in the DDBJ/EMBL/Genbank database under the following accession numbers: Mav100, GCA_000696735.1; Mav101, GCA_000696755.1; MavA5, GCA_000696775.1; and Mav3388, GCA_000696775.1.

**RESULTS**

**Growth of *M. avium* isolates in mice**

To compare growth rates of the clinical isolates in a mouse model of systemic infection, C57BL/6 mice were infected intravenously with approximately 2–3 × 10^6 bacteria and infection burden in the spleen was determined at one week and four weeks post-infection (Table 1). Mav101, Mav104 and MavA5 had an approximate log increase in bacterial load in the spleen from one to four weeks post-infection, whereas there was a decrease from 8.8 × 10^6 to 4.6 × 10^3 bacteria in mice infected with Mav100. Growth and virulence characteristics for Mav3388 (lung isolate) were not performed in this experiment. The growth in mice for Mav100, Mav101 and Mav104 was consistent with results from a study conducted by Saunders et al. [19].

**Genome sequence analysis of sequence strains**

The overall genome statistics for each sequenced genome, as well as previously sequenced Mav104, are shown in Table 2. The largest genome (besides Mav104) was Mav101 with 5,498,738 base pairs and the smallest was MavA5 with 4,870,506 base pairs. The sizes of these genomes are expected to be slightly larger than the given numbers as the genomes are broken up into contiguous sequences (contigs). Regions with high GC-content, not effectively captured by Illumina sequencing, and repetitive transposon sequences contributed to the inability to join the contigs. Unique coding sequences present in the genome of Mav104 compared to the other three sequenced isolates exhibit the same pattern as found by Uchiya et al., when comparing *M. avium* TH135 to Mav104 [7]. As Mav100 contained several contigs that could not be placed accurately in the genome assembly, these sequences were concatenated to the 3' end of the genome sequence; these contigs contain 442 coding sequences unique to Mav100 (Fig. S1). There were many significant differences found among the strains studied (except for Mav101 and Mav104).

**Functional analysis**

Comparative orthologous gene analysis performed using the InParanoid tool [26] revealed that there were 4023 genes shared between the three sequenced isolates and Mav104. To determine the functional breakdown of this core genome, we used Blast2Go to classify these genes into molecular function, biological process and/or cellular component (Fig. 1a–c, respectively). Interestingly, a large number of genes involved in plasma membrane, cell wall, or being integral to the membrane were identified, many of which may be responsible for the distinctive cell wall of mycobacterial species.

To characterize the functional differences between the sequenced strains, the non-core genes of each strain were compared to the core genome (Fig. 2). These results show that the Mav104 genome contains genes involved in transposase activity (including DNA recombination) as well as various metabolic activities. GO term enrichment analysis of the genes specific to Mav100 and Mav3388 revealed enrichment in genes involved in various nucleic acid binding functions, indicating there may be several genes involved in transcriptional differences between these strains. An analysis of the genes found only in MavA5 revealed no enrichment of any GO terms (data not shown). The large numbers of unique genes found in each isolate provides a valuable pool for further study into the

**Table 1. Growth characteristics of *M. avium* subsp. hominissuis strains in C57BL/6 mice**

<table>
<thead>
<tr>
<th>Strain</th>
<th>Spleen c.f.u./organ after one week</th>
<th>Spleen c.f.u./organ after four weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mav104</td>
<td>4.2 ± 0.5 × 10^7</td>
<td>6.8 ± 0.4 × 10^8</td>
</tr>
<tr>
<td>Mav101</td>
<td>5.3 ± 0.4 × 10^7</td>
<td>7.9 ± 0.6 × 10^8</td>
</tr>
<tr>
<td>MavA5</td>
<td>4.8 ± 0.6 × 10^8</td>
<td>3.8 ± 0.5 × 10^8</td>
</tr>
<tr>
<td>Mav100</td>
<td>8.8 ± 0.5 × 10^5</td>
<td>4.6 ± 0.8 × 10^5</td>
</tr>
</tbody>
</table>

**Table 2. Genome statistics**

<table>
<thead>
<tr>
<th>Strain</th>
<th>Genome size (bp)</th>
<th>Contigs</th>
<th>CDS</th>
<th>Non-core CDS</th>
<th>Strain-specific CDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mav104</td>
<td>5,475,491</td>
<td>1</td>
<td>5120</td>
<td>1097</td>
<td>534</td>
</tr>
<tr>
<td>Mav101</td>
<td>5,465,242</td>
<td>156</td>
<td>5190</td>
<td>1169</td>
<td>72 (534)</td>
</tr>
<tr>
<td>Mav100</td>
<td>5,498,738</td>
<td>341</td>
<td>5188</td>
<td>1165</td>
<td>640</td>
</tr>
<tr>
<td>MavA5</td>
<td>4,870,506</td>
<td>50</td>
<td>4509</td>
<td>486</td>
<td>126</td>
</tr>
<tr>
<td>Mav3388</td>
<td>5,052,467</td>
<td>128</td>
<td>4688</td>
<td>665</td>
<td>194</td>
</tr>
</tbody>
</table>
Fig. 1. Core genome GO term counts. GO terms for all genes found in the shared core genome were determined using the Blast2GO program. For each panel, the counts for each GO term are given on the X-axis, and the GO term is given on the Y-axis. GO terms were for genes involved in molecular function (a), biological process (b) and cellular component (c).
Fig. 2. Differential GO term enrichment of non-core or novel genes. Fisher's exact test ($P<0.05$) was used to determine whether there was an enrichment of GO terms between the core genome (green bars, all panels) and the non-core genes (blue bars) in Mav104 (a), the novel genes found in Mav100 (b) and the novel genes found in Mav3388 (c). The percentage of the tested sequences of each GO term is shown on the X-axis, and the GO term is given on the Y-axis.
interaction between gene content and both environmental fitness and pathogenesis.

**Virulence gene content between strains**

Previous work by several laboratories has shown the importance of certain mycobacterial virulence gene families in the pathogenesis of mycobacterial species. These include the mammalian cell entry (mce) proteins, the mycobacterial membrane proteins (mmp), the ppe family proteins and the esx family proteins. To determine whether differing virulence gene content of our sequenced isolates could be associated with differing virulence in mice, we compared the presence or absence of these virulence determinants between isolates. In addition, we compared various virulence-associated genes that have been previously reported in *M. avium* studies.

**mce genes**

The *M. tuberculosis* mce proteins are a family of proteins that have been shown to be expressed during infection in humans [29], and the expression of these genes in *Escherichia coli* confers the ability of the bacterium to enter and survive within host macrophages [30]. Eleven mce genes, found primarily within two operons (MAV_2532 – MAV_3537 and MAV_5047 – MAV_5051), are not present in the other sequenced isolates and are unique to Mav104 (Fig. 3). Further analysis detected six putative mce genes not previously identified in Mav104, but all are present in the sequenced strains. As similar growth was observed in mice infected with Mav104 and MavA5 (Table 1), the presence of the 11 unique mce genes in Mav104 indicates that these genes are not necessarily crucial for successful *M. avium* infection, or that additional genes within MavA5 can perform similar functions to compensate for the loss of these mce genes. Furthermore, despite lower growth observed in mice infected with Mav100 (Table 1), the genome contained seven putative novel mce genes lacking in the other strains.

**mmp genes**

A distinctive feature common to all mycobacterial species is the high lipid content of their cell wall [31]. The mmp, termed mmpL and mmpS, are a family of proteins that have been shown to be involved in the transport of lipid metabolites for the biosynthesis of cell wall lipids [7, 32, 33]. It is possible that differing cell wall structure or constituents between the sequenced strains could help explain the observed variation in virulence in vivo. Fig. 4(a) shows the mmp gene content of the sequence strains, with both mmpL and mmpS genes included. MavA5 has gene sequence for all homologues of mmp genes found in Mav104, including MAV_3801, which is disrupted in Mav104. Similarly, Mav100 and Mav3388 contain the entire suite of mmp genes found in Mav104, with only MAV_1287 disrupted in strain Mav100 and MAV_0084 missing from strain Mav3388. Mav104 and MavA5 share six novel mmp genes, one of which (homologues of MAV_100_18425) is also present in the Mav3388 genome. Mav3388 contains four novel mmp genes that are not found in any of the other sequenced isolates. These novel mmp genes found in the sequenced isolates may play a role in these isolates having a modified cell wall, but these differences cannot necessarily explain differences in the virulence observed between the strains.

**ppe genes**

A comparative genomic study of mycobacterial species by Marri et al. showed that one of the major differences was in the genes that constitute the polymorphic ppe proteins [34]. The ppe gene family of acidic, glycine-rich proteins is named for the ppe motifs found near the N-termini of these proteins. To date no precise function is known for any members of this family of proteins, but some members have been associated with virulence [17] and have also been found to be antigens recognized by the host immune system [35]. Similar to the mce and mmp genes, the majority of ppe genes are present in all of the sequenced genomes (Fig. 4b).

A single ppe gene (MAV_0790) is unique to Mav104, while three ppe genes, disrupted in Mav104, are intact in Mav100 and MavA5. MavA5 contains two ppe genes, absent from Mav104, that it shares with the two other newly sequenced genomes, while Mav3388 and Mav100 contain two and three unique ppe genes, respectively. Mav3388 also contains disrupted copies of two ppe genes, MAV_2039 and MAV_2909, that are present in Mav100 and MavA5.

**esx genes**

The esx genes are contained in five regions of difference, named ESX-1 to ESX-5, that have been described in *M. tuberculosis* and contain the genes involved in both encoding secreted proteins as well as requisite genes for the type VII secretion system [36]. Comparative analyses revealed that *M. bovis* BCG is lacking ESX-1, which contributes to its attenuation during infection [37]. Mav104 does not contain ESX-1, but contains the other four esx regions [17]. We compared ESX-1 to ESX-5 between Mav104, Mav100, MavA5 and Mav3388 (Fig. 5). Similar to Mav104, MavA5 and Mav3388 possess Rv3876, but lack the rest of ESX-1. Interestingly, Mav100 contains roughly half of the genes contained in ESX-1. All four strains contain complete ESX-2 to ESX-4 regions, with the exception of Mav104; it has disrupted copies of Rv3992c and Rv3994c in ESX-2, and Rv0284 in ESX-3. All four strains contain an intact ESX-5 region, with only Rv1792 encoding a CFP-10 protein absent.

Of other virulence-related genes: among the intestinal mucosal invasion-related genes that have been identified [10, 38–40], all are present in all the strains sequenced with the exception of MAV_4658, which is missing in Mav3388 (Fig. 6a). This suggests that the ability to interact with the intestinal mucosa is conserved. Similarly, the majority of genes associated with survival and replication in macrophages are conserved [41–43], with the exception of MAV_2941 and MAV_4320, which are in Mav104 but not the other three strains (Fig. 6b). A pathogenicity island, previously identified in Mav104 [44], is not represented in the other strains (Fig. 6c). Genes associated with biofilm formation [13] were all conserved between the strains, except MAV_1385, which is disrupted in Mav104 (Fig. 6d). Similarly, MAV_3013, previously shown to be important for micro-aggregate formation [11], was seen in all
Fig. 3. Comparing mce gene content of each strain. Presence or absence of the mce genes is shown using coloured boxes. If the candidate mce gene is present in the Mav104 genome then the Mav104 gene name is shown on the left, otherwise the gene name from the appropriate strain is used.
Fig. 4. Comparing the mmp and ppe gene content of each strain. Presence or absence of the virulence genes is shown using coloured boxes. (a) mmp gene content. (b) ppe gene content. If the candidate virulence gene is present in the Mav104 genome then the Mav104 gene name is shown on the left, otherwise the gene name from the appropriate strain is used.
DISCUSSION

The genomes of four clinical isolates of *M. avium* were sequenced, annotated and compared to the published genome of Mav104. Three isolates, Mav101, Mav100 and MavA5, were obtained from blood samples (although the MavA5 blood infection was secondary to an initial lung infection), while the fourth (Mav3388) was obtained from a lung sample. Growth in C57Bl/6 mice was evaluated for Mav104, Mav100, Mav101 and MavA5 at one and four weeks post-infection (Table 1). Here we observed that three of the strains demonstrated similar c.f.u. at both time points, while Mav100 c.f.u. decreased at four weeks post-infection, raising the possibility that Mav100 does not contain important virulence factors responsible for fitness in mammals. The disparate observations of Mav100 being cleared from the spleen of infected mice while the other strains grew will certainly be explored in the future, utilizing the diverse genome regions identified in this study. Despite the genome of MavA5 being smaller and having significant differences from Mav104 and Mav101, all three strains behave similarly in mice. This suggests that the observed genomic differences of the sequenced strains (Fig. 6e). Lastly, genes associated with virulence in mice [15] were conserved between the strains (Fig. 6f).

**Fig. 5.** Examining the *esx* gene content of each strain. Although the ESX-1 region is mostly absent from the *M. avium* strains, the sequenced strains contain nearly all of the genes in the ESX-2, ESX-3, ESX-4 and ESX-5 regions.
### Fig. 6. Comparing other known *M. avium* virulence genes for each strain: (a) intestinal invasion-related genes; (b) macrophage virulence-related genes; (c) Mav104 pathogenicity island; (d) genes associated with biofilm formation; (e) MAV_3013, a gene involved with micro-aggregate formation; (f) decreased virulence in C57Bl/6 mice.

<table>
<thead>
<tr>
<th>Gene</th>
<th>MAV_104</th>
<th>MAV_100</th>
<th>MAV_A5</th>
<th>MAV_3388</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAV_1190</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_3679</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_4139</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_4671</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_5138</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_4324</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_4658</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_3616</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_0216</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_0383</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_0378</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_4891</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_0631</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_0024</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_1177</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_2925</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_1365</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_4478</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_0996</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_0660</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_2509</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_1094</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_2941</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_4320</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_2416</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_2417</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_2418</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_2419</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_2420</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_0996</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_0660</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_2509</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_1094</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_2941</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAV_4320</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Gene present, complete**
- **Gene present, disrupted**
- **Gene absent**
may be associated with phenotypes other than virulence, such as survival in the environment. Alternatively, bacterial strains may contain different genes in the genome that perform comparable function. It is important to consider that the mice were infected intravenously, which bypasses many host defences that may have a different impact between the strains.

Virulence is a complex phenotype, with many factors affecting a successful infection. Two important factors for \textit{M. avium} infection include the route of infection and production of biofilms. Mav104, Mav101 and MavA5 behave quite differently with respect to biofilm production, where MavA5 forms robust biofilm containing abundant extracellular DNA (eDNA), while Mav104 and Mav101 produce less robust biofilms with small amounts of eDNA [21, 45]. The analysis of the bacterial genomes indicated that all the sequenced strains have the genes identified with biofilm formation, except for MAV_1385, which is disrupted in Mav104. We recently reported that MavA5 utilizes bicarbonate to trigger eDNA export, and through transposon mutagenesis found a unique genomic region associated with the process [46]. More work will need to be done to see whether these other \textit{M. avium} strains also respond to bicarbonate, to determine exactly what genes in this region are conserved, and to study whether conserved genes in these different strains perform similar or different functions. Mav100 has all the genes currently associated with biofilm formation, but its glycopeptidolipids (GPL) have been shown to differ from the other strains [21]. Work by the Kolter laboratory, in fact, demonstrated a direct link between sliding motility, biofilm formation and the amount of GPL [47]. More recently, it became evident that the proteins encoded by the genes MAV_3013 and MAV_0831 are very important in the formation of microaggregates in the lung and the ability of microaggregates to invade the mucosa [11, 12]. Both genes are present in all the sequenced strains. \textit{mce} genes are also involved in the pathogenesis of mycobacteria, being associated with cell uptake [30], the ability to survive in mice [48] and more recently in cholesterol uptake and utilization inside mammalian cells [49]. Work on \textit{Mycobacterium smegmatis} demonstrated that \textit{mce} genes are implicated in maintenance of the surface properties of the cell [50]. In the sequenced strains it is clear that some of the \textit{mce} operons in Mav100, MavA5 and Mav3388 are absent, and that the Mav100 strain contains at least seven \textit{mce} genes that are not present in the other strains. Their functions are currently unknown, but definitely deserve attention.

Lipid transport is very important for the survival and virulence of mycobacteria and is linked with surface \textit{mpp} proteins [51]. Strains Mav100 and Mav3388 have a group of \textit{mpp} genes different from Mav101 and Mav104, though the majority of the \textit{mpp} genes are present in the five compared strains. \textit{mpp} proteins are essential for lipid transport across the membrane, secretion of host-modulating lipids and maintenance of the bacterial cell surface. The fact that Mav100 has seven \textit{mpp} proteins that differ from the virulent strains Mav101 and Mav104 is intriguing.

\textit{ppe} proteins, which are mostly found in mycobacteria, have been associated with antigenic properties and pathogenesis [52]. They all share a relatively conserved 180-amino acid N-terminal domain and are divided into subgroups on the basis of their variable C-terminal domains. Some \textit{ppe} proteins are expressed on the surface of mycobacteria [53–55], while in other cases the proteins are exported and have functions related to pathogenesis. MAV_2928 has a role in \textit{M. avium} pathogenesis and can modify the mycobacterial vacuole membrane [17, 43]. MAV_2928 is present in all the sequenced strains, including the less virulent Mav100. MAV_0790 is only present in Mav104 and Mav101, but not in the other sequenced strains. It is also noteworthy that some of the \textit{ppe} genes have mutations that either disrupt protein synthesis or alter the protein.

A major secretion system in mycobacteria is the type VII, which is mostly contained in the \textit{exs} regions in the genome. \textit{M. avium}, in contrast to \textit{M. tuberculosis} and \textit{Mycobacterium marinum}, lacks the ESX-1 region, which in \textit{M. tuberculosis} has been associated with important virulence characteristics [44]. \textit{M. avium} strains, nonetheless, contain regions ESX-2, ESX-3, ESX-4 and ESX-5. Among these regions, ESX-5 has been shown to be associated with virulence, as well as ESX-3 [17, 43]. All the strains sequenced have the majority of the genes in the regions ESX-2 to ESX-5.

Other genes associated with the invasion of epithelial cells or survival within macrophages are mostly conserved between the sequenced strains, with the exception of genes that are involved in amoeba infection [44, 56]. These genes are mostly present in Mav104 and Mav101 but are absent in Mav100, MavA5 and Mav3388 (although Mav3388 and MavA5 are fully pathogenic). This is an interesting observation, and may indicate that \textit{M. avium} strains adapted to the human host are the ones that contain genes acquired during the infection of protozoa in the environment. There are very likely more undescribed genes involved with both amoeba and macrophage infection, which strains MavA5 and Mav3388 probably have, and supplement their pathogenicity. This introduces a hypothesis that \textit{M. avium} is an environmentally adapted organism that is inefficient in infecting mammals; however, its natural evolution has resulted in strains that are increasingly well adapted to infect mammalian hosts.

In summary, numerous clinical and \textit{in vitro} observations have been made, suggesting a large degree of variation in the virulence of \textit{M. avium} strains. Early sequence data demonstrated that genomic differences may play a role in the variability of \textit{M. avium} infection. Here, we sequenced four diverse \textit{M. avium} clinical isolates and performed a comparative genomic analysis. We reported on the presence or absence of putative virulence factors between the strains, and future work will focus on elucidating the function of...
these genes. The information presented here adds to the understanding of *M. avium* pathogenesis.

**Funding information**
This work was supported by grant AI041399 of the National Institutes of Health to LEB. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

**Conflicts of interest**
The authors declare that there are no conflicts of interest.

**Ethical statement**
The research reported in the manuscript was performed adhering to scientific ethical principles. All the clinical isolates were obtained from a clinical laboratory approximately 20 years ago, and the identities of the patients were never revealed. This work has been approved by the Biosafety Committee of Oregon State University. The animal work conducted adhered to IACUC guidelines, and was approved institutionally before the experiment was conducted.

**References**


