INTRODUCTION

The mycobacterial envelope contains a unique array of glycolipids (Brennan, 1988; Aspinall et al., 1995). Members of the Mycobacterium avium complex (MAC) are distinguished from other members of this genus by the presence of highly antigenic cell surface molecules called glycopeptidolipids (GPLs) and the oligosaccharide portion of the serovar-specific GPL defines the 28 serovars. Previously, the genomic region (ser2) encoding the enzymes responsible for the glycosylation of the lipopeptide core to generate the serovar-2-specific GPLs has been described. In this work, the ser2 gene clusters of M. avium serovar 2 strains 2151 and TMC 724 were fully sequenced and compared to the homologous regions of M. avium serovar 1 strain 104, M. avium subsp. paratuberculosis and M. avium subsp. silvaticum. It was also determined that 104Rg, a mutant of strain 104 that produces truncated GPLs, lost several GPL biosynthesis genes by deletion. This comparison, together with analysis of protein similarities, supports a biosynthetic model in which serovar-2-specific GPLs are synthesized from a serovar-1-specific GPL intermediate that is derived from a non-specific GPL precursor. We also identified a gene encoding an enzyme that is necessary for the biosynthesis of serovar-3- and 9-specific GPLs, but not serovar-2-specific GPLs, suggesting that the different serovars may have evolved from the acquisition or loss of genetic information. In addition, a subcluster of genes for the biosynthesis and transfer of fucose, which are needed to make serovar-specific GPLs such as those of serovar 2, is found in the non-GPL-producing M. avium subspecies paratuberculosis and silvaticum.

Abbreviations: 6-dTal, 6-deoxytalose; Fuc, fucose; GPLs, glycopeptidolipids; MAC, Mycobacterium avium complex; Me, methyl; nsGPLs, non-specific glycopeptidolipids; Rha, rhamnose; ssGPLs, serovar-specific glycopeptidolipids.

The GenBank/EMBL/DDBJ accession numbers for the sequences reported in this paper are AF125999 and AF143772.
serovar-2-specific GPLs are synthesized from nsGPLs via a serovar-1-specific GPL intermediate, and raise interesting questions about the evolutionary mechanism for creating the different serovars.

**METHODS**

**Growth of bacterial strains and DNA isolation.** *E. coli* DH5α was used for propagation of all recombinant plasmids. Cells were grown in Luria–Bertani medium and recombinants were selected by the addition of ampicillin (100 μg ml⁻¹) or kanamycin (25 μg ml⁻¹). Recombinant plasmids were isolated from *E. coli* using a QiAprep Spin Miniprep Kit (Qiagen) according to the manufacturer’s instructions.

*M. avium* strain 104 and its rough derivative (104Rg) were from J. Torrelles and D. Chatterjee (Torrelles et al., 2002). These strains were grown on Middlebrook 7H11 agar (Difco) containing 10 % oleic acid–albumin–glucose–catalase (OADC). After 2 weeks growth at 37 °C, a single colony was inoculated Middlebrook 7H9 broth (Difco) containing 10 % OADC. The 1 ml starting culture was up-scaled several times to a final volume of 200 ml. Cells were harvested by centrifugation at 3000 g for 30 min. For isolation of genomic DNA, harvested cells were resuspended in TES buffer (50 mM Tris, pH 8.0; 10 mM EDTA, pH 8.0; 100 mM NaCl) and stored on ice. Cells were lysed by vortexing with 0.5 mm zirconium beads for 2 min. SDS (final concentration 1 %) was added to the lysed cells and kept on ice for 15 min. The lysate was treated with proteinase K (final concentration 1 %) and kept on ice for 15 min, then extracted with phenol/chloroform/isooamyl alcohol (25:24:1). The aqueous phase was loaded on an equilibrated Genomic Tip 100 (Qiagen) and the genomic DNA was eluted with TE (10 mM Tris, pH 8.0; 1 mM EDTA, pH 8.0), according to the manufacturer’s instructions.

**Cloning, restriction endonuclease mapping and sequencing of ser2 and flanking regions from *M. avium* strains 2151 and TMC 724.** The recombinant plasmids used in this study, and the procedures used for their derivation are shown in Table 1. These plasmids were characterized by digestion with restriction endonucleases Smal, KpnI, PstI, EcoRI, EcoRV, SacI, NotI and BamHI (Invitrogen), and subclones were constructed by ligation of restriction fragments into the respective sites of the pBluescript II SK(−) vector (Stratagene) using T4 DNA ligase (Invitrogen). These subclones were used as double-stranded DNA templates for DNA sequencing with the pBluescript T3 and M13–20 primers. Custom primers were synthesized as necessary to resolve sequence ambiguities. Sequencing of DNA was performed by Macromolecular Resources Facility, Colorado State University. Contiguous DNA sequences, ORF analysis and codon usage were determined with SeqMan 3.0 software (Gene Codes Corporation) and FramePlot 2.3beta (http://www.ncbi.nih.gov/∼june/cgi-bin/frameplot.pl) (Ishikawa & Hotta, 1999). Identification of the putative function of each ORF was achieved via similarity searches between the deduced amino acid sequences and known proteins using BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) and multi-alignments were generated using MultAlin (http://www.toulouse.inra.fr/multalin.html) (Corpet, 1988).

**PCR analysis, cloning and sequencing of the ser1 and flanking regions remaining in *M. avium* strain 104Rg.** Forward primer P1 (5′-CCGGCCGTTCTGGTGGAAGTG-3′) and reverse primer P2 (5′-GATGCCCCAAGCATTCTCTT-3′) were synthesized by the Macromolecular Resources Facility at Colorado State University. PCR amplification was performed using a 2400 GeneAmp PCR System (Perkin–Elmer) and Taq polymerase (Invitrogen). The PCR programme used with the above primers was 28 cycles of 94·0 °C for 1 min, 63·4 °C for 1 min and 72·0 °C for 2·5 min. Prior to the first cycle, the starting temperature of 94·0 °C was held for 5 min, and at the end of the last cycle, a temperature of 72·0 °C was held for 7 min. The PCR product was resolved on a 0·8 % agarose gel and stained with ethidium bromide.

The 3690 bp PCR product was purified from the gel using the QiAquick Gel Extraction Kit (Qiagen), according to the manufacturer’s instructions, cloned into pGEM-Teasy (Promega) and the insert was sequenced using the SP6 and T7 primers. Subclones for DNA sequencing were generated with SstI, Ncol and HinI (Invitrogen). Sequencing of DNA was performed by the Macromolecular Resources Facility at Colorado State University. DNA sequence alignments, ORFs and codon usage were determined with Sequencher 3.0 software (Gene Codes Corporation). The site of the genomic deletion was facilitated by comparing the sequence data to the genome sequence data from *M. avium* strain 104 provided by The Institute for Genomic Research (http://www.tigr.org).

**Nucleotide sequence accession numbers.** The sequence data presented in this article are available in GenBank at AF143772 (**M. avium** 2151) and AF125999 (**M. avium** TMC 724). The previously published rfa sequence data (Eckstein et al., 1998) at AF060183 are included in AF125999, and the previously published 151601 sequence data (Eckstein et al., 2000) at AF060182 are included in AF143772.

<table>
<thead>
<tr>
<th>Table 1. Recombinant plasmids used for DNA sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Plasmid</strong></td>
</tr>
<tr>
<td>pITB11.1</td>
</tr>
<tr>
<td>pITB11.2</td>
</tr>
<tr>
<td>pITB71.1</td>
</tr>
<tr>
<td>pITB230.1</td>
</tr>
<tr>
<td>pITB230.2</td>
</tr>
<tr>
<td>pITB232.2</td>
</tr>
<tr>
<td>pITB232.3</td>
</tr>
</tbody>
</table>

*Recombinant cosmids pITB11 and pITB71 are overlapping clones derived from *M. avium* TMC 724 genomic DNA (Belisle et al., 1991), and pITB230 and pITB232 are overlapping clones derived from *M. avium* 2151 genomic DNA (Belisle et al., 1993b). †Restriction fragments were subcloned into the HindIII or Dral/HindIII site of vector pBluescript II.
### Table 2. Analysis of the ORFs in the ser2 regions of *M. avium* serovar 2 strains 2151 and TMC 724

<table>
<thead>
<tr>
<th>Type/name of predicted protein or IS element</th>
<th>Coding position in serovar 2 strain</th>
<th>Predicted molecular mass (kDa)</th>
<th>Predicted pI</th>
<th>Predicted function (GenBank accession no. and name of most closely related non-<em>M. avium</em> protein) or note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycosyltransferases</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RtfA</td>
<td>12 339–13 625</td>
<td>47.1</td>
<td>7.33</td>
<td>Rhamnosyltransferase (Q50583; Rv1524)</td>
</tr>
<tr>
<td>GtfA</td>
<td>10 658–11 935</td>
<td>46.8</td>
<td>6.36</td>
<td>Glycosyltransferase (AF192151.4; Gtf2)</td>
</tr>
<tr>
<td>GtfB</td>
<td>18 984–17 728</td>
<td>45.8</td>
<td>8.42 (8.08)</td>
<td>Glycosyltransferase (AF192151.4; Gtf2)</td>
</tr>
<tr>
<td>GtfC</td>
<td>24 002–24 811</td>
<td>29.9</td>
<td>8.87</td>
<td>Glycosyltransferase (P71793; Rv1514c)</td>
</tr>
<tr>
<td>GtfD</td>
<td>28 153–28 953</td>
<td>30.2</td>
<td>7.32</td>
<td>Glycosyltransferase (B70670; Rv2957)</td>
</tr>
<tr>
<td>Methyl- and acetyltransferases</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MtfA</td>
<td>8321–9448</td>
<td>43.3</td>
<td>7.72</td>
<td>Methyltransferase (AB032523.1; AveBVII)</td>
</tr>
<tr>
<td>MtfB</td>
<td>9715–10 536</td>
<td>30.7</td>
<td>5.86</td>
<td>Methyltransferase (D16097.1; MycF)</td>
</tr>
<tr>
<td>MtfC</td>
<td>13 727–14 527</td>
<td>30.0</td>
<td>6.91</td>
<td>Methyltransferase (D16097.1; MycF)</td>
</tr>
<tr>
<td>MtfD</td>
<td>14 627–15 445</td>
<td>30.6</td>
<td>5.62</td>
<td>Methyltransferase (AF192151.3; Mtf1)</td>
</tr>
<tr>
<td>MtfE</td>
<td>19 126–19 779</td>
<td>23.1</td>
<td>8.20</td>
<td>Methyltransferase (U81296.1; Pcm)</td>
</tr>
<tr>
<td>MtfF</td>
<td>27 286–28 008</td>
<td>26.4</td>
<td>5.40</td>
<td>Methyltransferase (A70670; Rv2956)</td>
</tr>
<tr>
<td>AtfA</td>
<td>Not found</td>
<td>40.1</td>
<td>9.87</td>
<td>Acetyltransferase (AF192151; Atf1)</td>
</tr>
<tr>
<td>D-Glucose dehydrogenase</td>
<td></td>
<td></td>
<td></td>
<td>Acetyltransferase (AF192151; Atf1)</td>
</tr>
<tr>
<td>D-Mannose dehydratase (B70714; GmdA)</td>
<td></td>
<td></td>
<td></td>
<td>Acetyltransferase (AF192151; Atf1)</td>
</tr>
<tr>
<td>6-Deoxy-4-keto-D-mannose reductase/epimerase</td>
<td></td>
<td></td>
<td></td>
<td>Acetyltransferase (AF192151; Atf1)</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DhgA</td>
<td>Disrupted</td>
<td>27.5</td>
<td>9.29</td>
<td>Dehydrogenase (AL132973.1; SCF91.07)</td>
</tr>
<tr>
<td>HlpA</td>
<td>Disrupted</td>
<td>37.2</td>
<td>6.00</td>
<td>Haemolysin (AF052516.1; Hhp)</td>
</tr>
</tbody>
</table>
RESULTS

Sequence analysis and definition of the ser2 gene cluster

The ser2 gene clusters of *M. avium* serovar 2 strains 2151 and TMC 724, as well as their flanking regions, were sequenced and the ORFs were determined (Table 2). Elucidation of putative gene function was facilitated by comparing this sequence information to the homologous regions in the genome sequence data from *M. avium* serovar 1 strain 104 provided by The Institute for Genomic Research (http://www.tigr.org). The data were also compared to the GS element of *M. avium* subsp. *paratuberculosis* and *M. avium* subsp. *silvaticum* (Tizard et al., 1998; Bull et al., 2000). The alignments of the five genetic maps of these regions are shown in Fig. 1.

The ser2 gene cluster is defined as the genomic region encoding the enzymic machinery necessary for the glycosylation of the lipopeptide core that results in the serovar-2-specific GPL (Belisle et al., 1993a). This cluster was

---

**Fig. 1.** Restriction endonuclease map (a) and genetic organization of the gene clusters involved in the glycosylation of the lipopeptide core of *M. avium* strain 104 (a), *M. avium* strain 2151 (b), *M. avium* strain TMC 724 (c), *M. avium* subsp. *silvaticum* (d) and *M. avium* subsp. *paratuberculosis* (e). The boxes represent the following homologous gene clusters (from left to right as shown in panel b): gene cluster upstream of ser1 (left crosshatching), ser1 gene cluster (black), ser3 gene cluster (white), fbt gene cluster (grey), drt gene cluster (right crosshatching). Genes encoding putative proteins with unknown functions are labelled with orf and letters or numbers. The names of the ORFs in parentheses in (d) and (e) are from Bull et al. (2000) and Tizard et al. (1998). In (e), *orfA'* is the ORF lacking 66 nt from the 3' end of *orfA* (a and b), and *gtfC'* is the ORF lacking 258 nt from the 5' end of *gtfC* (b, c and d). Restriction sites are: H, *HindIII*; K, *KpnI*; P, *PstI*.
previously delineated by restriction mapping, Southern blot analysis and the characterization of naturally occurring rough morphotypes that could produce only the non-glycosylated lipopeptide core (Belisle et al., 1993a, b; Eckstein et al., 2000). Based on the sequence data and ORF analyses, we propose that the genetic definition of the ser2 gene cluster be all of the ORFs from mtfA to mtfE, and gtfC to gtfD (Fig. 1b, c), and that the ser2 region is composed of two functional subclusters designated ser1 and fbt that are flanked by genes that do not appear to be involved in GPL biosynthesis (see Fig. 1 and legend). As explained below, the ser1 and fbt subclusters can be assigned possible functions with regard to the synthesis of the serovar-2-specific GPLs, while the intervening ser3' region appears to be an anomaly.

**Definition and proposed role for the ser1 subcluster**

The lipopeptide cores of all MAC GPLs are glycosylated at two positions: the C-terminal L-alaninol is glycosylated with a mono- or dimethylated Rha (it can also be trimethylated (MtfA, MtfB, MtfC and MtfD) (Table 2). The deduced ser1 specific GPL.

The simplest of the ssGPLs is that of serovar 1 (Table 3) which contains a single Rha residue linked to 6-dTal. This basic oligosaccharide structure is in all other ssGPLs except for those of serovars 5 and 10/11 (Aspinall et al., 1995) and so the genes for its glycosylation and methylation should be present in all the relevant serovars. The ser1 region (indicated by the solid black box in Fig. 1) was common to both serovar 1 (strain 104) and serovar 2 (strains 2151 and TMC 724), and the ORFs encompassed by this region are proposed to participate in the biosynthesis of the serovar-1-specific GPL.

The ser1 subcluster encodes two putative glycosyltransferases (GtfA, GtfB), a known rhamnosyltransferase (RtfA) (Eckstein et al., 1998) and four putative methyltransferases (MtfA, MtfB, MtfC and MtfD) (Table 2). The deduced amino acid sequences of GtfA, GtfB and RtfA are highly similar to putative glycosyltransferases of *Mycobacterium tuberculosis* (Rv1524, Rv1526c) (Cole et al., 1998) and of *Mycobacterium leprae* (L518_C2_147) (Eiglemeier et al., 1998). They also show high levels of identity (63.4–65.1%) and similarity (84.9–86.8%) to one another, with GtfB and RtfA being the most similar. RtfA transfers Rha to 6-dTal (Eckstein et al., 1998) and by analogy we propose that GtfB transfers Rha to the C-terminal L-alaninol of the lipopeptide core. This leaves GtfA as the likely candidate for adding 6-dTal to the D-alle-threonine.

The roles for GtfA and GtfB proposed above and the order in which these glycosylation steps occur were further examined by genetic analysis of strain 104Rg, a naturally occurring rough derivative of *M. avium* strain 104. Torrelles et al. (2002) determined that 104Rg produces truncated GPLs that lack the 6-dTal attached to D-alle-threonine, but retain the methylated Rha on the L-alaninol. In the present study, PCR analysis of 104Rg genomic DNA in comparison to that of the parental strain indicated that 104Rg contains gtfB, but not gtfA (data not shown). Given the presence of two copies of IS999 flanking the gtfA region in strain 104 (Fig. 1a), we reasoned that the absence of gtfA in 104Rg might be due to deletion, based on our previous finding that homologous recombination between direct repeats of IS1601 resulted in the deletion of the ser2 gene cluster to produce the Rg-1 morphotype of strain 2151 (Eckstein et al., 2000). However, these particular copies of IS999 are inverted repeats rather than direct repeats in the genome sequence of strain 104 (http://www.tigr.org) and homologous recombination between them should result in inversion rather than deletion. Nevertheless, based on the PCR results, the deletion hypothesis was tested by using primers based on orfA and rtfA sequences flanking the left and right (Fig. 1a) copies of IS999, respectively, to amplify 104Rg genomic DNA. A 3690 bp PCR product was obtained and this was cloned and sequenced. The DNA sequence indicates that 104Rg lost the 10 kb region containing orfB-gtfA while retaining one copy of IS999 (the left copy in Fig. 1a). It is possible that the 28 bp inverted repeats at the ends of IS999 mediated the deletion event since Mahairas et al. (1996) identified 24 and 12 bp repeats that were involved in the deletion of the RD2 and RD3 regions, respectively, in *Mycobacterium bovis* BCG. The region deleted from strain 104 includes three genes (mtfA and mtfB as well as gtfA) from the ser1 subcluster. Since 104Rg produces a lipopeptide core with methylated Rha on the L-alaninol (Torrelles et al., 2002), and we found that it has lost gtfA while retaining gtfB, it follows that GtfB is likely to transfer Rha to the C-terminal L-alaninol of the lipopeptide core before GtfA adds 6-dTal to the D-alle-threonine. Thus, the proposed biosynthetic scheme begins with GtfB (Fig. 2).

As noted above, 104Rg also lost mtfA and mtfB. The methyltransferases form two different groups based on either genetic location or deduced amino acid similarities. The mtfA and mtfB genes are physically separated from the

---

**Table 3. Structural similarity of the oligosaccharides of ssGPLs of *M. avium* serovars 1, 2, 3 and 9**

<table>
<thead>
<tr>
<th>Serovar</th>
<th>Oligosaccharide structure linked to D-alle-threonine</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>z1-6-dTal-(2→1)-z5-4-Rha</td>
</tr>
<tr>
<td>2</td>
<td>z1-6-dTal-(2→1)-z5-4-Rha-(3→1)-2,3-di-O-Me-z1-L-Fuc</td>
</tr>
<tr>
<td>3</td>
<td>z1-6-dTal-(2→1)-z5-4-Rha-(3→1)-2,3-di-O-Me-z1-L-Fuc-(4→1)-β-D-GlcA-(4→1)-2,3-di-O-Me-z1-Fuc</td>
</tr>
<tr>
<td>9</td>
<td>z1-6-dTal-(2→1)-z5-4-Rha-(3→1)-2,3-di-O-Me-z1-L-Fuc-(4→1)-β-D-GlcA-(4→1)-4-O-Ac-2,3-di-O-Me-z1-L-Fuc</td>
</tr>
</tbody>
</table>

GlcA, Glucuronic acid; Ac, acetyl.
mtfC and mtfD genes by gtfA and rtfA (Fig. 1). MtfA and MtfD are most similar to each other (37.5% identity and 65.6% similarity), as are MtfB and MtfC (85.3% identity and 94.4% similarity). Only MtfB and MtfC showed significant similarity to the known mycinamin III-O-methyltransferase of Micromonospora griseorubida (Inouye et al., 1994). MtfD is highly similar (90.6% identity) to a methyltransferase (Mtf1 or MeTase1) in M. smegmatis that adds a methyl (Me) group to the 3-position of the Rha attached to L-alaninol (Patterson et al., 2000). Mtf1 is not highly related (53.0% similarity) to MtfC. It is thus proposed that MtfC is responsible for the generation of 4-O-Me-Rha and MtfD is responsible for the 3-O-Me-Rha to produce nsGPL (Fig. 2). Since strain 104 also produces an nsGPL containing 3-O-Me-6-dTal that is missing in 104Rg (Torrelles et al., 2002), our genetic analyses support a proposed pathway for the biosynthesis of the serovar-2-specific GPL in which GtfB, MtfC and MtfD carry out the first two steps, namely the addition of the dimethylated Rha to L-alaninol, while GtfA and MtfA (or MtfB) are most likely involved in the addition of 6-dTal or 3-O-Me-6-dTal to D-allo-threonine (Fig. 2). The final action of the above-mentioned RtfA, namely the addition of Rha to 6-dTal, would thus lead to the formation of the serovar-1-specific GPL as shown in Fig. 2.

**The anomalous ser3’ gene cluster**

The ser3’ subcluster (indicated by the white box in Fig. 1) was so-named because it encodes a putative D-glucose dehydrogenase that would be needed to make the D-glucuronic acid found in the ssGPLs of M. avium serovars 3 or 9 (Table 3). The deduced amino acid sequence demonstrated a high degree of similarity (69 and 65%) to known D-glucose dehydrogenases of Bacillus subtilis (Hilt et al., 1991) and Pseudomonas aeruginosa (van Schie et al., 1985), respectively. The presence of this subcluster in the serovar 2 strains is anomalous because D-glucuronic acid is not found in the serovar-2-specific GPL.
Proposed role for the fbt subcluster in the de novo biosynthesis and transfer of L-fucose (L-Fuc)

The presence of dimethylfucose is the only difference between the serovar-2-specific and serovar-1-specific GPLs (Table 3), and so the genes for Fuc synthesis and addition should be present in serovar 2 strains, but not serovar 1. The region defined as fbt (Fuc biosynthesis and transfer; indicated by the grey box in Fig. 1) fulfills this criterion and appears to contain the genes needed for the de novo biosynthesis and transfer of L-Fuc. The deduced amino acid sequences of gtfC, mdhA, mtfA, mtfF and gtfD demonstrated a high degree of similarity (52 % identity, 67 % similarity) to proteins involved in the de novo synthesis and transfer of L-Fuc in *E. coli* (Stevenson et al., 1996; Andrianopoulos et al., 1998). Interestingly, this gene cluster is identical to the GS element of *M. avium* subsp. *paratuberculosis* and subsp. *silvaticum* (Tizard et al., 1998; Bull et al., 2000) although neither of these subspecies expresses GPLs (see Discussion).

It is possible that one of the two putative glycosyltransferase genes (*gtfC* and *gtfD*) belongs to the *ser3* rather than the *fbt* gene cluster because a transferase would be needed for the glucuronic acid found in the ssGPLs of serovars 3 or 9. However, *mtfE* is clearly part of the second subcluster, despite its physical separation from the other *fbt* genes, based on the earlier work of Mills et al. (1994). In that study, transposon mutants of the cloned *ser2* region from *M. avium* strain 2151 were analysed in *M. smegmatis* to identify genes involved in ssGPL biosynthesis (the *M. avium* genes required for nsGPL synthesis could not be characterized because *M. smegmatis* makes nsGPLs). The loci responsible for making 2-O-Me-Fuc and 3-O-Me-Fuc map to the *mtfF* and *mtfE* genes, respectively.

Comparison of the *ser2* regions of strains 2151 and TMC 724

Comparison of the two *ser2* gene clusters and their flanking regions revealed that the nucleotide sequences and gene organization were strongly conserved and only a few nucleotide differences were observed. Most were silent mutations, but some affected the deduced amino acid sequences of *gtfB*, *mtfE* and *gdhA* (4, 2 and 2 aa, respectively). Besides these highly conserved characteristics, there were three major differences between the organization of the *ser2* gene clusters in *M. avium* TMC 724 and 2151. First, two ORFs encoding a putative dehydrogenase (*dhgA*) and a putative haemolytic protein (*hlpA*) are disrupted by IS1245 and IS1348, respectively, in *M. avium* 2151, but not in *M. avium* TMC 724. Second, a potential hot spot for insertion sequences was observed downstream of the *fbt* gene cluster in both *M. avium* strains, but IS1601 was found in 2151 while IS2534 was found in TMC 724. The site of integration of these two IS elements in the two strains differs by only 133 bp. Finally, *M. avium* TMC 724 contained a putative O-acetyltansferase gene (*atfA*) downstream of the *fbt* subcluster that is not present in the region sequenced for *M. avium* 2151.

**DISCUSSION**

In previous studies, detailed restriction endonuclease mapping and Southern blot analyses revealed restriction fragment length differences in the *ser2* gene clusters of *M. avium* serovar 2 strains 2151 and TMC 724 (Belisle et al., 1993b). This work shows that the DNA sequences of both *ser2* gene clusters and their flanking regions are nearly identical, with the above-mentioned differences in restriction patterns being due to IS elements. Two ORFs encoding a putative dehydrogenase (*dhgA*) and a putative haemolytic protein (*hlpA*) are disrupted by IS elements in strain 2151, but not in strain TMC 724. Given that strains 2151 and TMC 724 both produce ssGPLs (Belisle et al., 1991, 1993a), these genes must not be involved in the biosynthesis of serovar-2-specific GPLs. Maslow et al. (1999) reported a correlation between haemolytic activity and disseminated infection, whereby *M. avium* strains from patients with disseminated MAC infection demonstrated haemolytic activity while strains from patients with pneumonia lacked this characteristic. Based on this observation, one can speculate that HlpA (as well as DhgA) might contribute to the intracellular survival of the bacilli. Comparison of the growth characteristics of these two strains in macrophages shows that strain TMC 724 grows faster and produces more colony forming units than does strain 2151 (Pedrosa et al., 1994; Florido et al., 1997).

Both *ser2* gene clusters have the same organization of genes encoding the enzymic machinery for the biosynthesis of the ssGPL, and this type of genetic conservation has been observed in other systems, such as the gene clusters responsible for expression of *E. coli* group 1 K-antigens (Rahn et al., 1999). The nucleotide and deduced amino acid sequences of the ORFs in the *ser2* regions of both strains are identical, with three exceptions (*gtfB*, *mtfE* and *gdhA*) as noted earlier. Given the similarity in structure of Rha, Fuc and 6-dTal, one would expect a high degree of similarity in the enzymes transferring these 6-deoxyhexose sugars. Indeed, RtfA, GtfA and GtfB are very similar to one another. However, the putative fucosyltransferases GtfC and GtfD are not related to one another or the other three glycosyltransferases. None of the five glycosyltransferases contains the E-x(7)-E-A-x(18)-E hexosyltransferase motif identified by Henderson & Nataro (1999). However, this motif was based on the analysis of glucosyl-, galactosyl- and mannosyltransferases, and not 6-deoxyhexosyltransferases, suggesting that the loss of the hydroxyl group on the sixth carbon might be the reason for the mismatch. Multi-alignment of putative 6-deoxyhexosyltransferases yielded two new motifs (Fig. 3a, b). The first motif, G-[TS]-R-G-D-x-[EQ]-P-x-x-A-x(4)-L-x(3)-G-x-x-V (Fig. 3a), is in the N-terminal 60 aa of the transferase, and the second motif, [DE]-x(18)-[AG]-[VI]-[VI]-H-H-G-G-x-G-[TS]-T (Fig. 3b), is about 200 aa away. Interestingly, NovM, an enzyme that...
transfers a Rha that is C-methylated at the fifth carbon, does not exhibit the first motif, but contains the second one. This suggests that the fifth and the sixth carbons of the 6-deoxyhexoses play a major role in binding to the first motif.

RtfA, GtfA and GtfB contain both of the proposed 6-deoxyhexosyltransferase motifs while GtfC and GtfD do not. This supports our biosynthetic model (see below) in which GtfC or GtfD transfer methylated Fuc molecules and would thus have different binding and transfer motifs than the 6-deoxyhexosyltransferases.

Comparison of the deduced amino acid sequences of the six methyltransferases found in the ser2 gene cluster with putative enzymes responsible for the methylation of 6-deoxyhexoses revealed the motif

\[
\text{[LM]}-x(13)-[DEN]-x(10-11)-[GA]-x(6)-[DE]
\]

(Fig. 4). This motif is not one of the five highly conserved amino acid sequences (regions I–V) that were identified by Ibrahim et al. (1998) when they analysed several plant O-methyltransferases. These enzymes are responsible for the methylation of chalcones, flavanones, flavones, isoflavones, flavonols and anthocyanins, but not sugars, and so one might not expect any similarity to the methyltransferases in our study. However, regions I–IV are thought to be involved in \text{S}-adenosyl-L-methionine and metal binding, and \text{S}-adenosyl-L-methionine is the common methyl donor, so we expected our methyltransferases to demonstrate some homology to those proposed regions. One obvious explanation is that the binding site for the methyl donor is different in bacteria and plants.

All of this information was used to propose the biosynthesis cascade shown in Fig. 2. Briefly, GtfB adds Rha to the C-terminal L-alaninol of the lipopeptide core and this Rha is then dimethylated by MtfC and MtfD. Both the transfer of this particular 6-deoxyhexose, as well as the proposal that the sugar is not methylated until after it is
transferred, correlate with the high degree of similarity between GtfB and the known rhamnosyltransferase RtfA (Eckstein et al., 1998) as well as the structure of the truncated GPL made by 104Rg (Torrelles et al., 2002). GtfA demonstrated less homology to RtfA than did GtfB, so we hypothesized that GtfA is responsible for the transfer of 6-dTal to D-ALLO-threonine to produce the nsGPL. The nsGPL then serves as the substrate for RtfA which adds Rha to the 6-dTal to yield the serovar-1-specific GPL. In serovar 1, the 6-dTal can alternatively be methylated at the 3-position (Torrelles et al., 2002), a process that is assigned to Mtfa or MtFB, primarily because mtfta and mtfB are also found in serovar 4 (Krzywinska & Schorey, 2003) and are thus not serovar-specific. These methytransferases could also modify the fatty acid portion of the GPLs, as has been shown for Mt2 in M. smegmatis (Jeevarajah et al., 2002), and so this is the most speculative portion of the proposed pathway. In the final part of the biosynthetic model (Fig. 2), the de novo synthesis of Fuc by MdhA and MerA, and the methylation of Fuc by Mtfe and Mtff, are followed by the linkage of the dimethylated Fuc to the non-methylated Rha by GtfC or GtfD to generate the serovar-2-specific GPL.

The proposed pathway is similar to the biosynthetic cascades for other cell-wall surface markers. Stevenson et al. (1996) described a biosynthetic pathway for the extracellular colanic acid of E. coli K-12 in which the four sugars are synthesized de novo in the cytosol and then transferred to the lipid carrier in the inner membrane. Polymerization and acetylation are then performed in the outer membrane of the cell wall. Similar pathways have been proposed by Arakawa et al. (1995) for the biosynthesis of the serotype K2 capsular polysaccharide of Klebsiella pneumoniae, and by Hashimoto et al. (1993) for the synthesis of the Vi antigen in Salmonella typhi. The transport of the cell surface markers to the final destination is usually mediated by an ABC transporter system (Hashimoto et al., 1993; Stevenson et al., 1996). We could not identify the genes for such a transport system for the GPLs within the ser2 gene clusters of M. avium 2151 or TMC 724. However, a second transport mechanism was discussed by Stevenson et al. (1996) in which the lipid carrier involved in the synthesis of colanic acid acts as a transporter. They proposed that the lipid carrier is flipped from the cytoplasm into the periplasm. Further work is needed to determine whether this transport mechanism or the ABC transporter system is responsible for the transfer of the GPLs to the outer surface of the cell wall of M. avium. In this regard, it should be noted that Recht et al. (2000) have suggested that mtfc might be involved in both the biosynthesis and transport of GPLs in M. smegmatis.

Downstream of the ser2 gene cluster of M. avium TMC 724 is an ORF (atfa) encoding a putative O-acetyltransferase. Atfa contains all three of the acetyltransferase motifs proposed by Slauch et al. (1996). The atfa gene is not found in the analogous regions of strains 2151 or 104, but it is present within the GS gene cluster of M. avium subsp. paratuberculosis and subsp. silvaticum. Although Atfa is not necessary for the biosynthesis of ssGPLs, it might be involved in GPL modification since ssGPLs can sometimes be acetylated, methylated or sulfated (Aspinall et al., 1995; Chatterjee & Khoo, 2001). Recht & Kolter (2001) recently showed that the atf1 gene of M. smegmatis encodes an acetyltransferase that acylates one or two sites on the 6-dTal of ssGPL. The nucleotide sequences on either side of IS2534-atfa in strain 724 are identical to those surrounding the right copy of IS1601 (Fig. 1) in strain 2151. This suggests that atfa may have been inserted along with the IS element, although IS2534-atfa is not flanked by direct repeats.

The presence of repetitive elements in the ser regions raises interesting possibilities with regard to the evolution of the different serovars. The ser2 region is flanked by IS1601 in strain 2151, and the ser1 region is flanked by IS999 and RE (a repetitive element containing portions of IS999) in strain 104, suggesting that these gene clusters together with the flanking IS elements form ‘biosynthetic islands’. A similar gene organization was found for the cap gene cluster of H. influenzae (Kroll, 1992) and the K10 capsule gene cluster in E. coli (Clarke et al., 1999). The IS elements flanking the cap genes generated a duplication of the gene cluster resulting in enhanced production of the capsule polysaccharide (Kroll, 1992). Clarke et al. (1999) found that the K10 capsule genes are flanked by IS3 and prophage elements, leading them to speculate that these elements were involved in the acquisition of the capsule gene cluster at its present chromosomal location. Two observations from the present work suggest that the acquisition or loss of such biosynthetic islands has occurred in the MAC. First, the anomalous ser3 gene subcluster contains gdhA, a gene encoding a putative D-glucose dehydrogenase. This enzyme produces D-glucuronic acid, which is found only in the ssGPLs of M. avium serovars 3 or 9 (Aspinall et al., 1995) (Table 3). Thus, the ser2 gene cluster in M. avium TMC 724 and 2151 might represent an intermediate step in which a serovar 2 strain acquired new genetic information to produce serovars 3 and 9. The second observation is that the fbt gene cluster of M. avium 2151 and TMC 724 is identical to the GS element, an 8.9 kb genomic region reported to be present in M. avium subsp. paratuberculosis and M. avium subsp. silvaticum, but absent from some strains of M. avium subsp. avium (Tizard et al., 1998; Bull et al., 2000). Interestingly, the GS region from M. avium subsp. silvaticum contains exactly the same IS element in exactly the same site as in M. avium strain TMC 724 (IS1612 is identical to IS2534). This suggests that M. avium subsp. silvaticum might have been derived from a serovar 2 strain of M. avium such as strain TMC 724, or vice versa.

In summary, this work represents a comprehensive analysis of the genes responsible for the glycosylation of the GPLs of M. avium. The data presented show a strong conservation of gene sequences and organization together with significant
differences among strains that suggests that the acquisition or deletion of genes has resulted in different serovars with specific haptenic oligosaccharides as well as different subspecies. This work also provides a template for future research to confirm gene function in the pathway for GPL biosynthesis and the role of GPLs in the biology of M. avium. One obstacle to these types of studies is the poor transformability of M. avium subspecies avium. Thus far, we have not been able to transform strain 2151 and so the genetic analyses of GPL biosynthesis in M. avium have been restricted to naturally occurring deletion mutants. However, J. N. Maslow, S.-W. Lee, T. M. Eckstein, J. M. Inamine and J. T. Belisle (unpublished) have recently used a transformation frequency of 4 \( \pm 4 \) [4 c.f.u. (\( \mu \)g DNA)\(^{-1} \)] with strain 724 to successfully carry out allelic exchange mutagenesis, and so it should be possible to test individual genes in the proposed GPL biosynthesis pathway in future studies.

ACKNOWLEDGEMENTS

This work was supported by grants AI-41925 and AI-51283 from the National Institute of Allergy and Infectious Diseases, National Institute of Health.

REFERENCES


