The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+C Gram-positive bacteria

Nico C Gey van Pittius*, Junaid Gamieldien†, Winston Hide†, Gordon D Brown‡, Roland J Siezen§ and Albert D Beyers*

Addresses: *US/MRC Centre for Molecular and Cellular Biology, Department of Medical Biochemistry, University of Stellenbosch, Tygerberg, 7505, South Africa. †South African National Bioinformatics Institute (SANBI), University of the Western Cape, Bellville, 7535, South Africa. ‡Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK. §Center for Molecular and Biomolecular Informatics, University of Nijmegen, 6525ED Nijmegen, The Netherlands.

Correspondence: Nico C Gey van Pittius. E-mail: ngvp@sun.ac.za

Published: 19 September 2001

*Genome Biology* 2001, 2(10):research0044.1–0044.18

The electronic version of this article is the complete one and can be found online at [http://genomebiology.com/2001/2/10/research/0044](http://genomebiology.com/2001/2/10/research/0044)

© 2001 Gey van Pittius et al., licensee BioMed Central Ltd

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Abstract

**Background:** The genome of *Mycobacterium tuberculosis* H37Rv has five copies of a cluster of genes known as the ESAT-6 loci. These clusters contain members of the CFP-10 (hp) and ESAT-6 (esat-6) gene families (encoding secreted T-cell antigens that lack detectable secretion signals) as well as genes encoding secreted, cell-wall-associated subtilisin-like serine proteases, putative ABC transporters, ATP-binding proteins and other membrane-associated proteins. These membrane-associated and energy-providing proteins may function to secrete members of the ESAT-6 and CFP-10 protein families, and the proteases may be involved in processing the secreted peptide.

**Results:** Finished and unfinished genome sequencing data of 98 publicly available microbial genomes has been analyzed for the presence of orthologs of the ESAT-6 loci. The multiple duplicates of the ESAT-6 gene cluster found in the genome of *M. tuberculosis* H37Rv are also conserved in the genomes of other mycobacteria, for example *M. tuberculosis* CDC1551, *M. tuberculosis* 210, *M. bovis*, *M. leprae*, *M. avium*, and the avirulent strain *M. smegmatis*. Phylogenetic analyses of the resulting sequences have established the duplication order of the gene clusters and demonstrated that the gene cluster known as region 4 (Rv3444c-3450c) is ancestral. Region 4 is also the only region for which an ortholog could be found in the genomes of *Corynebacterium diphtheriae* and *Streptomyces coelicolor*.

**Conclusions:** Comparative genomic analysis revealed that the presence of the ESAT-6 gene cluster is a feature of some high-G+C Gram-positive bacteria. Multiple duplications of this cluster have occurred and are maintained only within the genomes of members of the genus *Mycobacterium*.

**Background**

*Mycobacterium tuberculosis* remains a serious threat to human health and in spite of significant investment in research on this organism, the mechanisms of its pathogenicity are still not clearly understood. One strategy used to determine these mechanisms is to compare the presence and absence of genes in different species (for example, virulent and avirulent) and extrapolate these differences to variation...
in phenotype. The genomes of *M. tuberculosis* H37Rv, *M. tuberculosis* H37Ra, *M. bovis* and the attenuated *M. bovis* BCG have been compared in different combinations using a variety of methods (subtractive genomic hybridization [1], bacterial artificial chromosome (BAC) restriction profile analysis [2-5], BAC arrays [6], DNA microarrays [7] and Southern blotting [8]). This has identified a number of regions of difference (RD) between the various organisms.

One of these regions, designated the RD1 (region of difference 1) deletion region [1], is a 9,505 bp region absent in all *M. bovis* BCG strains. RD1 is commonly thought to be the primary deletion that occurred during the serial passage of *M. bovis* by Calmette and Guérin between 1908 and 1921, and is thus thought possibly to be responsible for the primary attenuation of *M. bovis* to *M. bovis* BCG [5,7]. Consequently, the genes contained in RD1 have been the object of a number of studies focusing on diagnosis of *M. tuberculosis* infection, the search for efficient vaccine candidates and virulence [9-12]. RD1 encompasses the genes Rv3871 to Rv3879c (annotation according to [13]), which include the genes for the 6 kDa early-secreted antigenic target ESAT-6 (esx or esat-6) and L45 homologous protein CFP-10 (*lhp*) [14,15]. The esat-6 and *lhp* genes are situated immediately adjacent to each other and encode potent T-cell antigens that are secreted but lack detectable secretion signals [16,17].

During the genome sequencing of *M. tuberculosis* H37Rv, Cole et al. [13] identified at least 11 additional genes encoding small proteins of approximately 100 amino acids that share sequence similarities with ESAT-6, and grouped them into the *esat-6* gene family. In addition, they found several small genes with similarity to *lhp* (which encodes the protein CFP-10) that are also situated directly adjacent to the *esat-6* family genes. Sequence analyses indicated that the *lhp* family members belong to and extend this *esat-6* gene family. It was also found that the *lhp* gene is co-transcribed and thus forms part of an operon with *esat-6* [15].

The genes encoding the originally annotated CFP-10 and ESAT-6 proteins within the RD1 deletion region lie in a cluster of 12 other genes (encompassing the deletion region), which seems to have been duplicated five times in the genome of *M. tuberculosis*. The duplicated gene clusters have been previously described as the ESAT-6 loci in an analysis of the proteome of *M. tuberculosis* [18]. An examination of the sets of genes in the clusters reveals that each of the clusters also contains (in addition to a copy of *esat-6* and *lhp*), genes encoding putative ABC transporters (integral inner-membrane proteins), ATP-binding proteins, subtilisin-like membrane-anchored cell-wall-associated serine proteases (the myosins [19]), and other amino-terminal membrane-associated proteins [18].

We have compared sequences to establish the relationship between the multiple copies of the ESAT-6 gene cluster. Our results show that the ESAT-6 gene cluster is of ancient origin, is present in, and restricted to, the genomes of other members of the high G+C Gram-positive bacteria such as *Corynebacterium diphtheriae* and *Streptomyces coelicolor*, and is duplicated multiple times in *M. tuberculosis* and other mycobacteria. We discuss the conservation of this gene cluster in the context of its possible functional importance and its use in diagnosis of mycobacterial infection.

**Results**

**Individual gene families and genomic organization in *M. tuberculosis***

The five ESAT-6 gene clusters present in *Mycobacterium tuberculosis* H37Rv were named region 1 (Rv3866-Rv3883c), 2 (Rv3884c-Rv3895c), 3 (Rv0282-Rv0292), 4 (Rv3444c-Rv3450c) and 5 (Rv1782-Rv1798), consistent with the arbitrary numbering system used previously to classify the five myosin (subtilisin-like serine protease) genes identified from these regions [19]. Orthologs of the ESAT-6 gene clusters of *M. tuberculosis* H37Rv could be identified in the genomes of eight other strains and species belonging to the genus *Mycobacterium*, as well as in two species belonging to other genera (Table 1). Up to 12 different genes representing different gene families were identified in the five gene cluster regions and were designated families A to L according to their position in region 1 (Table 2).

Figure 1 shows a schematic representation of the genomic organization of the gene families present in each of the five ESAT-6 cluster regions of *M. tuberculosis*. Annotations and descriptions of single genes in these regions can be found at [20]. Regions 1 and 2 are situated directly adjacent to each other in the genome and are transcribed in opposite directions. In both regions 1 and 5 the large gene belonging to family D (encoding the ATPase protein) has been disrupted by an insertion (Figure 1). This insertion has resulted in an in-frame stop codon, giving rise to two smaller genes (containing all the motifs of the larger homolog) located directly adjacent to each other. The gene positions of members of families C, D, G and H are maintained in all five regions (see Figure 1), whereas most of the families that are not present in region 4 seem to be more flexible with regard to their position within the gene clusters (families A, B, I and L). There are also genes present within some of the ESAT-6 gene clusters that do not have any homologs in the other clusters, suggesting subsequent insertions or deletions from the ancestral region (indicated by black arrows in Figure 1, see also Table 2).

The *esat-6/lhp* operon is not only present in the ESAT-6 gene clusters, but is distributed as six additional copies in the genome of *M. tuberculosis* (Figure 2). In four cases, the *esat-6/lhp* operon is flanked by *ppe* and *pe* genes (encoding proteins that have proline-proline-glutamic acid (PPE) and proline-glutamic acid (PE) motifs, respectively), indicating...
ESAT-6 gene cluster identification in other mycobacteria

Table 2 presents the results of the similarity searches and all available data for the 12 identified gene families present in the different regions. All the mycobacteria currently being sequenced contain multiple copies of these regions in their genomes. As these different copies are also found in the same respective genomic locations (corresponding flanking genes) in all the mycobacteria, it indicates that the duplication events took place prior to the divergence of the different species.

M. tuberculosis CDC1551, M. tuberculosis 210 and M. bovis

The genomes of the M. tuberculosis CDC1551 and 210 clinical strains as well as the genome of M. bovis contain all five of the ESAT-6 gene cluster regions present in the genome of M. tuberculosis H37Rv (sharing between 99 and 100% similarity to M. tuberculosis H37Rv at protein level). It is interesting to note, however, that two of the genes present in region 2 in CDC1551 (MT4000 and MT4001) contain frameshifts in their sequences, indicating that they and the rest of the region may no longer be functional in CDC1551. Part of region 2 (a 2,405 bp fragment containing Rv3888c, Rv3888c and Rv3889c) is also deleted in certain strains of M. bovis only, including the strain AF2122/97 that is currently being sequenced [21]. An in-frame stop codon found in Rv1792 (family G) is also present in the orthologs in CDC1551 (MT1841) and strain 210 (MTB1966), indicating that the mutation may have taken place before divergence of the three strains. Two of the H37Rv genes as well as the strain 210 family D genes (in regions 1 and 5) have acquired in-frame stop codons, resulting in two genes lying adjacent to each other, whereas the family D Rv1783 and Rv1784 orthologs in CDC1551 are still one intact gene (MT1833). The orthologs of this gene in M. bovis (MB771.1D), M. leprae (ML1543), M. avium (MA221D), and M. paratuberculosis (MP1783) are also intact, implying that the mutation in the H37Rv and strain 210 orthologs must have occurred after divergence of the three M. tuberculosis strains.

M. leprae

Figure 3 shows a schematic representation of the genomic organization of the respective gene families present in each of the five ESAT-6 gene clusters of M. leprae. The genome sequence of M. leprae contains functional copies of two of the five ESAT-6 gene cluster regions (regions 1 and 3, which have between 50 and 70% similarity to M. tuberculosis H37Rv at protein level). Most of the genes from region 2 are deleted, and all the remaining genes in this region have
## Table 2

Presence of genes in gene clusters of all available finished and unfinished genome sequences

| Gene family | Description | Protein size (in Mtb) | ESAT-6 cluster region | M. tuberculosis H37Rv | M. tuberculosis CDC1551 (CSU93) | M. tuberculosis \(^{+}\) 210 | M. bovis \(^{+}\) AF2122/97 (spoligotype 9) | M. bovis \(^{+}\) BCG Pasteur 1173P2 |
|-------------|-------------|-----------------------|-----------------------|-----------------------|------------------------|----------------------|--------------------------|---------------------|
| A           | ABC transporter family signature, 19-27% homology | 283 | 1 | Rv3866 | MT3980 | ND | MB851A | No sequence data |
|             |             | 276 | 2 | Rv3889c | MT4004 | MTB12A | MB727.3A | No sequence data |
|             |             | 295 | 3 | Rv0289 | MT0302 | MTB203A | MB548A | No sequence data |
|             |             | 300 | 5 | Rv1794 | MT1843 | MTB196A | MB557A | No sequence data |
| B           | AAA+ class ATPases, CBXX/CFQX family, SpoVII, 1x ATP/GTP-binding site, 29-39% homology | 573 | 1 | Rv3868 | MT3981 | MTB44B | MB851B | No sequence data |
|             |             | 619 | 2 | Rv3884c | MT3999 | MTB12B | MB727.1B | No sequence data |
|             |             | 631 | 3 | Rv0282 | MT0295 | MTB23B | MB672B | No sequence data |
|             |             | 610 | 5 | Rv1798 | MT1847 | MTB196B | MB542B | No sequence data |
| C           | Amino-terminal transmembrane protein, ATP/GTP-binding motif, 31-41% homology | 480 | 1 | Rv3869 | MT3982 | MTB44C | MB851C | No sequence data |
|             |             | 495 | 2 | Rv3895c | MT4011 | MTB36C | MB780.1C | No sequence data |
|             |             | 538 | 3 | Rv0283 | MT0296 | MTB23C | MB672C | No sequence data |
|             |             | 470 | 4 | Rv3450c | MT3556 | MTB45C | MB493.1C | No sequence data |
|             |             | 506 | 5 | Rv1782 | MT1832 | MTB46C | MB771.1C | No sequence data |
| D           | DNA segregation ATPase, ftsK chromosome, partitioning protein, SpoIIIE, yukA, 3x ATP/GTP-binding sites, 28-39% homology | 747 | 1 | Rv3870+71 | MT3983+85 | MTB44D+Db | MB851D | MB851D |
|             |             | 1396 | 2 | Rv3894c | MT4010 | MTB3D | MB780.1D | No sequence data |
|             |             | 1330 | 3 | Rv0284 | MT0297 | MTB23D | MB672D | No sequence data |
|             |             | 1236 | 4 | Rv3447c | MT3553 | MTB45D | MB855.1D | No sequence data |
|             |             | 435 | 5 | Rv1783+84 | MT1833 | MTB46D+Db | MB771.1D | No sequence data |
| E           | PE, 18-90% homology | 99 | 1 | Rv3872 | MT3986 | MTB44E | MB851E | Deleted |
|             |             | 77 | 2 | Rv3893c | MT4008 | MTB3E | MB780.1E | No sequence data |
|             |             | 102 | 3 | Rv0285 | MT0298 | MTB23E | MB389E | No sequence data |
|             |             | 99 & 99 | 5 | Rv1788 & 91 | MT1837 & 40 | MTB196E & Eb | MB771.0E & MB557E | No sequence data |
| F           | PPE, 19-88% homology | 368 | 1 | Rv3873 | MT3987 | MTB44F | MB851F | Deleted |
|             |             | 399 | 2 | Rv3892c | MT4007 | MTB3F | MB780.1F | No sequence data |
|             |             | 513 | 3 | Rv0286 | MT0299 | MTB472F | MB528F | No sequence data |
|             |             | 365 | 4 | Rv1787 & 89 & 90 | MT1836 & 38 & 39 | MTB196A & Fb & Fc | MB771.0Fa & Fb & MB557F | No sequence data |
| G           | lhp or CFP-10, also MTSA-10, grouped into ESAT-6 family, potent secreted T-cell antigens, 9-32% homology | 100 | 1 | Rv3874 | MT3988 | MTB44G | MB851G | Deleted |
|             |             | 107 | 2 | Rv3891c | MT4006 | MTB12G | MB727.3G | No sequence data |
|             |             | 97 | 3 | Rv0287 | MT0300 | MTB472G | MB548G | No sequence data |
|             |             | 125 | 4 | Rv3445c | MT3550 | MTB45G | MB585.0G | No sequence data |
|             |             | 98 | 5 | Rv1792 (Stop) | MT1841 (Stop) | MTB196G (Stop) | MB557G | No sequence data |
Table 2 (continued)

| Gene family | Description | Protein size (in M. tb) | ESAT-6 cluster region | M. tuberculosis | M. tuberculosis* | M. tuberculosis H37Rv | M. bovis* | M. bovis* AF212/97 | M. bovis* CDC1551 (CSU93) | M. paratuberculosis K10 | M. avium* 104 | M. smegmatis* MC2 155 | C. diphtheriae* NCTC13129 | S. coelicolor A3 (2) |
|-------------|-------------|------------------------|-----------------------|----------------|-----------------|----------------------|----------|-------------------|------------------------|---------------------|----------------|-------------------|-------------------|----------------|
| H | ESAT-6 family, cfp7, L45 or l-esat, also Mtb B9 family, potent secreted T-cell antigens, 15-27% homology | 95 | 1 | Rv3875 | MT3989 | MTB44H | MB851H | Deleted |
| | cfp7, L45 or l-esat, also Mtb B9 family, potent secreted T-cell antigens, 15-27% homology | 95 | 2 | Rv3890c | MT4005 | MTB12H | MB727.3H | No sequence data |
| | cfp7, L45 or l-esat, also Mtb B9 family, potent secreted T-cell antigens, 15-27% homology | 96 | 3 | Rv0288 | MT0301 | MTB203H | MB548H | No sequence data |
| | cfp7, L45 or l-esat, also Mtb B9 family, potent secreted T-cell antigens, 15-27% homology | 100 | 4 | Rv344c | MT3549 | MTB45H | MB585.0H | No sequence data |
| | cfp7, L45 or l-esat, also Mtb B9 family, potent secreted T-cell antigens, 15-27% homology | 94 | 5 | Rv1793 | MT1842 | MTB196H | MB557H | No sequence data |
| I | ATPases involved in chromosome partitioning, 1 ATP/ - ATP-binding motif, - 33% homology | 666 | 1 | Rv3876 | MT3990 | MTB60I | MB477I | Deleted |
| | ATPases involved in chromosome partitioning, 1 ATP/ - ATP-binding motif, - 33% homology | 341 | 2 | Rv3888c | MT4003 | MTB12I | Deleted # | No sequence data |
| | ATPases involved in chromosome partitioning, 1 ATP/ - ATP-binding motif, - 33% homology | - | 3 | No duplication | No duplication | No duplication | No duplication | No duplication |
| | ATPases involved in chromosome partitioning, 1 ATP/ - ATP-binding motif, - 33% homology | - | 4 | No duplication | No duplication | No duplication | No duplication | No duplication |
| J | Integral inner membrane protein, binding-protein-dependent transport systems inner membrane component signature, putative transporter protein, 19-27% homology | 511 | 1 | Rv3877 | MT3991 | MTB369J | MB477J | Deleted |
| | Integral inner membrane protein, binding-protein-dependent transport systems inner membrane component signature, putative transporter protein, 19-27% homology | 509 | 2 | Rv3887c | MT4002 | MTB12J | MB727.3J | No sequence data |
| | Integral inner membrane protein, binding-protein-dependent transport systems inner membrane component signature, putative transporter protein, 19-27% homology | 472 | 3 | Rv0290 | MT0303 | MTB203J | MB548J | No sequence data |
| | Integral inner membrane protein, binding-protein-dependent transport systems inner membrane component signature, putative transporter protein, 19-27% homology | 467 | 4 | Rv3448 | MT3554 | MTB45J | MB585.1J | No sequence data |
| | Integral inner membrane protein, binding-protein-dependent transport systems inner membrane component signature, putative transporter protein, 19-27% homology | 503 | 5 | Rv1795 | MT1844 | MTB196J | MB506J | No sequence data |
| K | Mycosins, subtilisin-like cell-wall associated serine proteases, 43-49% homology | 446 | 1 | Rv3883c | MT3998 | MTB12Ka | MB727.0K | No sequence data |
| | Mycosins, subtilisin-like cell-wall associated serine proteases, 43-49% homology | 550 | 2 | Rv3886c | MT4001 (Frame) | MTB12Kb | MB727.2K | No sequence data |
| | Mycosins, subtilisin-like cell-wall associated serine proteases, 43-49% homology | 461 | 3 | Rv0291 | MT0304 | MTB203K | MB548K | No sequence data |
| | Mycosins, subtilisin-like cell-wall associated serine proteases, 43-49% homology | 455 | 4 | Rv3449 | MT3555 | MTB45K | MB585.1K | No sequence data |
| | Mycosins, subtilisin-like cell-wall associated serine proteases, 43-49% homology | 585 | 5 | Rv1796 | MT1845 | MTB196K | MB506K | No sequence data |
| L | 2x amino-terminal transmembrane protein, 16-27% homology | 462 | 1 | Rv3882c | MT3997 | MTB12La | MB727.0L | No sequence data |
| | 2x amino-terminal transmembrane protein, 16-27% homology | 537 | 2 | Rv3885c | MT4000 (Frame) | MTB12Lb | MB727.2L | No sequence data |
| | 2x amino-terminal transmembrane protein, 16-27% homology | 331 | 3 | Rv0292 | MT0305 | MTB203L | MB694.0L | No sequence data |
| | 2x amino-terminal transmembrane protein, 16-27% homology | 406 | 4 | Rv1797 | MT1846 | MTB196L | MB542L | No sequence data |
| | 2x amino-terminal transmembrane protein, 16-27% homology | 585 | 5 | Rv1798 | MT1845 | MTB196K | MB506K | No sequence data |

Presence and names of genes in each species:

| Gene family | Description | Protein size (in M. tb) | ESAT-6 cluster region | M. leprae TN | M. avium* 104 | M. paratuberculosis* K 10 | M. smegmatis* MC2 155 | C. diphtheriae* NCTC13129 | S. coelicolor A3 (2) |
|-------------|-------------|------------------------|-----------------------|----------------|-----------------|----------------------|-------------------|-------------------|----------------|
| A | ABC transporter family signature, 19-27% homology | 283 | 1 | ML0057(pseudo) | ND | ND | MS29A | ND | ND |
| | ABC transporter family signature, 19-27% homology | 276 | 2 | MLabc (pseudo)* | MA138A | MP3889c | ND | ND | ND |
| | ABC transporter family signature, 19-27% homology | 295 | 3 | ML2530 | MA141A | MP0289 | MS32A | ND | ND |
| | ABC transporter family signature, 19-27% homology | - | 4 | No duplication | No duplication | No duplication | No duplication | No duplication | No duplication |
| | ABC transporter family signature, 19-27% homology | 300 | 5 | ML1540 | MA310A | MA138A | MP3889c | ND | ND |
| B | AAA+ class ATPases, CBXX/CFQX family, SpoVK, 1x ATP/GTP binding site, 29-39% homology | 573 | 1 | ML0055 | ND | ND | MS29A | ND | ND |
| | AAA+ class ATPases, CBXX/CFQX family, SpoVK, 1x ATP/GTP binding site, 29-39% homology | 619 | 2 | ML0039(pseudo) | MA177B | MP3884c | ND | ND | ND |
| | AAA+ class ATPases, CBXX/CFQX family, SpoVK, 1x ATP/GTP binding site, 29-39% homology | 631 | 3 | ML2537 | MA788 | MP0282 | MS32A | ND | ND |
| | AAA+ class ATPases, CBXX/CFQX family, SpoVK, 1x ATP/GTP binding site, 29-39% homology | - | 4 | No duplication | No duplication | No duplication | No duplication | No duplication | No duplication |
Table 2 (continued)

| Gene family | Description | Protein size (in M.b) | ESAT-6 cluster region | Presence and names of genes in each species |
|-------------|-------------|-----------------------|-----------------------|--------------------------------------------|
| C           | Amino-terminal transmembrane protein, possible motif, 31-41% homology | 480 | 1 | ML0054 | ND | ND | MS29C | ND | ND |
|             |                         | 495 | 2 | Deleted | MA144C | MP3895c | ND | ND | ND |
|             |                         | 538 | 3 | ML2536 | MA78C | MP0283 | MS32C | ND | ND |
|             |                         | 470 | 4 | Deleted | MA94C | MP3450c | MSBC | CORDmem | SC3C3.07 |
|             |                         | 506 | 5 | ML1544 | MA221C | MP1782 | ND | ND | ND |
| D           | DNA segregation ATPase, ftsK chromosome partitioning protein, SpoIIE, yukA, 3x ATP/GTP-binding sites, 2x amino-terminal transmembrane protein, 28-39% homology | 747+591 | 1 | ML0053+52 | ND | ND | MS29D (Stop$) | ND | ND |
|             |                         | 1396 | 2 | Deleted | MA144D | MP3894c | ND | ND | ND |
|             |                         | 1330 | 3 | ML2535 | MA78D | MP0284 | MS32D | ND | ND |
|             |                         | 1236 | 4 | Deleted | MA504D | MP3447c | MS8D | CORDyuk | SC3C3.20c |
|             |                         | 435+932 | 5 | ML1543 | MA221D | MP1783 | ND | ND | ND |
| E           | PE, 18-90% homology | 99 | 1 | Deleted | ND | ND | MS29E | ND | ND |
|             |                         | 77 | 2 | Deleted | MA138E | MP3893c | ND | ND | ND |
|             |                         | 102 | 3 | ML2534 | MA78E | MP0285 | MS32E | ND | ND |
|             |                         | - | 4 | No | No | No | No | No | No |
|             |                         | 99 & 99 | 5 | duplication | duplication | duplication | duplication | duplication | duplication |
| F           | PPE, 19-88% homology | 368 | 1 | ML0051 | ND | ND | MS29F | ND | ND |
|             |                         | 399 | 2 | Deleted | MA138F | MP3892c | ND | ND | ND |
|             |                         | 513 | 3 | ML2533 (pseudo) | MA78F | MP0286 | MS32F | ND | ND |
|             |                         | - | 4 | No | No | No | No | No | No |
|             |                         | 365, 393 & 350 | 5 | duplication | duplication | duplication | duplication | duplication | duplication |
| G           | lhp or CFP-10, also MTSA-10, grouped into ESAT-6 family, potent secreted T-cell antigens, 9-32% homology | 100 | 1 | ML0050 | ND | ND | MS29G | ND | SC3C3.10 |
|             |                         | 107 | 2 | Deleted | MA138G | MP3891c $ | ND | ND | SC3C3.11(c) |
|             |                         | 97 | 3 | ML2532 | MA141G | MP0287 | MS32G | ND | ND |
|             |                         | 125 | 4 | Deleted | MA319G | MP3445c | MS8G | CORDcfp10 | ND |
|             |                         | 98 | 5 | MLcfp (pseudo)$ | MA310G | MP1792 | ND | ND | ND |
| H           | ESAT-6 family, cfp7, L45 or l-est, also Mtb9.9 family, potent secreted T-cell antigens, 15-27% homology | 95 | 1 | ML0049 | ND | ND | MS29H | ND | SC3C3.10 |
|             |                         | 95 | 2 | ML0034 (pseudo) | MA138H | MP3890c $ | ND | ND | SC3C3.11f |
|             |                         | 96 | 3 | ML2531 | MA141H | MP0288 | MS32H | ND | ND |
|             |                         | 100 | 4 | ML0363 | MA319H | MP3444c | MS8H | CORDdesat6 | ND |
|             |                         | 94 | 5 | MLesat (pseudo)$ | MA310H | MP1793 | ND | ND | ND |
| I           | ATPases involved in chromosome | 666 | 1 | ML0048 | ND | ND | MS29I | ND | SC3C3.03c |
|             |                         | 341 | 2 | ML0035 (pseudo) | MA138I | MP3888c | ND | ND | ND |
### Table 2 (continued)

| Gene family | Description | Protein size (in Mt.b) | ESAT-6 cluster region | Presence and names of genes in each species |
|-------------|-------------|------------------------|-----------------------|--------------------------------------------|
|             |             |                        |                       | M. leprae | M. avium* | M. paratuberculosis* | M. smegmatis* | C. diphtheriae* | S. coelicolor |
|             |             |                        |                       | TN | 104 | K | 10 | MC² | 155 | NCTC13129 | A3 (2) |
| **partitioning, 1x** | ATP/GTP-binding motif, 33% homology | - | 3 | No duplication | No duplication | No duplication | No duplication | No duplication | No duplication | No duplication |
| **J** | Integral inner membrane protein, binding-protein-dependent transport systems inner membrane component signature, putative transporter protein, 19-27% homology | 511 | 1 | ML0047 | ND | ND | MS29J | ND | ND | ND |
| | | 509 | 2 | ML0036 (pseudo) | MA138J | MP3887c | ND | ND | ND | ND |
| | | 472 | 3 | ML2529 | MA141J | MP0290 | ND | MS32J | ND | ND |
| | | 467 | 4 | Deleted | MA504J | MP3448 | MS8J | CORDransporter | SC3C3.21 |
| | | 503 | 5 | ML1539 | MA310J | MP1795 | ND | ND | ND | ND |
| **K** | Mycosins, subtilisin-likecell-wall associated serine proteases, 43-49% homology | 446 | 1 | ML0041 | ND | ND | MS65K | ND | ND | ND |
| | | 550 | 2 | ML0037 (pseudo) | MA177K | MP3886c | ND | ND | ND | ND |
| | | 461 | 3 | ML2528 | MA141K | MP0291 | ND | MS32K | ND | ND |
| | | 455 | 4 | Deleted | MA439K | MP3449 | MS8K | CORDsub | SC3C3.17c |
| | | 585 | 5 | ML1538 | MA310K | MP1796 | ND | ND | ND | ND |
| **L** | 2x amino-terminal transmembrane protein, 16-27% homology | 462 | 1 | ML0042 | ND | ND | MS65L | ND | ND | ND |
| | | 537 | 2 | ML0038 (pseudo) | MA177L | MP3885c | ND | ND | ND | ND |
| | | 331 | 3 | ML2527 | MA81L | MP0292 | ND | MS32L | ND | ND |
| | | - | 4 | No duplication | No duplication | No duplication | No duplication | No duplication | No duplication |
| | | 406 | 5 | ML1537 | MA310L | MP1797 | ND | ND | ND | ND |

#### Other region-specific genes of known functions (not assigned to a family)

| Region 5 (not present in M. smegmatis, C. diphtheriae and S. coelicolor) | Rv1785c | Probable member of the cytochrome P450 family (pseudogene in M. leprae) |
|-----------------------------|---------|---------------------------------------------------------------------|
| Rv1786 | Probable ferredoxin (pseudogene in M. leprae) |

#### Other region-specific genes of unknown functions (not assigned to a family)

| Region 1 (deleted in M. avium and M. paratuberculosis, not present in C. diphtheriae and S. coelicolor) | Rv3867 | Unknown, annotated as part of MT3980 (Rv3866) in M. tuberculosis CDC1551 sequence with a frameshift (functional in M. leprae) |
|--------------------------------------------------|---------|---------------------------------------------------------------------|
| Rv3878 | Unknown, some similarity to PPE family, deleted with RD1 deletion region in M. bovis BCG (pseudogene in M. leprae) |
| Rv3879c | Unknown, repetitive, highly proline-rich N-terminus, deleted with RD1 deletion region in M. bovis BCG (pseudogene in M. leprae) |
| Rv3880c | Unknown (functional in M. leprae) |
| Rv3881c | Unknown (pseudogene in M. leprae) |

| Region 4 (not present in S. coelicolor) | Rv3446c | Unknown, may contain a possible ABC transporter signature (deleted in M. leprae) |

*Names of genes of these organisms were given arbitrarily by the authors of this paper. †Gene not identified by BLAST, data obtained from [1]. GenBank accession no. U34484 and AAC44033. ‡The gene is present in the sequence, but not annotated (name given arbitrarily by authors of this paper). ††Genes identified by BLAST as well as data obtained from GenBank, accession no. AJ250013. ‡‡Orthologs in S. coelicolor are equally similar to family G and H. ND, Not detected - not necessarily absent from genome but possibly not detected because of unfinished sequencing process. No duplication, no duplication of this gene is present in this region. No sequence data, no sequence data is available for this organism, published deletion information is included ([1] and others). Deleted, deleted from the genome of this particular species or strain ( # = deleted in only some strains of this species). Frame, frameshift. Stop, in-frame stop codon. Stop$, stop codon corresponds to stop codon in M. tuberculosis H37Rv, which splits gene into Rv3870 and Rv3871. Pseudo, confirmed pseudogene due to multiple frameshifts and stop codons.
Figure 1
Schematic representation of the genomic organization of the genes present in the five ESAT-6 gene cluster regions of *Mycobacterium tuberculosis* H37Rv as well as the regions in *C. diphtheriae* and *S. coelicolor*. ORFs are represented as blocked arrows showing the direction of transcription, with the different colors reflecting the specific gene family and the length of the arrow reflecting the relative lengths of the genes. Annotations of *M. tuberculosis* H37Rv genes are according to Cole et al. [13]. Black arrows indicate unconserved genes present in these regions. Gaps between genes do not represent physical gaps between genes on the genome, but have been inserted to aid in indicating conservation among gene positions. Gene families were named arbitrarily according to their position in *M. tuberculosis* H37Rv region 1. The regions were named after the numbering system of Brown et al. [19] used arbitrarily for the five mycosin (subtilisin-like serine protease) genes identified from these regions (family K). *M. tuberculosis* regions are shown in order of suggested duplication events (see phylogenetic results) and not by numbering. The results of the analyses of the primary features of these genes and their corresponding proteins are included in a short summary at the bottom of the figure (see also Table 2).

become pseudogenes as a result of extensive point mutations. This is in contrast to the genes from region 1 (which lies directly adjacent to region 2), which contains no pseudogenes. It is thus conceivable that these clusters should function as a unit, and that genes could become non-functional when part of the unit is disrupted. Furthermore, all the genes immediately flanking the putative functional regions, as well as five of the eight genes only present in one of the regions as depicted in Table 2 (the Rv1785c, Rv1786, Rv3878, Rv3879c and Rv3881c orthologs ML1542, ML1541, ML0046, ML0045 and ML0043), are probable pseudogenes, indicating that the genes present in the functional clusters are being maintained as a unit.

*M. avium* and *M. paratuberculosis*
The genomes of the *M. avium* strain 104 and the closely related species *M. paratuberculosis* (or *M. avium* subsp. *paratuberculosis*) has revealed four of the five ESAT-6 gene cluster regions (sharing between 65 and 75% similarity to *M. tuberculosis* H37Rv at protein level), with region 1 being absent in both species (Figure 4). Closer inspection of the gene sequence surrounding region 1 in both these species
has revealed a deletion of the region containing region 1 and some upstream flanking genes (from the Rv3861 gene ortholog up to and including the Rv3883c ortholog). This deletion coincided with the insertion of a \( \pm 2,292 \) bp sequence containing the genes for a putative hydroxylase (\( \pm 818 \) bp) and the sigl sigma factor (\( \pm 824 \) bp). The presence of this sequence in both genomes (99% DNA sequence identity) indicates that the insertion/deletion may have occurred before the divergence of the two species. The genes from the remaining ESAT-6 gene cluster regions that are present in \textit{M. avium} and \textit{M. paratuberculosis} contain no stop codons or frameshifts and thus appear to be functional.

\textbf{M. smegmatis}

The genome sequence of the avirulent, fast-growing mycobacterial species \textit{M. smegmatis} contains three of the five ESAT-6 gene cluster regions, namely regions 1, 3 and 4 (sharing between 60 and 75% similarity to \textit{M. tuberculosis} H37Rv at protein level), with regions 2 and 5 being absent (Figure 5). No deletions, frameshifts or stop codons were identified in any of the genes present in the regions 1, 3 and 4 and therefore it is concluded that these regions are functional.

\textbf{ESAT-6 gene cluster identification in bacteria other than the mycobacteria}

\textit{Corynebacterium diphtheriae}

The genome sequence of the closely related \textit{C. diphtheriae} has revealed a copy of the region 4 ESAT-6 gene cluster (Figure 1, see Table 3 for percentage similarity between sequences), situated in the same genomic location as in the mycobacteria (indicated by the large stretch of flanking genes homologous to the genes flanking region 4 in \textit{M. tuberculosis} H37Rv). All the genes present within this cluster appear to be fully functional, as no deletions, stop codons or frameshifts were identified. No duplications of the gene cluster could be detected in the genome of this organism.

\textbf{Streptomyces coelicolor}

The \textit{S. coelicolor} genome has revealed distinct orthologs for four of the six most conserved genes from the ESAT-6 gene cluster regions located in close proximity to each other (Figure 1). These genes show the highest similarity to the corresponding orthologs in region 4 of \textit{M. tuberculosis} (see Table 3 for percentage similarity between sequences). There is also a very distinct ortholog (SC3C3.02g) of the region 1 family I gene (Rv3876) in the \textit{S. coelicolor} region. There is no homolog for this gene in region 4 of \textit{M. tuberculosis}. A sequence-similarity search using the sequences of the other two proteins encoded in region 4, namely ESAT-6 (Rv3444c) and CFP-10 (Rv3445c), has also revealed some similarity to two small genes situated within the same region in the genome of \textit{S. coelicolor} (Table 3, Figure 1). These genes (SC3C3.10 and SC3C3.11) encode small proteins (124 and 103 amino acids) of unknown function, are very similar to each other, and lie adjacent to each other, similar to the observation for the \textit{esat-6/lhp} operon. The sequences of both these proteins also contain the motif W-X-G, a feature present in most of the ESAT-6 and CFP-10 proteins. The higher degree of similarity between the genes from region 4 of the mycobacteria (and \textit{C. diphtheriae}) and those present in the region in \textit{S. coelicolor} suggests that region 4 may be the ancestral region in the mycobacteria, although a number of differences between these regions do exist.

\textbf{Taxonomy}

It is evident from the taxonomy (Figure 6) of the different species of bacteria in which copies of the ESAT-6 gene clusters could be found, that the presence of these clusters appears to be a specific characteristic of the high G+C
Figure 3
Schematic representation of the genomic organization of the genes present in the five ESAT-6 gene cluster regions of *Mycobacterium leprae*. ORFs are represented as blocked arrows showing the direction of transcription, with the different colors reflecting the specific gene family and the length of the arrow reflecting the relative lengths of the genes as in Figure 1. Black arrows indicate unconserved genes present in these regions, while open arrows indicate pseudogenes. Annotations of *M. leprae* genes are according to Cole et al. [25].

Gram-positive Actinobacteria, and that multiple copies thereof are only found in the mycobacteria. No copies of the clusters could be found in the completed genome sequence of *Bacillus subtilis* and that of other related species, which also form part of the Firmicutes (Gram-positive bacteria), but fall under the *Bacillus/Clostridium* group (low G+C Gram-positive bacteria). No copies of these clusters could be found in the genomes of any other bacteria or organism outside of the Firmicutes and thus the ESAT-6 gene clusters appear to be unique to the Actinobacteria.

**Phylogeny of the ESAT-6 gene cluster**
To calculate the phylogenetic relationships between the five duplicated ESAT-6 gene cluster regions in *M. tuberculosis* and to identify the ancestral region, detailed phylogenetic analyses were performed on each of the six protein families present in all five of these regions (families C, D, G, H, J and K). Figure 7a shows a neighbor-joining tree of the protein sequences of the ATP/GTP-binding protein family (family D) from the ESAT-6 gene clusters of mycobacteria and *C. diphtheriae*, with the protein ortholog from *S. coelicolor* as the outgroup. This tree is representative of all six trees that were drawn using the six families (data for the other trees are not shown). To confirm the results obtained with the *S. coelicolor* orthologs as outgroups, the same analyses were done using the *C. diphtheriae* orthologs as outgroups, with comparable results (data not shown). This tree topology was not due to systematic error, as trees drawn using the FITCH algorithm gave the same results (data not shown). To confirm the basic structure of the trees and to verify that this structure is not influenced by the choice of outgroup, unrooted trees without any outgroup were constructed using the KITSCH algorithm, once again with comparable results (data not shown). To further verify the relationships among these clusters, the conserved sequences of all six proteins from *M. tuberculosis* were combined into one protein sequence and the same analysis performed (Figure 7b).

To investigate whether the non-conserved protein families (those that are not present in region 4 of the mycobacteria, *C. diphtheriae* or *S. coelicolor*) show the same basic phylogenetic relationships as the conserved families (present in all five regions), an analysis was done on the AAA+ class ATPase family (family B). This family does not have a homolog in region 4 and there is also no *C. diphtheriae* or *S. coelicolor* ortholog to use as outgroup. The tree constructed from the
Figure 4
Schematic representation of the genomic organization of the genes present in the four ESAT-6 gene cluster regions of *Mycobacterium avium* and *Mycobacterium paratuberculosis*, as well as the flanking genes of the region 1 deletion. ORFs are represented as blocked arrows showing the direction of transcription, with the different colors reflecting the specific gene family and the length of the arrow reflecting the relative lengths of the genes as in Figure 1. Black arrows indicate unconserved genes present in these regions. *M. avium* and *M. paratuberculosis* genes were arbitrarily annotated by the authors of this paper.

Data from this family clearly showed once again that regions 2 and 5, and region 1 and 3, respectively, are phylogenetically closer to each other (data not shown).

Neighbor joining, FITCH, KITSC and concatenated sequence comparison analyses all supported a single phylogeny that indicated that region 4 seems to be the most ancient of the mycobacterial ESAT-6 gene cluster regions. Region 4 is also the closest region to the *S. coelicolor* and *C. diphtheriae* regions. The order of duplication seems to extend from region 4, through 1 and 3 to regions 2 and 5. The phylogenetic relationships between corresponding clusters in the different mycobacteria are maintained throughout the different protein-family trees, and agree with the proposed phylogenetic order (or taxonomic position) of the mycobacterial species according to 16S rRNA data (see Figure 6).

As the genome of *M. tuberculosis* contains 11 copies of the esat-6/lhp gene pair that appears to be duplicated together, phylogenetic trees were constructed using the ESAT-6 or CFP-10 proteins separately (data not shown), or in combination as one ESAT-6/CFP-10 protein (Figure 7c). Using the combined *C. diphtheriae* ESAT-6/CFP-10 ortholog protein as outgroup, the same organization of duplication events was obtained with regions 1, 3, 2 and lastly 5 being duplicated from the ancient region 4. The other copies of the esat-6/lhp operon pair that are present in the *M. tuberculosis* genome sequence, but are not part of the ESAT-6 gene cluster regions, seem to have arisen from singular duplication events originating from different cluster regions. It is interesting to note that esat-6 and lhp from region 5 seem to be highly prone to duplication, as there are four additional copies of these two genes present in the genome, compared to just one additional copy originating from region 4 and region 3, respectively. These four gene duplicates of esat-6 and lhp from region 5 are also nearly identical (93-100% similarity at protein level), indicating their recent duplication.

**Discussion**
It was recently estimated in an *in silico* analysis of the genome sequence of *M. tuberculosis* H37Rv, that 52% of the proteome has been derived from gene duplication events [18]. One such involves the formation of multiple copies of the genes for the secreted T-cell antigens ESAT-6 and
CFP-10 [14,16,17] together with a number of associated genes. A total of twelve gene families were identified in five regions (which were termed the ESAT-6 loci).

Phylogenetic analyses of the protein sequences of the six most conserved gene families, present within the five regions, predict that region 4 (Rv3444c to Rv3450c) is the ancestral region. Region 4 also contains the least number of proteins (only 6 compared to the 12 of region 1 (Rv3866c-Rv3883c) and region 2 (Rv3884c-Rv3895c)), and does not contain the genes for PE and PPE, which may have been inserted into this region after the first duplication. Phylogenetic analyses using different methods and protein family data also suggests that subsequent duplications took place in the following order: region 1 (Rv3866c-Rv3883c) → 3 (Rv0282c-Rv0292c) → 2 (Rv3884c-Rv3895c) → 5 (Rv1782c-Rv1798c). Furthermore, these analyses support the taxonomic order observed for the mycobacteria, with M. smegmatis being taxonomically the farthest removed from M. tuberculosis. The presence of a copy of region 4 and its flanking genes in C. diphtheriae strengthens the taxonomic data that implies that the corynebacteria and mycobacteria have a common ancestor. It appears that C. diphtheriae diverged from the mycobacteria before the multiple duplications of the ESAT-6 gene cluster, as only one copy of this cluster could be identified in the genome of this organism.

The loss of region 1 from the genomes of the species M. avium and M. paratuberculosis (belonging to the M. avium complex) is confirmed by clinical data showing that patients seronegative for the human immunodeficiency virus (HIV) and infected with mycobacteria belonging to the M. avium complex do not respond to ESAT-6 from region 1, but do recognize purified protein derivative (PPD) and M. avium sensitins [22]. The genes for ESAT-6 and CFP-10 (esat-6 and lhp) in region 1 are also not found in M. bovis BCG and have thus been the focus of recent research because of their application as diagnostic markers to differentiate between BCG vaccination and M. tuberculosis, M. bovis or M. avium infection (see for example [17,23]). In this study we have found several copies of the ESAT-6 and CFP-10 genes (with differing degrees of similarity) in the genomes of different mycobacteria (80% and 71% protein sequence similarity for ESAT-6 and CFP-10 respectively from region 1 in avirulent M. smegmatis), as well as orthologs in species

| M. tuberculosis region 4 proteins | Family | Percentage similarity | C. diphtheriae | S. coelicolor |
|----------------------------------|--------|-----------------------|----------------|-------------|
| Rv3450c                          | C      | 47%                   | 36%            |             |
| Rv3447c                          | D      | 53%                   |                |             |
| Rv3445c                          | G      | 47%                   | 47 and 51%*    |             |
| Rv3444c                          | H      | 58%                   | 41 and 44%*    |             |
| Rv3448                           | J      | 33%                   | 45%            |             |
| Rv3449                           | K      | 49%                   | 45 and 47%     |             |

* Orthologs in S. coelicolor are equally similar to families G and H.
outside the mycobacteria; care should therefore be taken when using these proteins for diagnostic purposes. It will be important to look at the protein sequence similarity between the copies of ESAT-6 and CFP-10 of different virulent and environmental mycobacterial species before a member of these immunodominant protein families can be chosen as a definite marker of *M. tuberculosis* infection. Studies to determine the production of interferon-γ in response to exposure to ESAT-6 and CFP-10 from environmental mycobacteria (for example *M. smegmatis*) by peripheral blood mononuclear cells from infected patients have not been done. Until these results are available, indicating that the T-cell responses against these proteins are not comparable to those against the *M. tuberculosis* proteins, care should be taken with claims regarding the potential diagnostic value of these antigens.

Most of the sequences of the genes belonging to the ESAT-6 gene cluster regions contain no stop codons or frameshifts and thus appear to be functional. This is significant when placed in the context of a bacterium such as *M. leprae*, as it is hypothesized that the genome of *M. leprae* may contain the minimal gene set required by a pathogenic mycobacterium [5,24,25] and that the activities of some functional genes once present in the genome of *M. leprae* have been silenced (they became pseudogenes through multiple stop
Figure 7
Phylogenetic trees showing the relationships between the five duplicated gene cluster regions. (a) Neighbor-joining phylogenetic tree of all available protein sequences of the ATP/GTP-binding protein family (family D in Table 2) with the protein ortholog of *Streptomyces coelicolor* as the outgroup. This tree is representative of all the trees drawn using the six most conserved proteins in these regions as well as using the protein ortholog of *Corynebacterium diphtheriae* as the outgroup. (b) Neighbor-joining phylogenetic tree of all six conserved proteins from the *M. tuberculosis* gene clusters combined into one protein per region. The combined protein of *C. diphtheriae* was used as the outgroup. (c) Neighbor-joining phylogenetic tree of the ESAT-6 and CFP-10 protein families combined (family G and H), using the combined protein of *C. diphtheriae* as the outgroup.
codon mutations and frameshifts) because they are no longer needed for the bacterium’s intracellular survival [13]. It appears that \textit{M. leprae} contains at least two functional copies of the ESAT-6 gene cluster in its genome (regions 1 and 3). The \textit{M. leprae} ESAT-6 copy from region 1 (the L45-antigen or L-ESAT antigen from clone L45) was shown to be strongly reactive to sera from leprosy patients [26], providing experimental evidence that at least one of the cluster regions is functional in \textit{M. leprae}.

As most of the genes present within the ESAT-6 gene cluster regions encode proteins that are predicted to be associated with transport and energy-providing systems, we hypothesize that these proteins may be involved in the secretion of a substrate across the mycobacterial cell wall. It is well known that the T-cell antigens ESAT-6 and CFP-10 are found in short-term culture filtrates (ST-CF) of \textit{M. tuberculosis}, although the mechanism of secretion is unknown, as these proteins do not possess any of the usual Sec-dependent secretion signals [14-16]. It is therefore possible that the genes in the ESAT-6 gene cluster regions act together to provide a system for the secretion of ESAT-6 and CFP-10. There is evidence for the processing of the TB10.4 protein (the ESAT-6 family member from region 3) to a lower molecular weight product [27], suggesting a possible role for the cell-wall-associated mycosin proteases [19] in the suggested transport system. Most of region 1 is situated in the RD1 deletion region of \textit{M. bovis} BCG, possibly explaining the absence of expression of the mycosin-1 gene (Rv3883c) in BCG [19].

The hypothesis that an interdependent functional relationship exists between the proteins encoded in these regions is further supported by the \textit{M. leprae} sequence data, which shows that deletions of parts of the ESAT-6 gene cluster region 2 apparently caused the remaining genes in the region to become pseudogenes. Furthermore, Wards and co-workers [12] produced an \textit{M. bovis} knockout mutant of the ATPase gene Rv3871 (family D) in the ESAT-6 gene cluster region 1, resulting in a strain that did not sensitize guinea pigs to an ESAT-6 skin test. These results indicate a close relationship between the genes contained within these regions.

Wards \textit{et al.} [12] showed that an \textit{esat-6}/\textit{fp} knockout mutant of \textit{M. bovis} was less virulent than its parent if gross pathology, histopathology and mycobacterial culture from tissues were taken into account. These results, combined with the fact that multiple copies of the ESAT-6 gene clusters are found in all the mycobacteria, clearly indicate that they form an important part of the mycobacterial genome. The presence of multiple duplications of the ESAT-6 gene cluster regions in the mycobacteria may be a significant difference between the members of this genus and other high GC Gram-positives. Although the function of this cluster is presently unknown, there is sufficient evidence to indicate that it is of crucial importance to the mycobacteria and needs to be investigated further.

### Materials and methods

#### Genome sequence data and analyses

Annotations and descriptions of individual genes as well as gene and protein sequences of individual organisms were obtained from the publicly available finished and unfinished genome sequence databases listed in Table 1. Preliminary sequence data for \textit{M. tuberculosis} 210, \textit{M. avium} 104 and \textit{M. smegmatis} MC\textsuperscript{5} 155 was obtained from The Institute for Genomic Research (TIGR) website [28]. Preliminary sequence data for \textit{M. paratuberculosis} K10 was obtained from the University of Minnesota \textit{M. paratuberculosis} website [29]. Preliminary sequence data for \textit{M. bovis} AF2122/47 (spoligotype 9), \textit{C. diphtheriae} NCTC13129 and \textit{S. coelicolor} A\textsubscript{3} (2), was obtained from the Sanger Centre website [30]. All gene and protein sequences were subjected to analysis with the following programs to confirm annotation and to look for additional information: SignalP V2.0.02 [31,32], ClustalW WWW server at the European Bioinformatics Institute [33,34], TMHMM v0.1 [35,36], MOTIF [37] and BLASTP [38,39]. No data, progress report or BLAST search function is available for the genome sequencing of \textit{M. bovis} BCG Pasteur 1173P2 at the Pasteur Institute, but information concerning genome deletions was obtained from published data [1-3,5-7] and from the Pasteur Institute website [40].

#### Analyses of similar gene clusters

BLAST similarity searches [38], using the BLAST 2.0 program with tblastn and the BLOSUM-62 weight matrix, were used to identify stretches of DNA containing putative ORFs homologous to the genes found in the \textit{M. tuberculosis} ESAT-6 gene cluster regions from finished and unfinished genome sequences available at the National Center for Biotechnology Information (NCBI) website [41]. A total of 98 finished and unfinished genome sequences (35 from Gram-positive species) were used in the analysis, as summarized in Table 4. Where applicable, BLAST servers in database search services of individual sequencing centers were also used for protein identification. The Sanger Centre and The Institute for Genomic Research (TIGR) use the program WU-BLAST version 2.0 [42], while the University of Minnesota uses BLASTN with supplied defaults [43]. Sequences were only admitted to analysis when found to be part of one of the five gene clusters. In other words, no single homologous genes in the mycobacteria or other organisms (for example \textit{B. subtilis}) that did not form part of a similar gene cluster were considered for the analyses, to exclude any potential unassociated similarity that could lead to false positives.

Contig sequences corresponding to the gene clusters were obtained from their respective genome databases and used in further analyses. The Genetics Computer Group (Wisconsin Package Version 10.0, Genetics Computer Group (GCC), Madison, Wisconsin) program FRAMESEARCH was used to obtain whole sequence ORFs from the contigs. These ORFs were translated to protein sequences with the program
Table 4

Publicly available finished and unfinished genome sequence databases used in this study

| Species                  | Finished Genome Sequences | Unfinished Genome Sequences |
|--------------------------|---------------------------|-----------------------------|
| Acidithiobacillus ferrooxidans |                           |                             |
| Actinobacillus actinomycetemcomitans |                     |                             |
| Aquifex aeolicus          |                           |                             |
| Bacillus anthracis        |                           |                             |
| Bacillus halodurans       |                           |                             |
| Bacillus subtilis         |                           |                             |
| Bordetella bronchiseptica |                           |                             |
| Bordetella parapertussis  |                           |                             |
| Bordetella pertussis      |                           |                             |
| Borrelia burgdorferi      |                           |                             |
| Brucella melitensis biovar Suis |               |                             |
| Buchnera sp. APS          |                           |                             |
| Burkholderia mallei      |                           |                             |
| Burkholderia pseudomallei |                           |                             |
| Campylobacter jejuni NCTC 11168 |                 |                             |
| Carboxydothermus hydrogenoformans |           |                             |
| Caulobacter crescentus   |                           |                             |
| Chlamydia muridarum      |                           |                             |
| Chlamydia pneumoniae     |                           |                             |
| Chlamydia trachomatis D/UW-3/CX |               |                             |
| Chlamydophila pneumoniae AR39 |             |                             |
| Chlamydophila psittaci   |                           |                             |
| Chlorobium tepidum       |                           |                             |
| Clostridium acetobutylicum |                         |                             |
| Clostridium difficile    |                           |                             |
| Corynebacterium diphtheriae |                     |                             |
| Coviella burnetii        |                           |                             |
| Dehalococcoides ethenogenes |                   |                             |
| Desulfovibrio vulgaris   |                           |                             |
| Deinococcus radiodurans  |                           |                             |
| Escherichia coli K-12 MG1655 |                 |                             |
| Escherichia coli O157:H7 |                           |                             |
| Escherichia coli O157:H7 |                           |                             |
| Enterococcus faecalis    |                           |                             |
| Geobacter sulfurreducens |                           |                             |
| Haemophilus ducreyi      |                           |                             |
| Haemophilus influenzae Rd |                           |                             |
| Helicobacter pylori 26695 |                          |                             |
| Helicobacter pylori J99  |                           |                             |
| Klebsiella pneumoniae    |                           |                             |
| Lactococcus lactis subsp. lactis |           |                             |
| Legionella pneumophila   |                           |                             |
| Listeria monocytogenes   |                           |                             |
| Mesorhizobium loti       |                           |                             |
| Methylococcus capsulatus |                           |                             |
| Mycobacterium avium     |                           |                             |
| Mycobacterium avium subsp. paratuberculosis | |                       |
| Mycobacterium avium     |                           |                             |
| Mycobacterium leprae     |                           |                             |
| Mycobacterium smegmatis |                           |                             |
| Mycobacterium tuberculosis 210 |               |                             |
| Mycobacterium tuberculosis CDC1551 |           |                             |
| Mycobacterium tuberculosis H37Rv |           |                             |
| Mycoplasma genitalium G37 |                          |                             |
| Mycoplasma pneumoniae M129 |                        |                             |
| Neisseria gonorhoeae    |                           |                             |
| Neisseria meningitidis MC58 |                      |                             |
| Neisseria meningitidis Z2491 |                       |                             |
| Pasteurella multocida PM70 |                         |                             |
| Parphyromonas gingivalis WB3 |                     |                             |
| Pseudomonas aeruginosa   |                           |                             |
| Pseudomonas putida KT2440 |                          |                             |
| Pseudomonas putida PRS1  |                           |                             |
| Pseudomonas syringae pv. tomato |                  |                             |
| Rickettsia prowazekii    |                           |                             |
| Rhodobacter sphaeroides |                           |                             |
| Salmonella dublin        |                           |                             |
| Salmonella enteritidis   |                           |                             |
| Salmonella paratyphi     |                           |                             |
| Salmonella typhi         |                           |                             |
| Salmonella typhimurium LT2 |                        |                             |
| Shewanella putrescens    |                           |                             |
| Sinorhizobium meliloti   |                           |                             |
| Staphylococcus aureus CGL |                          |                             |
| Staphylococcus aureus MRSA |                        |                             |
| Staphylococcus aureus MSSA |                         |                             |
| Staphylococcus aureus Mu50 |                        |                             |
| Staphylococcus aureus N315 |                       |                             |
| Staphylococcus aureus NCTC 8325 |               |                             |
| Staphylococcus epidermidis |                      |                             |
| Streptococcus equi       |                           |                             |
| Streptococcus gordoni    |                           |                             |
| Streptococcus mutans     |                           |                             |
| Streptococcus pneumoniae |                           |                             |
| Streptococcus pyogenes   |                           |                             |
| Streptococcus pyogenes Manfredo |         |                             |
| Streptomyces coelicolor A3(2) |                |                             |
| Synechocystis PCC6803    |                           |                             |
| Thermotoga maritima      |                           |                             |
| Treponema denticola      |                           |                             |
| Treponema pallidum       |                           |                             |
| Ureaplasma urealyticum   |                           |                             |
| Vibrio cholera           |                           |                             |
| Wolbachia               |                           |                             |
| Xylella fastidiosa      |                           |                             |
| Yersinia enterocolitica  |                           |                             |
| Yersinia pestis          |                           |                             |

Finished genome sequences are indicated in bold, Gram-positive species are underlined.

TRANSLATE (also from GCG). All multiple sequence alignments and phylogenetic analyses were conducted on the protein level with these translated protein sequences.

Multiple sequence alignments

Multiple sequence alignments were performed on separate gene families belonging to the different clusters using ClustalW 1.5 [33] with the default parameters. The alignments were manually checked for errors and refined where appropriate. Multiple sequence alignments were also manually edited in some analyses during which unaligned regions (inserts) were removed (resulting in so-called edited alignments).

Phylogenetic trees

Bootstrapping resampling of the data sets were performed on the edited alignments, which generated 100 randomly chosen subsets of the multiple sequence alignment. Pairwise distances were determined with PROTDIST using the Dayhoff PAM matrix and neighbor-joining phylogenetic trees were calculated using NEIGHBOR (PHYLIP 3.5. [44]). In the case of each family of proteins, the C. diphtheriae sequence was first used as the outgroup after which the S. coelicolor sequence was used. Further phylogenetic analyses were performed using the programs FITCH and KITSCH with and without the outgroups respectively. A majority rule and strict consensus tree of all bootstrapped sequences were obtained using CONSENSE. The same analyses as described above were performed on a combined protein consisting of the edited aligned sequences of all six conserved proteins in these gene clusters as well as a combined protein constructed from the edited aligned sequences of all available ESAT-6 and CFP-10 family members. Finally, to confirm the results obtained with the single proteins, an analysis was performed with whole, unedited aligned sequences of the six most conserved proteins, using the program Paup 4.0b4a [45], during which negative branches were collapsed and 1,000 subsets were generated for bootstrapping resampling of the data. The consensus trees of all the above were drawn using the program Treeview 1.5 [46].

Acknowledgements

We are indebted to Rob Warren for continued advice, support, and critical reading of the manuscript. Sequencing of M. paratuberculosis was accomplished with support from USDA and Minnesota Agricultural
1. Mahairas GG, Sabo PJ, Hickey MJ, Singh DC, Stover CK: Analysis of the proteome of Mycobacterium tuberculosis in silico. Tuberc Lung Dis 1999, 79:329-342.

2. Philipp WJ, Nair S, Guglielmi G, Lagranderie M, Gicquel B, Cole ST: Toward the development of diagnostic assays to discriminate between Mycobacterium bovis Infection and Bacille Calmette-Guerin vaccination in cattle. Clin Infect Dis 2000, 30:124-133.

3. Sanger Centre [http://www.sanger.ac.uk/Projects/M_tuberculosis/]

4. University of Minnesota M. paratuberculosis [http://www.ncbi.nlm.nih.gov/Microb_blast/unfinishedgenome.html]

5. NCBI BLAST Server [http://blast.ncbi.nlm.nih.gov/Blast.cgi?Jform=0]

6. Yeast [http://www.ncbi.nlm.nih.gov/blast/blast.cgi?Jform=0]

7. Pasteur Institute [http://www.pasteur.fr/recherche/unites/Lgbm/Deletion.html]

8. NCBI BLAST Server [http://www.ncbi.nlm.nih.gov/blast/blast.cgi]

9. BLASTP [http://www.ncbi.nlm.nih.gov/blast/blast.cgi?Form=0]

10. Pasteur Institute [http://www.pasteur.fr/recherche/unites/Lgbm/Deletion.html]

11. NCBI BLAST Server [http://www.ncbi.nlm.nih.gov/blast/blast.cgi]

12. WU-BLAST version 2.0 [http://www.ncbi.nlm.nih.gov/blast/wustl/]

13. University of Minnesota BLAST Server [http://www.ncbi.nlm.nih.gov/blast/blast.cgi]

14. PHYLIP - Phylogeny Inference Package (Version 3.6) [http://evolution.genetics.washington.edu/phylip/phylip.html]

15. TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 1996, 12:357-358.

16. The Sanger Centre: M. tuberculosis Genome Project [http://www.sanger.ac.uk/Projects/M_tuberculosis/]

17. Reference publication: D. S. Wilson [http://www.ncbi.nlm.nih.gov/Genbank/]

18. A. P. P. and A. M. W. [http://www.ncbi.nlm.nih.gov/Genbank/]

19. Mycobacterium tuberculosis [http://www.ncbi.nlm.nih.gov/]
48. TIGR Mycobacterium tuberculosis CDC1551 Information
   [http://www.tigr.org/tigr-scripts/CMR2/GenomePage3.spl?database=gmc]
49. TIGR BLAST Search Engine for Unfinished Microbial Genomes
   [http://www.tigr.org/cgi-bin/BlastSearch/blast.cgi]
50. The Sanger Centre: M. bovis Genome Project
    [http://www.sanger.ac.uk/Projects/M_bovis/]
51. Pasteur Institute Mycogenomics
    [http://www.pasteur.fr/recherche/unites/Lgmb/mycogenomics.html]
52. The Sanger Centre: Mycobacterium leprae genome project
    [http://www.sanger.ac.uk/Projects/M_leprae/]
53. Pasteur Institute Leproma
    [http://genolist.pasteur.fr/Leproma]
54. The Sanger Centre: C. diphtheriae Genome Project
    [http://www.sanger.ac.uk/Projects/C_diphtheriae/]
55. The Sanger Centre: S. coelicolor Genome Project
    [http://www.sanger.ac.uk/Projects/S_coelicolor/]
56. Shinnick TM, Good RC: Mycobacterial taxonomy. Eur J Clin Microbial Infect Dis 1994, 13:884-901.