A novel polymorphic genetic locus in members of the *Mycobacterium tuberculosis* complex

Jean Rauzier,1 Eamonn Gormley,2 M. Cristina Gutierrez,3 Eric Kassa-Kelembho,1,4 Laurie J. Sandall,2 Chris Dupont,2 Brigitte Gicquel1 and Alan Murray1,2

**INTRODUCTION**

The *Mycobacterium tuberculosis* complex comprises closely related species (*M. tuberculosis*, *M. bovis*, *M. africanum*, *M. microti* and the *M. bovis* BCG vaccine strains) that are difficult to differentiate. Polymorphic genetic regions have been identified between these species and used to differentiate *M. tuberculosis* and *M. bovis* (Rodriguez et al., 1995). In addition, several polymorphic regions are also used to differentiate strains within a species to follow transmission in various settings using molecular epidemiology methods (Yang et al., 1995a, b).

We have recently demonstrated that the *Mycobacterium paratuberculosis* P<sub>AN</sub> promoter region can be used as a probe to identify the major pathogenic groups of mycobacteria (Gormley et al., 1997). In a study on a limited number of strains from the *M. tuberculosis* complex...
complex, the $P_{AN}$ probe hybridized to the same sized EcoRI fragment in $M$. tuberculosis and $M$. bovis BCG, and to a different sized EcoRI fragment in $M$. bovis.

The $P_{AN}$ probe also detected RFLPs in other pathogenic mycobacteria suggesting that the target sequence was located in or very close to a region that contained genetic differences in these species. In $M$. paratuberculosis, the $P_{AN}$ promoter lies adjacent to, and outside, the 3' end of one copy of the IS900 (Murray et al., 1992). $P_{AN}$ is able to drive expression of an ORF (ORF2) within the IS900 located in or very close to a region that contained genetic differences from the IS900. Since the IS900 is unique to $M$. paratuberculosis, the ORF adjacent to $P_{AN}$ in $M$. paratuberculosis is clearly different to the ORF adjacent to the $P_{AN}$ equivalent in the $M$. tuberculosis complex.

This study was initiated to characterize the counterpart of the $P_{AN}$-IS900 genetic region of $M$. paratuberculosis in members of the $M$. tuberculosis complex and to determine the nature of the RFLP detected by $P_{AN}$. Bacteriophage lambda libraries of $M$. bovis and $M$. bovis BCG were screened with $P_{AN}$ and the hybridizing fragments, and flanking regions subcloned and sequenced. The equivalent $M$. tuberculosis sequence was obtained through the $M$. tuberculosis genome sequencing project and compared. Analysis of the nucleotide sequence is presented; it was found that the observed RFLP that distinguished a wild-type $M$. bovis strain from $M$. tuberculosis and $M$. bovis BCG was generated by a 2.4 kb fragment that was present in $M$. tuberculosis and $M$. bovis BCG but absent in many $M$. bovis strains.

**METHODS**

**Mycobacterial strains.** The distribution of the 2405 bp fragment in DNA from 106 strains of the $M$. tuberculosis complex and 26 atypical mycobacteria was investigated. The former included 27 strains of $M$. tuberculosis isolated from patients in Paris hospitals, 43 $M$. tuberculosis strains isolated from patients in the Central African Republic, 25 $M$. bovis strains isolated from animals in France, 2 $M$. africanum strains, 2 $M$. microti strains and 7 $M$. bovis BCG daughter strains (Montreal, Russian, Japanese 172, Prague, Glaxo 1077, Danish 1313 and Pasteur 1173P2). The strains were identified by standard phenotypical characteristics (David et al., 1989). Specific features for $M$. bovis were: slowly growing mycobacteria, nitrate-negative, niacin-negative, sensitive to 2-thiophene-carboxylic acid hydrazide and resistant to pyrazinamide. Genetic confirmation of identification was based on the absence of the mtb-40 sequence in the chromosomal DNA from the strains (Del Portillo et al., 1991) and on the presence of an allele-specific polymorphism at nt 285 of the oxyR gene (Espinosa de los Monteros et al., 1998) which differentiates $M$. bovis (adenine) from the other members of the $M$. tuberculosis complex (guanine). The atypical mycobacteria included were: $M$. asiaticum, $M$. aewum, $M$. avium, $M$. brumae, $M$. chelonae, $M$. flavescens, $M$. fortuitum, $M$. gastro, $M$. gordoniae, $M$. haemophilum, $M$. intracellularare, $M$. kansasi, $M$. malmoense, $M$. marinum, $M$. paratuberculosis, $M$. porcine, $M$. rhodiniae, $M$. scrofulaceum, $M$. senegalense, $M$. simiae, $M$. smeagmatis, $M$. szulgai, $M$. terrae, $M$. triviale, $M$. vaccae and $M$. xenopi.

**Construction and screening of genomic libraries to isolate the $P_{AN}$ equivalent genetic region from $M$. bovis and BCG.** Chromosomal DNA from $M$. bovis (KML, New Zealand field isolate) and BCG (Pasteur, ATCC 35734) was partially digested with EcoRI and size-fractionated on a 1% agarose gel. Fragments in the size range 10–23 kb were cloned into the bacteriophage lambda DASH vector (Stratagene). The titre of the libraries after amplification was estimated to be 10^9 p.f.u. ml⁻¹. Phage plaques (approx. 300 per plate) were transferred to Hybond-N nylon membrane (Amersham) for screening with $P_{AN}$. The $P_{AN}$ sequence (Gormley et al., 1997) was labelled with ³²P using the Radprime labelling protocol (BRL) and used as a probe at high stringency using the Multiprime rapid hybridization system (Amersham). Bacteriophage clones that hybridized with $P_{AN}$ were selected and digested with Sall to obtain smaller fragments, which were separated on a 1% agarose gel. Following Southern blotting, the fragments were probed with ³²P-labelled $P_{AN}$ and those that gave a positive signal were purified using the GeneClean 11 kit (Bio101) and subcloned into pUC18 for sequencing.

**DNA sequencing and analysis.** Sequencing was carried out using the dye deoxy chain-termination method with the Big Dye Terminator Cycle Sequencing kit (PE Applied Biosystems) on a GeneAmp PCR System 9600 (Perkin Elmer), and run on a DNA analysis system model 373 (Applied Biosystems). The sequencing strategy involved using universal, forward and reverse primers to get preliminary data on the SalI fragments and then designing primers to newly acquired sequence. In addition, the larger inserts were partially digested with Sau3A, the DNA purified using a GeneClean 11 kit, the fragments shotgun-cloned into pUC18 and sequenced using forward and reverse primers. The sequence was assembled and processed using DNA Strider (CEA, France) and the University of Wisconsin Genetics Computer Group (GCG) packages. The BLAST algorithm (BLASTN and BLASTX) (Alschul et al., 1990) was used to search for sequence similarities. DNA Strider and the GCG package were also used for secondary structure work.

**PCR conditions and primers.** PCR amplifications using primers AM9 (5'-CGGAGGCCACCGTGATCGACGG-3') and AM10 (5'-GTTTCGACCCGCAGGCGGCAGGG-3'), which flank the region containing the 2405 bp fragment, were performed using the Expand Long Template PCR System (Boehringer Mannheim) in a final volume of 50 µl in buffer 3 containing 2.25 mM MgCl₂ and detergents plus 10% (w/v) DMSO, 0.04 mM of each dNTP, 25 pmol of each primer and 2.5 U enzyme. The PCR conditions consisted of one cycle of denaturation (95 °C, 2 min), followed by 10 cycles of amplification consisting of denaturation (95 °C, 30 s), annealing (62 °C, 30 s) and primer extension (68 °C, 150 s), followed by 20 cycles of 95 °C for 30 s, 62 °C for 30 s, 68 °C for 150 s plus an extension of 20 s per cycle, and final extension at 68 °C for 7 min. PCR amplifications using primers AM30 (5'-GAGCA- CGACGTGATGAAAGGC-3') and AM31 (5'-AGTGCGCATC- GAACGGACATTG-3'), which bind within the 2405 bp fragment, were performed using 2.5 U AmpliTaq (Perkin Elmer) with 1× the supplied buffer in a final volume of 50 µl containing 1.5 mM MgCl₂, 10% DMSO, 0.2 mM (of each) deoxyoligonucleotide triphosphate and 20 pmol of each primer. Initial denaturation was achieved by 1 cycle at 94 °C for 5 min and amplification by 30 cycles at 94 °C for 1 min, 65 °C for 1 min and 72 °C for 1 min. This was followed by a final extension step of 72 °C for 10 min.

**Southern blot analysis.** Chromosomal DNA isolated from $M$. tuberculosis and $M$. bovis was digested with Poulf, electrophoresed on a 1% agarose gel and transferred onto nylon.
membrane (Hybond-N+; Amersham). Probe labelling, hybridization and signal detection were carried out with the Enhanced Chemiluminescence System (Amersham) according to the manufacturer's instructions. The probes were: (1) a 434 bp amplification product from an M. bovis strain lacking the 2405 bp, obtained using the primers AM9/AM10 (external probe); (2) an 800 bp segment in the 2405 bp fragment that does not include the Puvall site of the fragment (internal probe).

RESULTS

DNA sequence analysis of the P_AN-binding region

DNA fragments from M. bovis and M. bovis BCG that contained the genetic region similar to the P_AN promoter region of M. paratuberculosis were isolated from phage lambda libraries of each species. The segments similar to the P_AN promoter were termed P_AN regions. Restriction fragments suitable for sequencing were prepared and subcloned into pUC18. The DNA sequence of the fragments was determined for the two species on both strands using universal primers and primers designed to the newly generated sequence. The resulting sequence from M. bovis was 6559 bp in length and the BCG fragment was 8964 bp in length. Comparison of the two sequences by computer alignment showed that they were identical at the 5' and 3' ends; however, within the BCG sequence (at nt 1896) was a 2405 bp segment that was absent in M. bovis (Fig. 1). The G + C content of the 2405 bp segment was 62.6 mol% compared to a mean of 66.05 mol% for the 6.5 kb of flanking sequence.

P_AN promoter sequence

The P_AN regions of M. bovis and M. bovis BCG (nt 781–953) were shown to be identical. To determine the similarity of the M. bovis/BCG P_AN with the counterpart sequence in M. tuberculosis, a search of the Sanger database was performed. It was found that P_AN in M. tuberculosis is identical to that in BCG and M. bovis (100% identity was found to be a fragment on the M. tuberculosis cosmid SCY15F10). Comparison of the M. bovis/BCG P_AN sequence with P_AN from M. paratuberculosis showed 70% similarity (Fig. 2). The Shine–Dalgarno sequence, transcriptional start site (+1) and the −10 and −35 promoter regions are conserved. In M. paratuberculosis, the initiation codon is ATG separated by 6 nt from the ribosome-binding site and in M. bovis and BCG, GTG could be an initiation codon. A putative 321 bp ORF was identified downstream from P_AN in M. bovis and BCG. The predicted 107 amino acid sequence from this ORF showed no significant similarity with any other sequence in the databases.

Characteristics of the BCG 2405 bp fragment

To determine if the 2405 bp fragment was present in the M. tuberculosis genome, a BLAST search of the Sanger database (a laboratory-adapted virulent strain, M. tuberculosis H37Rv) and TIGR database (a virulent clinical isolate of M. tuberculosis; Valway et al., 1998) was performed. Alignment was obtained to Sanger cosmid SCY15F10 (99–100% identity in four segments) and to the TIGR contig 7741 (91–100% identity in four segments), indicating the presence of the fragment in two M. tuberculosis strains. The fragment did not contain any sequences with similarity to known transposable elements. Computer programs were used to identify ORFs within the fragment. Three ORFs in three frames were found as follows: in-frame 0, Orf7 (221 codons), which continues outside of the fragment; frame +1, Orf6 (246 codons); and frame +2, Orf5 (207 codons), which starts outside the fragment and ends within it (Fig. 3a). Orf5 shows 100% identity to the hypothetical M. tuberculosis protein Rv3889c which is of unknown function and Orf7 shows 100% identity to Rv3887c, which also is of unknown function but its hydrophilic and hydrophobic domains suggest that it is a probable membrane protein (Cole et al., 1998). No significant ORFs were found on the complementary strand. Sequence comparisons using the BLAST and FASTA algorithms showed that Orf6 had some similarity to an ATP-binding protein. Alignment of Orf6 with the Methanococcus jannaschii cell-division-inhibitor homologue and the Treponema pallidum putative ATP-binding protein shows a region of similarity, SGKGG, which is a consensus sequence for binding to ATP (Fig. 4). Orf6 possesses an N-terminal hydrophobic segment consistent with a lipoprotein signal sequence. Orf6 also has similarity to part of the 1D sequence located in the RD1 region of M. bovis and M. tuberculosis (Mahairas et al., 1996). RD1 is a 9.5 kb DNA segment that is found in virulent laboratory and clinical isolates of M. bovis and M. tuberculosis. A search for repetitive or inverted repeat sequences within the 2405 bp fragment identified a 29 bp palindromic sequence (pal3) in the 3' region of BCG.
the fragment that was close to a series of eight 10 bp sequences with some similarity to major polymorphic tandem repeat (MPTR) sequences (Hermans et al., 1992). The eight MPTR-like sequences shared between 5 and 8 bp homology to palindromic sequence (pal) and a 20 bp sequence (pa12) flanking the fragment were identified. Flanking sequences of the 2405 bp fragment are shown in Fig. 3(a).

Flanking sequences
A search for sequences with the potential for forming secondary structures that may provide a mechanism for the deletion of the 2405 bp fragment from M. bovis was undertaken using computer programs. Two imperfect palindromes flanking the fragment were identified. A 43 bp sequence (pal1) 238 nt 5' of the 2405 bp fragment and a 26 bp sequence (pal2) 79 nt 3' to the fragment were identified (Fig. 3b). No repetition consistent with known deletion mechanisms was found. ORFs surrounding the 2405 bp fragment are shown in Fig. 3(a). Orf8 shows 100% identity to Rv3886c, a probable serine protease of M. tuberculosis (Cole et al., 1998). Orf8 also shows similarity to a serine protease of Streptomyces coelicolor (U33176) (Siezen et al., 1991) and Rv3883c, a hypothetical protein of M. tuberculosis (that shows weak similarity to a variety of proteases) and ORF1H (nt 16900-15530) from the RD1 region of M. tuberculosis (Mahairs et al., 1996). Note that the similarity of Orf8 to the sequence in RD1 is to ORF1H and not ORF1J, the predicted serine protease described by Mahairs et al. (1996). Alignment of Orf8, Rv3883c and U33176 shows a segment of 86 codons in Orf8 that is absent in Rv3883c and the serine protease of S. coelicolor (data not shown). This sequence does not have any characteristics of an intron. No significant similarity was found with any of the other ORFs flanking the 2405 bp fragment.

Genetic localization of the sequence
The complete BCG sequence described in this work including the 2405 bp fragment can be aligned with...
A novel polymorphic locus in the TB complex

Fig. 4. Alignment of Orf6 with c64368 (a cell-division-inhibitor homologue of Methanococcus jannaschii) and tpu36839 (a putative ATP-binding protein of Treponema pallidum). The consensus for ATP binding is shown in bold and the putative lipoprotein signal sequence is underlined.

Fig. 5. Alignment of the ORF map of M. bovis BCG to the equivalent region of the M. tuberculosis genome including the ORFs described by Cole et al. (1998), showing the close proximity of the RD1 region.

the sequence of cosmid SCY1SF10 from the Sanger database (http://www.sanger.ac.uk:80/Projects/M._tuberculosis/).

Alignment of SCY1SF10 with the 3' end of the cosmid containing RD1 (including the sequence of the putative serine protease) showed the close proximity of the 3' end of our sequence to the RD1 segment (Fig. 5).

Distribution of the 2405 bp fragment in mycobacterial strains

The presence or absence of the 2405 bp fragment in a range of mycobacterial strains was investigated with PCR using flanking primers AM9/AM10 and internal primers AM30/AM31. The PCR with AM30/AM31 using DNA from 70 M. tuberculosis strains, 7 M. bovis BCG daughter strains, 2 M. africanum strains and 2 M. microti strains gave a 257 bp product (data not shown). In contrast to our preliminary results with M. bovis strain KML (which does not possess the 2405 bp fragment), we observed that primers AM30/AM31 gave a 257 bp product with DNA from 7 out of 25 of the M. bovis strains. The remaining 18 M. bovis strains gave no product (data not shown). To confirm the absence of the 2405 bp fragment, the PCR was repeated using primers AM9/AM10, which give a 2839 bp product (large segment) when the 2405 bp fragment is present and a 434 bp product (small segment) when it is absent. All of the M. tuberculosis strains and the 7 M. bovis strains that gave a product with AM30/AM31 gave the large segment and the 18 M. bovis strains that gave no product with AM30/AM31 gave the small segment (data not shown).

As a further check on the validity of the PCR results, Southern blots of PvuII-digested DNA from M. tuberculosis and M. bovis were hybridized with two DNA probes. With the external probe (Fig. 6a) two patterns were obtained. Strains carrying the 2405 fragment showed two bands as a consequence of the presence of a PvuII site at position 2243 within this fragment. Strains without the 2405 fragment gave only one band, which corresponds to a PvuII fragment. The internal probe (Fig. 6b) hybridized only with the strains carrying the fragment. The results were entirely consistent with the PCR data.

The distribution of the 2405 bp fragment in mycobacterial species other than the M. tuberculosis complex was also investigated by PCR using primers AM30/AM31. The DNA from 26 different mycobacterial
species was tested and the 257 bp fragment failed to amplify any of them, suggesting that they did not possess the 2405 bp fragment.

DISCUSSION

The study of polymorphic regions in closely related organisms has generated useful information for epidemiology studies. It can also provide information on the genetic differences between related pathogenic and non-pathogenic organisms. This can lead to a better understanding of the genes required for virulence and prospective genetic targets for drug and vaccine development. In this study, we have used the \( P_{AN} \) promoter from \( M. \) \textit{paratuberculosis} as a probe to identify a polymorphic region in \( M. \) \textit{biovius} that is close to the \( P_{AN} \) region.

In the first instance, we used the \( P_{AN} \) promoter from \( M. \) \textit{paratuberculosis} to identify similar sequences in the equivalent regions of \( M. \) \textit{biovius}, \( M. \) \textit{biovius} BCG and \( M. \) \textit{tuberculosis}. It was found that the \( P_{AN} \) region in \( M. \) \textit{biovius}, BCG Pasteur and \( M. \) \textit{tuberculosis} is identical and overall there is a 70% similarity to the \( M. \) \textit{paratuberculosis} \( P_{AN} \) sequence. Preliminary expression studies with the \( P_{AN} \) promoter from BCG indicate that the transcriptional start site is in the same position as in \( P_{AN} \) from \( M. \) \textit{paratuberculosis} (data not shown). However, the distance to the putative GTG start codon in BCG is 50 nt compared with 40 nt to the ATG start codon in \( M. \) \textit{paratuberculosis}. Previous work has shown \( P_{AN} \) from \( M. \) \textit{paratuberculosis} to be a functional promoter on plasmids cloned into BCG and \( M. \) \textit{smeqmatiis} (Murray \textit{et al.}, 1992), and the relative strength of activity is influenced by the host (Timm \textit{et al.}, 1994).

To investigate the nature of the RFLP observed between \( M. \) \textit{biovius} and BCG when \( P_{AN} \) was used to probe EcoRI digests of genomic DNA (Gormley \textit{et al.}, 1997), we sequenced the flanking region of \( P_{AN} \) from these two species. It was discovered that less than 1 kb downstream of \( P_{AN} \) in BCG was a 2405 bp fragment that was absent in \( M. \) \textit{biovius} KML. A restriction map of the fragment revealed an internal EcoRI site that would generate a different sized \( P_{AN} \)-binding fragment in BCG compared with \( M. \) \textit{biovius}, and thus generate the RFLP (see Fig. 1 for the location of the EcoRI site). This banding pattern was observed with three New Zealand \( M. \) \textit{biovius} field strains isolated from cattle with tuberculosis and three \( M. \) \textit{biovius} ATCC strains (Gormley \textit{et al.}, 1997). To investigate this observation further, we screened 25 \( M. \) \textit{biovius} strains isolated from animals in France. Unexpectedly, it was found that 7 out of 25 (28%) of these strains possessed the fragment. Thus, absence of the fragment is not a universal feature of \( M. \) \textit{biovius} strains. The original BCG vaccine strain developed by Calmette and Guerin between 1908 and 1921 was presumably derived from an \( M. \) \textit{biovius} isolate that possessed the fragment.

Sequence analysis of the 2405 bp fragment and its flanking region did not provide any clues to explain its polymorphic nature. It does not possess any characteristics of known transposable elements and the flanking sequences do not have any obvious features that suggest a deletion mechanism. The significance of the three palindromic sequences (pal1 and pal2 some distance outside the fragment, and pal3 within the fragment) is unknown. Within the fragment at the 3' end (in Orf7) there is a series of eight MPTR-like sequences showing 5–8 bp homology to the consensus MPTR sequence 5'-GCCGGTGTGG (Hermans \textit{et al.}, 1992). However, unlike the sequence described by Hermans \textit{et al.} (1992), Orf7 does not appear to belong to the PPE protein family, which is characterized by the presence of multiple, tandem copies of the motif Asn-X-Gly-X-Gly-Asn-X-Gly (Cole \textit{et al.}, 1998). A BLAST search with Orf7 shows 100% identity to the \( M. \) \textit{tuberculosis} hypothetical protein Rv3887c, with no similarity to any of the described PPE proteins (data not shown). It should also be noted that unlike previously described MPTR sequences in \( M. \) \textit{tuberculosis}, the majority of the repeats are not separated by a 5 bp spacer segment. The significance (if any) of the MPTR-like sequences is unknown. There is no evidence to suggest that these sequences are involved in genetic reorganization in this region, but it is interesting to note that MPTR sequences.
do have significant homology to the *Escherichia coli* and bacteriophage lambda DNA sequence chi, which is a prokaryotic signal for homologous recombination (Smith et al., 1981). Indeed, the chromosomal location of the fragment does appear to be susceptible to deletions. The 240.5 bp fragment is in close proximity to the RD1 region, which is absent in all BCG substrains (Mahairas et al., 1996).

This is the first report of a genetic region of *M. bovis* BCG that is not universally present in *M. bovis* strains. Organisms that possess the fragment are members of the *M. tuberculosis* complex and the polymorphic nature of this region in *M. bovis* will provide an additional genetic marker for strain differentiation within this species. A more extensive investigation of the distribution of the fragment in the *M. tuberculosis* complex using isolates from different geographical locations is under way.

ACKNOWLEDGEMENTS

The financial support of the Agricultural and Marketing Research and Development Trust of New Zealand (AGMARDT) is gratefully acknowledged. M. C. G. was the recipient of a postdoctoral fellowship from the Spanish Ministry of Education and Culture.

REFERENCES


Received 18 September 1998; revised 23 February 1999; accepted 12 March 1999.