Short Communication

A fragment of 21 ORFs around the direct repeat (DR) region of *Mycobacterium tuberculosis* is absent from the other sequenced mycobacterial genomes: implications for the evolution of the DR region

Karina Caimi and Angel Cataldi*

Institute of Biotechnology, CICVyA-INTA Castelar, Argentina

*Correspondence to:
Angel Cataldi, Instituto de Biotecnología, CICVyA/INTA, Los Reseros y Las Cabañas, 1712 Castelar, Argentina.
E-mail: acataldi@cnia.inta.gov.ar

Received: 24 July 2003
Revised: 15 December 2003
Accepted: 22 December 2003

Abstract

The direct repeat (DR) region is a singular locus of the *Mycobacterium tuberculosis* complex genome. This region consists of 36 bp repetitive sequences separated by non-repetitive unique spacer sequences. Around this region there are several genes coding for proteins of unknown function. To determine whether the *M. smegmatis*, *M. avium*, *M. marinum* and *M. leprae* genomes contain sequences and ORFs similar to those of the DR locus of the *M. tuberculosis* complex, we analysed the corresponding regions in these species. As a first step, some conserved genes that flank the DR genes \[Rv2785c (rpsO), Rv2786c (ribF), Rv2790c (ltp1), Rv2793c (truB), Rv2800, Rv2825, Rv2828, Rv2831 (echA16), Rv2838 (rbfA) and Rv2845 (proS)\] were used as markers to locate the corresponding orthologues in *M. smegmatis*, *M. avium*, *M. marinum* and *M. leprae* in silico. Most of these *M. tuberculosis* marker genes have highly similar orthologues located in the same order and orientation in the other mycobacteria. In contrast, no orthologues were found for ORFs *Rv2801–Rv2824*, suggesting that these genes are unique to *M. tuberculosis* within the genus *Mycobacterium*. We observed that in *M. smegmatis* and *M. avium*, *Rv2800* and *Rv2825* are adjacent. This observation was experimentally confirmed by PCR. In conclusion, as the DR locus and the ORFs around it are absent in *M. smegmatis* and *M. avium* and, as it is possible that these species are older than *M. tuberculosis*, we postulated that the DR locus was acquired by the *M. tuberculosis* complex species or by an ancestor bacterium. Copyright © 2004 John Wiley & Sons, Ltd.

Keywords: *Mycobacterium*; DR region; insertion/deletion; tuberculosis

Introduction

The direct repeat (DR) region of the *Mycobacterium tuberculosis* complex consists of 36 bp repetitive sequences separated by non-repetitive unique spacer sequences whose lengths vary from 27 to 41 bp (Hermans *et al.*, 1991; Kamerbeek *et al.*, 1997; van Embden *et al.*, 2000; Caimi *et al.*, 2001). *M. tuberculosis* complex isolates are polymorphic (Fang *et al.*, 1998; Groenen *et al.*, 1993) at the DR region because they can differ in the presence or absence of the different spacers and adjacent DRs. In spoligotyping, a technique to distinguish isolates, the DR region is amplified by PCR using DR-specific biotinylated primers. The presence or absence of each of the spacers is determined by hybridizing the amplified DNA to a membrane containing oligonucleotides corresponding to each of these spacer sequences (Kamerbeek *et al.*, 1997; Sola *et al.*, 1999; van Soolingen *et al.*, 1993).
The function of the DR locus in *M. tuberculosis* is presently unknown, but its omnipresence in the *M. tuberculosis* complex and the conservation in many different isolates of the sequence and order of the DRs and spacers suggest that the region could have a biological function. The DR region in *M. tuberculosis* is a hot spot of integration of the IS6110 insertion element (Fang et al., 1998; Hermans et al., 1991). The polymorphisms in the DR associated with IS6110 insertions have been described in different strains, including some outbreak isolates (Fang et al., 1998; Groenen et al., 1993). The different species of the *M. tuberculosis* complex (*M. tuberculosis*, *M. bovis*, *M. africanum*, *M. microti* and *M. canetti*) show distinctive characteristics, such as the presence and absence of particular spacers (Caimi et al., 2001; van der Zanden et al., 2002). Interestingly, there is low homology between the spacers (with the exception of spacers 40 and 85), and they contain a huge amount of stop codons (unpublished observations), suggesting that they do not derive from a gene interrupted by the insertions of the 36 bp direct repeats. Notably, a unique set of spacers is present in *M. canetti*. No significant homology has been reported between the *M. tuberculosis* DR sequence and DNA from any other bacteria. After the recent sequencing of various bacterial genomes, it has been postulated that the DR sequence of *M. tuberculosis* belongs to the clustered regularly interspaced short palindromic repeats (CRISPR) family of repetitive DNA (Jansen et al., 2002). In spite of this important advance in the understanding of the ubiquity and structure of CRISPR-type sequences, their biological function, if any, remains unclear (Mojica et al., 1995). In consequence, the DR region remains as an enigma of the *M. tuberculosis* complex genome.

In this work we have analysed the genes surrounding the *M. tuberculosis* DR region in the *M. smegmatis*, *M. avium*, *M. marinum* and *M. leprae* genomes. The goal was to determine whether the equivalent regions have a similar ORF organization and to find any resemblances to the DR sequences.

**Material and methods**

**BLAST searches**

Amino acid sequence alignments were made by searching public databases via the Internet. For *M. smegmatis* and *M. avium* the BLASTp search facility at the TIGR Unfinished Microbial Genomes site (http://www.tigr.org/tdb/mdb/mdbinprogress.html) was used. The parameters used were: program, BLASTp; alignments, 5; matrix, blosum62; cut-off, none; strand, default expect, 0.1; description, 5; and filter, none. At the Pasteur Institute sites, BLASTp searches were performed for *M. leprae* (http://genolist.pasteur.fr/Leproma/) using the NCBI BLAST version 2 algorithm, and for *M. ulcerans* (http://genopole.pasteur.fr/MulceBururiList.html) using the parameters: −F; F; −e, 1e−3; −q, −1; −b, 20; −v, 20; and −S, 3. Finally, TBLASTn searches were performed at the Sanger Centre sites for *M. marinum* (http://www.sanger.ac.uk/Projects/M_marinum/) and *S. coelicolor* (http://www.sanger.ac.uk/Projects/S_coelicolor/) selecting genome sequence EMBL entries.

The Artemis software (release 4 for Macintosh, Sanger Center) was used to annotate the selected regions of *M. smegmatis*, *M. avium*, *M. leprae* and *M. marinum*.

BLASTp searches of ORFs identified by Artemis were performed using the NCBI services (http://www.ncbi.nlm.nih.gov/BLAST) against the nr database.

**PCR**

Amplifications were performed using Taq-DNA polymerase (Promega) under standard conditions, in a volume of 50 µl. Briefly, the concentration of dNTPs was 0.2 mM each, and 20 pmol of each primer were used. Cycling conditions were: one cycle of 94°C for 3 min, followed by 35 cycles of: 94°C for 1 min, 56°C for 1 min, 1 min at 72°C; followed by 10 min at 72°C. A total of 2 ng of genomic DNA was used as template. The *M. avium* genome was amplified with the primers: Maup2DR (CGGCCGTTAC-CGGCAGGATAC), which anneals to Rv2800-like gene, and MalowDR (GGATGGGCAGGCAGGATAC), which anneals to Rv2825-like gene. *M. smegmatis* was amplified with the primers MsupDR (TCCCCGTGCCCCGGATCTCTTA), which anneals to Rv2800-like gene, and MslowDR (CGACTACGGAGCGTGAAGAG), which anneals to Rv2825-like gene.
Results

As a first step, we identified the following genes that are described as conserved in the Tuberculist database and located around the DR region in the *M. tuberculosis* genome: Rv2785c (*rpsO*), Rv2786c (*ribF*), Rv2790c (*ltl1*), Rv2793c (*truB*), Rv2800, Rv2825, Rv2828, Rv2831 (*echA16*), Rv2838 (*rbfA*) and Rv2845 (*proS*) (Figure 1). As highly conserved counterparts of all these genes are present in several other bacterial species, they were used as markers to locate, *in silico*, the corresponding orthologues in the *M. smegmatis*, *M. avium*, *M. leprae* and *M. marinum* genomes, and thus identify the regions where DR-like sequences might be found. In *M. tuberculosis*, the DR region is located between Rv2800 and Rv2825.

BLASTp searches against four mycobacterial genomes (*M. smegmatis*, *M. avium*, *M. leprae* and *M. marinum*) indicate that in these bacteria most of the *M. tuberculosis* marker genes have highly similar orthologues, which are also found in the same order and orientation. Table 1 shows the data from *M. tuberculosis*, *M. smegmatis* and *M. avium*. Importantly, there are no orthologues for ORFs Rv2801–Rv2824 in either *M. smegmatis* or *M. avium*, suggesting that these genes are unique to *M. tuberculosis* within the genus *Mycobacterium*, at least in those mycobacteria whose genomes were sequenced. The absence of these ORFs is reflected by the distance between the orthologues of Rv2800 and Rv2825, which is 24.5 kb in *M. tuberculosis*; but only 2.2 kb in *M. avium* and 2.2 kb in *M. smegmatis*. Similar observations were made in *M. leprae* and *M. marinum* (Table 2). No repetitive sequences similar to the DR of *M. tuberculosis* were found.

As the *M. avium* genome is not yet annotated, the region between the orthologues of *truB* and *proS* was located by text phrase search in the whole genome text file provided by TIGR. This region, comprising 20527 bp, was selected, cut from the text file, and its ORFs analysed using Artemis. In turn, the *M. smegmatis* region between *truB* and *rbfA* was located in the same way on contig 3315 provided by TIGR, from which 19002 bp were cut and analysed with Artemis. Every ORF present in these regions was BLAST-searched against the Tuberculist database. We observed that orthologues of Rv2801–Rv2824 are not present in *M. smegmatis* and *M. avium*, and that Rv2800 and Rv2825 are adjacent in these genomes (Figure 1). Thus, a 24.5 kb segment of the *M. tuberculosis* genome that contains

![Figure 1. Schematic representation of ORFs in the DR region and surrounding area in *M. tuberculosis* and other mycobacteria. Mt, *M. tuberculosis*. Arrow blocks represent ORFs. The names of genes are given above the ORFs (as they appear in Tuberculist). Circles represent the DR region. To facilitate comparison, Rv2800 and Rv2825 are coloured grey and black, respectively.](image-url)
Rv2801–Rv2824 is completely absent from the *M. smegmatis* and *M. avium* genomes.

By examining in detail the putative Rv2800 and Rv2825 orthologues in *M. avium*, it was observed that these genes overlap at their 3′ ends by 41 bp in this species. In the *M. smegmatis* genome these genes are separated by 23 bp. We also noted that the *ugp* operon is located elsewhere in the *M. avium* genome and that the *dinF* gene is absent from the *M. avium* genome (data not shown). The *M. leprae* genome is annotated, so it was easier to analyse the genetic structure of this region. The orthologues of Rv2800 and Rv2825 are ML1550 and ML1551, respectively. Both are pseudogenes, and they are adjacent and divergently transcribed, as in the other species.

In *M. marinum*, there is an ORF between the Rv2800- and Rv2825-like genes (Figure 1). This ORF has limited similarity to the *otsB1* probable trehalose-6-phosphatase and is highly similar to a hypothetical protein from *Streptomyces lavendulae* (data not shown). In *Streptomyces coelicolor* there are no orthologues of the Rv2800, Rv2813, Rv2816 and Rv2825 genes and we found that the *truB* and *rbfA* genes are neighbours (data not shown).

Primers spanning the juncture of the orthologues of Rv2800 and Rv2825 in *M. smegmatis* and *M. avium* were designed (the upper primers hybridize to Rv2800 and the lower primers to Rv2825) and used to confirm by PCR the *in silico* observation that these genes are neighbours in these species. The amplicons observed (Figure 2) confirmed the expected sizes; around 1 kb in *M. avium* and 0.55 kb *M. smegmatis*. As expected, amplification was not observed in *M. tuberculosis* because the hybridization sites of the primers are too far apart.

**Discussion**

Despite the sequence conservation showed by most of the mycobacterial genomes so far sequenced, a genomic segment containing 21 genes and the DR region present in the *M. tuberculosis* complex is absent in *M. avium*, *M. smegmatis*, *M. marinum* and *M. leprae*. This region has probably been acquired by *M. tuberculosis*. This conclusion is based on the hypothesis that *M. avium* is older than *M. tuberculosis* (Sreevatsan et al., 1997). *M. avium* and *M. smegmatis* probably appeared earlier in the evolution of this clade than *M. tuberculosis* because the genomes of these bacteria are longer than that of *M. tuberculosis*. *M. marinum* also seems to have a large genome (http://www.sanger.ac.uk/Projects/M.marinus/). Progressive deletions are proposed to have accumulated during the evolution of the *M. tuberculosis* complex, leading to speciation (Brosch...
Table 2. Distances between ORFs that flank the DR region in *M. tuberculosis*

<table>
<thead>
<tr>
<th>Gene</th>
<th><em>M. avium</em> Position</th>
<th>Delta (kb)</th>
<th><em>M. tuberculosis</em> Position</th>
<th>Delta (kb)</th>
<th><em>M. smegmatis</em> Position</th>
<th>Delta (kb)</th>
<th><em>M. leprae</em> Position</th>
<th>Delta (kb)</th>
<th><em>M. marinum</em> Position</th>
<th>Delta (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rv2793/truB</td>
<td>3 810 970</td>
<td></td>
<td>3 102 367</td>
<td></td>
<td>1 55 877</td>
<td></td>
<td>1 870 999</td>
<td></td>
<td>405 450</td>
<td></td>
</tr>
<tr>
<td>Rv2800</td>
<td>3 813 339</td>
<td>2.4</td>
<td>3 108 416</td>
<td>6.0</td>
<td>1 59 842</td>
<td>4.0</td>
<td>1 874 326</td>
<td>3.3</td>
<td>400 324</td>
<td>5.1</td>
</tr>
<tr>
<td>Rv2825</td>
<td>3 815 567</td>
<td>2.2</td>
<td>3 132 895</td>
<td>24.5</td>
<td>1 62 020</td>
<td>2.2</td>
<td>1 875 360</td>
<td>1.0</td>
<td>394 666</td>
<td>5.6</td>
</tr>
<tr>
<td>Rv2831/echA16</td>
<td>3 815 922</td>
<td>0.35</td>
<td>3 135 791</td>
<td>2.9</td>
<td>1 62 357</td>
<td>0.33</td>
<td>1 876 260</td>
<td>0.9</td>
<td>392 937</td>
<td>1.0</td>
</tr>
<tr>
<td>Rv2838/rbfA</td>
<td>3 818 714</td>
<td>2.8</td>
<td>3 144 623</td>
<td>8.8</td>
<td>1 74 879</td>
<td>12.5</td>
<td>Invert</td>
<td></td>
<td>382 741</td>
<td>10.2</td>
</tr>
<tr>
<td>Rv2845/proS</td>
<td>3 825 596</td>
<td>6.8</td>
<td>3 151 205</td>
<td>6.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Distance with respect to the previous gene.*

"Located downstream of proS in *M. leprae*."

The molecular mechanism that has led to the insertion of this fragment in *M. tuberculosis* is unknown, but it is possible to speculate that the putative transposases present in the upstream flank of the DR could have played a role in
a hypothetical insertion into the *M. tuberculosis* chromosome. Significantly, *Rv2813* is highly similar (expect score $= e^{-117}$, 245/270 identical amino acids) to a gene found on an *M. celatum* plasmid, close to a double 18 bp tandem repeat (data not shown) with no function identified (Picardeau and Vincent, 1998; Picardeau *et al*., 2000; Le Dantec *et al*., 2001). All these features are reminiscent of the organization of the DR region in *M. tuberculosis* (2 × 18 bp = 36 bp) and suggest that the DR region could have been acquired from a mycobacterial plasmid, not necessarily from pCLP1. Several authors have proposed horizontal transfer as an explanation for genetic exchange amongst mycobacteria (Howard *et al*., 2002; Le Dantec *et al*., 2001; Parsons *et al*., 1998) but mycobacterial horizontal transfer is a controversial subject.

In summary, we propose that the DR region was acquired by the *M. tuberculosis* complex or by an ancestral bacterium. The genomes of other mycobacterial species are currently being sequenced, and by extending our analysis to them, we should be able to confirm or discard this hypothesis.

### References


