Characterization of a *Mycobacterium bovis* BCG insertion sequence related to the IS21 family

Philippe Lefe'vre, Martine Braibant, Jean Content, Philippe Gilot *

Department Virology, Pasteur Institute, rue Engeland 642, B-1180 Brussels, Belgium

Received 31 May 1999; received in revised form 12 July 1999; accepted 12 July 1999

Abstract

The structure and distribution of a *Mycobacterium bovis* BCG insertion element of the IS21 family were investigated. Several IS21-like elements found in mycobacterial genomes were separated in four types, following their nucleic acid similarities. The *M. bovis* BCG IS21 element is highly similar to IS1533 (class I), 70% similar to IS1534 (class II), 52% similar to IS1532 (class III) of *Mycobacterium tuberculosis*, and 54% similar to both an *Mycobacterium avium* serovar 2 and an *M. avium silvaticum* IS (class IV). The *M. bovis* BCG IS21 element of the class I appears to be present in a single copy in the genome of *M. bovis* BCG, *M. bovis*, *M. tuberculosis* and *Mycobacterium africanum* and to be absent from all other tested species of the Corynebacteria-Mycobacteria-Nocardia group. ß 1999 Federation of European Microbiological Societies. Published by Elsevier Science B.V. All rights reserved.

Keywords: Corynebacteria-Mycobacteria-Nocardia group; IstA; IstB

1. Introduction

Bacterial insertion sequences (IS) are small (<2.5 kb), phenotypically cryptic segments of DNA with a simple genetic organization. They are capable of inserting at multiple sites in target molecules and they generally encode no functions other than those involved in their mobility. These include recombinationally terminal inverted repeat (IR) and a transposase which recognizes and processes these ends. The bacterial IS can be classified in at least 17 families based on the arrangement of their open reading frames, on similarities of their transposases and of their ends, and on the fate of the nucleotide sequence of their target site [1].

The IS21 family is named after the prototypical IS21 element discovered as a constituent of the conjugated plasmid R68, a broad host range plasmid among the Gram-negative bacteria. Members of the IS21 family carry related terminal IR, whose lengths may vary between 11 bp and 50 bp. Several members carry multiple repeated sequences in the terminal IR, which may represent transposase binding sites. They exhibit two consecutive open reading frames: a long upstream frame designated *istA* and a shorter downstream frame designated *istB* [1]. In *Escherichia coli*, it was shown that *istA* encodes, in frame, a 46-kDa and a truncated 45-kDa protein, termed transposase and cointegrase, respectively. Transposase essentially catalyzes transposition of single IS21 elements while
cointegrase rarely carries out this reaction but is highly effective in a mechanism of replicon fusion between plasmids carrying an IS21 tandem duplication and a target replicon. Both IstA and IstB proteins are required for integration. IstA is responsible for the 3' end cleavage of IS21, while target joining is assisted by the 30-kDa helper protein IstB in the presence of ATP [2-6].

We previously described the cloning, sequencing and expression of a *Mycobacterium bovis* BCG gene encoding a 22-kDa protein present at the surface and in the culture fluid of the bacterium [7]. This protein was found to be homologous to a putative lipoprotein encoded by the *lppx* gene (Rv2945) of *M. tuberculosis* H37Rv ([8], EMBL Database Acc. no. Z83858). On the same cloned DNA fragment and in the opposite orientation we also found a complete 801 bp ORF preceded by an other incomplete ORF.

In this work we describe that these two ORFs encode proteins similar to the IstB and IstA proteins encoded by the IS21 insertion element of *E. coli*, and we analyze the presence and copy number of this insertion sequence among members of the Corynebacteria-Mycobacteria-Nocardia (CMN) group.

2. Materials and methods

2.1. DNA manipulation and sequence analysis

Genomic DNA was extracted as previously described [9]. All other DNA manipulations were carried out by standard methods [10]. DNA was sequenced on both strands as described previously [7]. Dot and Southern hybridization of chromosomal DNA were performed either with a 190 bp class I-*istA* internal probe (a *Sma*I restriction fragment extracted from the *M. bovis* BCG fragment cloned in plasmid C3-1 [7] which corresponds to nt 12901 to 13091 of the Rv2943 gene of *M. tuberculosis* H37Rv (EMBL Acc. no. Z83858)), or with a 371 bp class I-*istB* internal probe (an *SphI*-StuI restriction fragment, nt 577 to 944, Fig. 1). Restriction fragments used as probes were labelled by random priming with [α-32P]dCTP. The hybridization was conducted in the Rapid-hyb® buffer of Amersham according to the manufacturer’s instruction. After hybridization, the membrane was washed at high stringency (final wash at 65°C in a 0.1% SDS, 0.1 × SSC buffer), and autoradiographed at −70°C.

3. Results and discussion

3.1. Nucleotide sequence analysis

The *M. bovis* BCG 801 bp ORF previously cloned in plasmid pBluescript II SK+ (Stratagene) was sequenced on both strands (Fig. 1). Searches in protein databases (Blast P) revealed significant homologies of the 29374 Da protein encoded by this 801 bp ORF with the IstB protein encoded by the IS21 insertion sequence of *E. coli* [3], with three *M. tuberculosis* H37Rv gene products: Rv2944, Rv3638 and Rv3427 [8], with an *M. avium* serovar 2 protein (EMBL Acc. no. AF125999), with a *M. avium silvaticum* protein ([11], EMBL Database Acc. no. AJ223832), and with several transposases present in different bacterial genomes. Fig. 2 shows the alignment of the IstB-like proteins encoded by mycobacterial genomes with IstB of *E. coli* IS21. All these IstB proteins possess well conserved regions and in particular they carry (with the exception of Rv3638), a potential nucleotide triphosphate binding domain (P-loop, consensus [AG]-x(4)-G-K-[ST], boxed in Fig. 2), a motif well conserved amongst the IstB family.

It is worth noting that, whereas all the above proteins are rather homologous, the situation is not the same at the DNA level. The *M. bovis* BCG 801 bp ORF is highly similar to the Rv2944 gene (99.6% identity in 892 nt overlap) and moderately similar to the Rv3638 gene (65.6% identity in 756 nt overlap) but present a lower level of similarity with the other cited elements (52.8% identity in 690 nt overlap with Rv3427, 55.2% identity in 645 nt overlap with the *M. avium* serovar 2 and the *M. avium silvaticum* genes, and 55.0% nt in 642 nt overlap with *istB* of *E. coli*). Despite the very high homology between the *M. bovis* BCG 801 bp ORF and the Rv2944 ORF, there is an apparent discrepancy between the first 118 amino acids encoded by the *M. bovis* BCG 801 bp ORF and the first 90 amino acids encoded by Rv2944 (Fig. 2). This is due to the presence of an additional cytosine (after nt 611, Fig. 1) in the de-
posited sequence of the Rv2944 ORF and to the choice of the GTG in position 395 (Fig. 1) as the start codon of the protein encoded by Rv2944 [8]. The Rv2944 gene product as described by the *M. tuberculosis* genome sequencing team [8] is thus shorter than its *M. bovis* BCG homologue and in another reading frame till amino acid 92 (Fig. 2). It is not known if this additional cytosine is a real polymorphism between the two species or is a simple sequencing error.

In *E. coli*, the *istB* coding frame of IS21 is preceded by the *istA* ORF and the TGA stop codon of *istA* overlaps with the ATG start codon of *istB* [3]. Just upstream of the *M. bovis* BCG 801 bp ORF we also found an incomplete 310 bp ORF ending with a TGA codon (Fig. 1, nt 308 to 310). This TGA stop codon overlaps with the start codon of the 801 bp ORF, suggesting that translational coupling may occur.

Similarly the other five mycobacterial genes encoding proteins homologous to IstB of *E. coli* are preceded by ORF encoding proteins similar to IstA of *E. coli* (Fig. 3). The 102 amino acids polypeptide encoded by the incomplete 310 bp ORF of *M. bovis*...
BCG is 100% similar to the carboxyl end of the protein encoded by the *M. tuberculosis* Rv2943 ORF, and the protein encoded by Rv2943 shares 25.2% identical aa in 230 aa overlap with IstA of *E. coli*. Homology of the other proteins with IstA of *E. coli* are: 22% identical aa in 50 aa overlap for the polypeptide encoded by Rv3636, 26.9% identical aa in 108 aa overlap for the polypeptide encoded by Rv3637, 27.3% identical aa in 267 aa overlap for the protein encoded by Rv3428, 23.2% aa identical in 164 aa overlap for the *M. avium silvaticum* protein encoded by the incomplete istA ORF, and 23.1% identical aa in 360 aa overlap for the *M. avium* serovar 2 transposase. Whereas all those IstA-like proteins can be aligned, their overall similarity is nevertheless lower than the one found for the IstB-like proteins (result not shown, EMBL alignment accession number ds 38468).

As for istA of *E. coli* and for the *M. bovis* 301 bp incomplete ORF described above, the stop codon of the Rv3637 ORF, of the *M. avium* serovar 2 transposase gene and of istA of *M. avium silvaticum* overlaps with the start codon of their downstream istB-like ORF. Contrary to this usual organization in the IS21 family, the Rv3428 ORF is separated by 97 nt from the Rv3427 ORF, and the Rv2943 ORF is separated (with the restriction concerning the additional cytosine discussed above) by 85 nt from the...
A 17 bp separation of istB from istA was nevertheless also described for IS408, an insertion sequence of the IS21 family originating from Burkholderia cepacia [1].

All those mycobacterial elements were also found to possess inverted repeats at their ends [12] and possess thus all features of members of the IS21 family.

Based on their nucleic acid similarities, these mycobacterial IS21-like sequences can be separated in four types (Fig. 3).

The type I comprises the *M. bovis* BCG IS described in this work and the *M. tuberculosis* IS1533 (Rv2943 and Rv2944). These two IS are almost identical except for the previously cited additional cytosine in the sequence of the *M. tuberculosis* istB ORF and the change of a GCC istB codon (Fig. 1, nt 817 to 819) by a CGC codon.

The type II comprises the *M. tuberculosis* IS1534 (Rv3636-Rv3637, Rv3638) which possess around 70% nucleic acid identity with the type I mycobacterial IS21. IS1534 is possibly defective because only the first aa (Rv3636) and the last 166 aa (Rv3637) of this IstA-like protein seem to be produced by IS1534 (Fig. 3, [8]). This is due to three missing nucleotides in the deposited sequence of the *istA*-like gene of IS1534. Indeed, the DNA encoding the IstA-like protein of IS1533 (Rv2943) and of IS1534 (Rv3636 to Rv3637) can be aligned (70.9% identity in 1216 nt overlap) but three gaps in the DNA sequence covering Rv3636 to Rv3637 are necessary for performing this alignment. A CA dinucleotide and a C are missing 329 nt and 675 nt after the start codon of Rv3636, respectively. Nevertheless, if these nucleotides were added in these positions in the IS1534 *istA*-like gene, a complete ORF of 1236 nt could be derived and the 411 aa protein encoded by this ORF shows 70.3% identical aa in 407 aa overlap.

Fig. 3. Schematic representation of the mycobacterial DNA regions around the IS21 elements. The mycobacterial IS21 elements were aligned with IS21 of *E. coli*. Black arrows indicate direction of transcription of the ORF and the blank dotted part of the arrows shows unsequenced regions of the IS. Thin lines represent intergenic regions containing the inverted repeats of the insertion elements, and grey lines intergenic regions internal to IS21. Roman numerals indicate the four different classes of the mycobacterial IS21 elements.
with the protein encoded by Rv2943. Another argument for the non-functionality of IS1534 is that the IstB-like protein encoded by the Rv3638 ORF lacks the nucleotide triphosphate binding domain (Fig. 2) presumably necessary for the faithful transposition of IS21 [5].

The type III comprises the M. tuberculosis IS1532 (Rv3428, Rv3427) which possess around 52% nucleic acid identity with the type I mycobacterial IS21, whereas the type IV comprises the M. avium serovar 2 and the M. avium silvaticum insertion sequences which are very similar and which each possess around 54% nucleic acid identity with the type I mycobacterial IS21.

The presence of IS belonging to three different classes (I to III) of the IS21 family in M. tuberculosis genome is particular. These IS could have evolved in M. tuberculosis from a common ancestor first integrated in its genome but it is also conceivable that each of these classes I, II or III IS have evolved on their own in different bacterial species and were then acquired independently by horizontal transfer into M. tuberculosis. In this context, it is worth noting that the highest similarity of IS1532 (class III) is with IS1491 from Pseudomonas alcaligenes, the highest similarity of IS1533 (class I) is with IS5376 from Bacillus stearothermophilus and that the highest similarity of IS1534 (class II) is with IS21 from E. coli [12].

3.2. Distribution and copy number of the type I mycobacterial IS21 element

To determine the host range of the type I mycobacterial IS21-like elements amongst the CMN group, the total DNA from several different species of this group was extracted and first analyzed by dot hybridization under high stringency, using a 371 bp istB internal probe and a 190 bp istA internal probe. Only M. bovis, M. bovis BCG, M. tuberculosis, and M. africanum – the three species of the M. tubercu-
losis complex which are pathogenic for both humans and animals [13] – were able to hybridize with the istB and the istA probes (Fig. 4A).

The copy number of this IS in the genome of the four strains of the CMN group was then determined by Southern blot hybridization experiments. DNA isolated from these strains was digested with EcoRI, separated by agarose gel electrophoresis, transferred to a nylon membrane and hybridized with the istA and istB probes under high stringency. Both probes hybridized in each of these strains with a unique DNA fragment of approximately 6600 bp (Fig. 4B). Since there are no EcoRI sites within the type I mycobacterial IS21 element, a single hybridizing fragment was assumed to represent a single copy of the IS. This approach may nevertheless yield to an underestimation of the copy number of the IS since three IS copies would be able to reside on a DNA fragment of this length. The analysis of the M. tuberculosis H37Rv genome sequence has shown that IS1533 is present in only one copy and only present in one copy in this species [8]. We searched for EcoRI sites around IS1533 and we found that this IS is present on a 6630 bp EcoRI fragment. It can thus be assumed that the type I mycobacterial IS21-like elements are also only present in one copy in M. bovis, M. bovis BCG and M. africanum. Nevertheless, using PCR experiments, Gordon et al. recently described that IS1533 was absent from one of the 29 clinical isolates they examined [12].

In conclusion we here have shown that mycobacteria possess four different classes of IS21-like elements and that the class I IS21 element described in this work is present in a unique copy and only in members of the M. tuberculosis complex.

Acknowledgements

We thank L. De Wit for help provided in DNA sequencing, V. Pelicic (Pasteur Institute, Paris) for making preliminary experiments, M. Dufaux (Pasteur Institute, Brussels) for the gift of M. bovis, M. africanum and M. microti DNA and P. Brennan (Colorado State University, Fort Collins, CO) for the gift of M. leprae DNA. This work was supported by grants PL 96 2167 and 96 2134 from the European Economic Community (BIOMED 2)

References