Biotechnol Lett (2010) 32:1853–1862
DOI 10.1007/s10529-010-0357-y

ORIGINAL RESEARCH PAPER

Structural peculiarities of linear megaplasmid, pLMA1, from *Micrococcus luteus* interfere with pyrosequencing reads assembly

Martin Wagenknecht · Julián R. Dib · Andrea Thürmer · Rolf Daniel · María E. Farías · Friedhelm Meinhardt

Received: 29 June 2010 / Accepted: 9 July 2010 / Published online: 21 July 2010

Abstract Different strains of *Micrococcus luteus*, isolated from high-altitude Argentinean wetlands, were recently reported to harbour the linear plasmids pLMA1, pLMH5 and pLMV7, all of which with 5′-covalently attached terminal proteins. The link between pLMA1 and the host's erythromycin resistance as well as further presumptive qualities prompted us to perform a detailed characterization. When the 454 technology was applied for direct sequencing of gel-purified pLMA1, assembly of the reads was impossible. However, combined Sanger/454 sequencing of cloned pLMA1 fragments, covering altogether 23 kb of the 110-kb spanning plasmid, allowed numerous sequence repeats of varying in lengths to be identified thus rendering an explanation for the above 454 assembly failure. A large number of putative transposase genes were identified as well. Furthermore, a region with five putative iteron sequences is possibly involved in pLMA1 replication.

Keywords 454 Sequencing · Inverted repeats · Iteron · Linear plasmid · *Micrococcus* repetitive sequences · Transposase

Introduction

Microbial linear plasmids are rather widespread, particularly among various Gram-positive bacteria. They have been found in many *Streptomyces* spp., several rhodococci and mycobacteria, but also in *Arthrobacter nitroguajacolicus* Rü61a, *Planobispora rosea*, and the plant pathogen *Clavibacter michiganensis*. Such linear replicons belong to a class of genetic elements, which are characterized by terminal inverted repeats (TIRs) and terminal proteins (TPs) attached to each 5′-end. Most of the linear megaplasmids are conjugative and display rather low-copy numbers (see Wagenknecht and Meinhardt 2010, and references therein). Linear plasmids often encode nonessential functions but—under certain conditions—they may provide advantages attributes, such as heavy metal resistance or specific catabolic traits. Antibiotic resistance is rather seldomly found to be

Electronic supplementary material The online version of this article (doi:10.1007/s10529-010-0357-y) contains supplementary material, which is available to authorized users.

M. Wagenknecht · F. Meinhardt (✉)
Institut für Molekulare Mikrobiologie und Biotechnologie, Westfälische Wilhelms-Universität Münster, Corrensstr. 3, 48149 Münster, Germany
e-mail: meinhar@uni-muenster.de

J. R. Dib · M. E. Farías
Laboratorio de Investigaciones Microbiologicas de Lagunas Andinas (LIMLA), Planta Piloto de Procesos Industriales Microbiológicos-CONICET, Av. Belgrano y Pje. Caseros, 4000 Tucumán, Argentina

A. Thürmer · R. Daniel
Göttingen Genomics Laboratory, Institut für Mikrobiologie und Genetik, Georg-August-Universität Göttingen, Grisebachstr. 8, 37077 Göttingen, Germany

linear plasmid-encoded, as for methylenomycin of *Streptomyces* linear elements. [For a monograph on microbial linear plasmids see Meinhardt and Klassen (2007).]

We recently reported the isolation and characterization of the first linear plasmids in different strains of *Micrococcus luteus* (Dib et al. 2010a). All of them (pLMA1, pLMH5, and pLMV7) possess 5′-attached TPs. Host strains were isolated from high-altitude Argentinean wetlands, which are considered extreme and pristine environments characterized by high UV radiation, arsenic concentration, and salinity. Such bacteria displayed high UV tolerance, heavy metal resistance, and, unexpectedly, resistance against a number of antibiotics (Dib et al. 2008). As there is already a proven link between pLMA1 occurrence and erythromycin resistance (Dib et al. 2010a) we decided to focus on such plasmid.

One of the most powerful new sequencing technologies is the 454 pyrosequencing as such parallel noncloning pyrosequencing-based system is capable of delivering sequence information much faster than current Sanger sequencing platforms. 454 sequencers provide shorter reads (ideally ~350 bp on average versus ~800 bp for Sanger reads) but at greatly reduced per-base costs. Moreover, the 454 technology was shown to be capable of resolving hard stops for which Sanger sequencing is ineffective (Goldberg et al. 2006).

When we sequenced pLMA1 with the 454 technology, attempts to assemble the reads encountered insurmountable obstacles. However, applying both, the Sanger sequencing and the 454 technology for cloned pLMA1 fragments, long sequence stretches of the linear plasmid were obtained. Peculiar attributes, such as a large number of repetitive sequences and transposase encoding genes became obvious.

Materials and methods

Bacterial strains, plasmids, and growth conditions

M. luteus strains were cultivated as described in Dib et al. (2010a). Erythromycin 100 μg ml^{-1} was added to the medium to counteract loss of pLMA1. *Escherichia coli* NEB 5-alpha (New England Biolabs) and pUC18 were used for cloning *Pst*I restriction fragments.

DNA isolation

Extraction of *Pst*I-digested DNA fragments from agarose gels was achieved using the QIAquick Gel Extraction Kit (Qiagen). Isolation of bulk and pLMA1 DNA was done as described previously (Dib et al. 2010a, b).

Labeling of DNA probes and Southern hybridization

Probe fragments were PCR-amplified using Phusion Hot Start High-Fidelity DNA Polymerase (Finnzymes), primer pairs op15-revw5/op15-uniw3-1 (2,315 bp), op34-revw3-1/op34-uniw4 (1,597 bp), op54-revw2/op54-uniw2 (1,607 bp), op62-revw2/op62-uniw2 (999 bp), and op74-revw2/op74-uniw1 (1,456 bp), and pP86-15, pP62-34, pP45-54, pP37-62, and pP37-74, respectively, as template (for primer sequences see Supplementary Table S1). Probe labeling, capillary blotting, and hybridization/detection were done as described in Wagenknecht and Meinhardt (2010).

454 Sequencing of pLMA1, assembly, and mapping

Sequencing of gel-purified plasmid pLMA1 was done using the Genome Sequencer FLX system (Roche Applied Science). A single-stranded DNA shotgun library (ssDNA library) was generated from approximately 5 μg of isolated pLMA1 DNA. The DNA was fragmented by nebulization for 30 s and 1 bar. Further steps were done according to the Roche protocol. The size selection of the ssDNA library resulted in an average length of 469 bp. A total of 84,302 reads were achieved. The GS De Novo Assembler (Roche Applied Science) software package was used for sequence assembly. The GS Reference Mapper was used to align the reads from the 454 sequencing run to the cloned *Pst*I restriction fragments.

Sequencing of cloned *Pst*I restriction fragments by primer walking

Sequencing of the pUC18-cloned *Pst*I restriction fragments was done using BigDye Terminator 3.1 chemistry and an ABI Prism 3700 DNA Analyzer (Applied Biosystems). For sequencing primers see

Table S1 in the Supplementary material. Hard-stop events that occurred during primer walking were bypassed by an additional 10 min denaturation step combined with 1 M betaine prior to the sequencing reaction.

Prediction and annotation of ORFs and identification of repetitive sequences

An initial set of predicted protein-coding regions was identified using Artemis V11.22 (Sanger Institute). ORFs shorter than 50 amino acids and with overlaps of higher scoring regions were eliminated. Annotation was done considering a Blast (Altschul et al. 1990) search against Swissprot (Pearson 1994) and GenBank nr (Benson et al. 2004). Putative protein functions were assigned for hits with a full length alignment and appropriate similarity and confirmed by protein domain hits.

Repetitions within the nucleotide sequence of the cloned *Pst*I fragments of pLMA1 and the reference sequences were identified and analysed using Spectral Repeat Finder (Sharma et al. 2004) and Clone Manager (Sci-Ed Software).

Results and discussion

454 Pyrosequencing of pLMA1 and read assembly

Approx. 5 µg pLMA1 DNA, isolated by electroelution from preparative pulsed-field (PF) gels, were used to generate the single-stranded DNA shotgun library that was subsequently subjected to the 454 sequencing procedure. The sequencing runs yielded 84,302 single reads generating total sequence information of 8,703,704 bases. Considering 110 kb as the electrophoretically determined size of pLMA1 (Dib et al. 2010a), the obtained sequence data theoretically correspond to a 79-fold depth coverage. However, by de novo assembling of the single reads, it turned out that only 29% of all the sequences could be assembled. Moreover, from the 955 contigs obtained, the majority (~80%) displayed remarkable short sizes, ranging from 100 to 200 bp. About 16% of all contigs ranged from 201 to 400 bp; the maximum length was 684 bp only. All bioinformatic attempts to increase the contig lengths which, after repeated assembly, would allow for creating longer stretches of continuous pLMA1 sequences, failed. Thus, subsequent sequence analyses, such as ORF prediction and annotation, were virtually impossible.

Chromosomal DNA contamination of the plasmid DNA sample could have been the reason for above findings. However, test digests of the pLMA1 DNA sample with different restriction endonucleases revealed a distinct band pattern with a negligible smear in the gel (Supplementary Fig. S1), indicative of only faint amounts of contaminating chromosomal DNA.

Repetitive sequences may severely hamper de novo assembly of sequence data, as short read lengths, characteristic for 454 pyrosequencing, may make it impossible to span such repetitive elements. This situation is known from genome sequencing projects of bacteria (Goldberg et al. 2006) and in particular those of eukaryotes with highly repetitive genomes, such as barley (Wicker et al. 2006). The more and the longer sequence repeats are present in a given DNA molecule, such issue gains in importance. Though the employed Genome Sequencer FLX System (Standard Series, Roche) theoretically allows read lengths of 200–300 bases, the average read length obtained for pLMA1 was 103 bases, making the assembly of nucleotide sequences rich in repetitive elements even more difficult. It is a fairly common experience that a high GC content may lead to short read lengths, which agrees in the case of pLMA1 with the calculated average GC content of 74.2% of all reads obtained.

Cloning and Sanger sequencing of selected restriction fragments of pLMA1

Since the above 454 pyrosequencing did not yield long continuous stretches of pLMA1 nucleotide sequences and for checking that repeats caused the failure of the assembly, we cloned restriction fragments of pLMA1 and subjected them to Sanger sequencing.

Five of the fragments produced by *Pst*I cleavage (8.6, 6.2, 5.8, 4.5, and 3.7 kb; Supplementary Fig. S1), were excised from a preparative gel and cloned in *Pst*I-linearized pUC18. Plasmid DNA isolated from *E. coli* transformants was cut with *Pst*I, which released the insert. Each of the transformation reactions—with one exception—yielded transformants with plasmids exhibiting the expected insert sizes. Plasmids designated pP86-15, pP62-34, pP45-54, and pP37-62, harboring the 8.6, 6.2, 4.5, and the 3.7 kb *Pst*I restriction fragment, respectively, were used for insert sequencing by primer walking starting with the standard primers 'uni' and 'rev'. Regrettably, though repeatedly attempted, no hybrid plasmid containing the 5.8 kb insert was obtained.

As linear plasmids of actinomycetes possess TPs covalently attached to each 5′-end, proven also for pLMA1 (Dib et al. 2010a), and treatment with proteinase K (done prior to PFGE to allow the DNA to enter the gel) does not remove the linking amino acid residue of the TP (Yang et al. 2002), the 5.8 kb *Pst*I restriction fragment possibly carries one of the pLMA1 termini.

Plasmids isolated from PF gels routinely contain trace amounts of chromosomal DNA. Thus, before continuing insert sequencing by primer walking, sequence data obtained from the initial sequencing reactions using 'uni' and 'rev' were subjected to a BlastN analysis to verify the origin of the cloned *Pst*I fragments. The chromosomal sequence of *M. luteus* NCTC 2665 (Accession no. CP001628) served as the reference. No, or only partial, similarity was seen for the inserts of pP86-15, pP62-34, and pP45-54; however, the insert sequence of plasmid pP37-62 revealed a chromosomal origin of the cloned 3.7-kb *Pst*I fragment. A PCR analysis (not shown), using insert-specific primers and total DNA of the wild-type strain *M. luteus* A1 and of the pLMA1-deficient strain *M. luteus* A1-M1 as templates, confirmed above findings. In an additional Southern analysis (Fig. 1), we PCR-amplified probes deduced from the respective insert sequences (Fig. 2) and hybridized them with *Pst*I-digested total DNA of *M. luteus* strains A1 and A1-M1, and with likewise-cut pLMA1 DNA. Observed hybridization signals affirmed inserts of pP86-15, pP62-34, and pP45-54 being fragments of pLMA1 (Fig. 1a, b, and c), whereas the insert of pP37-62 indeed is a chromosomal fragment (Fig. 1e). We therefore checked remaining transformants obtained from the corresponding transformation experiment by isolating plasmids and sequencing the terminal regions of their inserts with the backbone primers. Analysing insert sequences in a BlastN search, a couple of plasmids exhibited identical inserts that did not show similarity with the reference chromosome. One of them, pP37-74, was selected for entire insert sequencing and subjected to further PCR and Southern analysis (Fig. 1d), which confirmed its origin from pLMA1.

During the sequencing reactions of pP86-15, pP62-34, and pP37-74 hard stops arose that were resolved as described in Material and methods. Regions with a potential to form such secondary structures are known to cause difficulties in Sanger sequencing, predominantly in DNA molecules with a high GC content (Goldberg et al. 2006).

Unexpectedly, hybridization of the probe deduced from the insert of pP45-54 revealed an additional signal of chromosomal origin with a size of approximately 2.4 kb (Fig. 1c). Comparison of nucleotide sequences of the probe and the *M. luteus* NCTC 2665 reference chromosome disclosed a homology stretch of 625 bp with 99.7% identity, present also on the chromosome of the pLMA1 host strain *M. luteus* A1.

Action of the restriction endonulcease *Pst*I is influenced by the nucleotide sequences neighbouring its recognition site in a way that adjacent GC runs significantly impede cleavage (Armstrong and Bauer 1982), rendering an explanation for the large but less intense signals observed in lanes containing pLMA1 DNA in addition to the expected ones (Fig. 1a–d, lane 6). Since the lanes in which total DNA of *M. luteus* A1 was separated contained lesser amounts of pLMA1, such signals are hardly observed (Fig. 1a–d, lane 4).

454 Reads map on Sanger-sequenced pLMA1 fragments

After finishing primer walking on the cloned *Pst*I fragments, we were now able to align numerous reads of the 454 sequencing to the Sanger-sequenced pLMA1 fragments. Since each of the pLMA1 fragments was abundantly covered by 454 reads, that also overlapped, the depth coverage obtained reached average values of 56–87-fold. The mapping results are summarized in Table 1. Thus, we could affirm the accuracy of the nucleotide sequences initially obtained by primer walking. Being consistent with a chromosomal origin, only a total of 4 reads mapped on the 3,721-bp fragment of pP37-62 (Table 1). Also, we aligned all 454 reads (84,302) to the *M. luteus* NCTC 2665 reference chromosome; only 7,432 (8.8%) of them were found to map. Thus, chromosomal DNA contamination can most likely be ruled out as the reason for the failure of the 454 reads assembly.

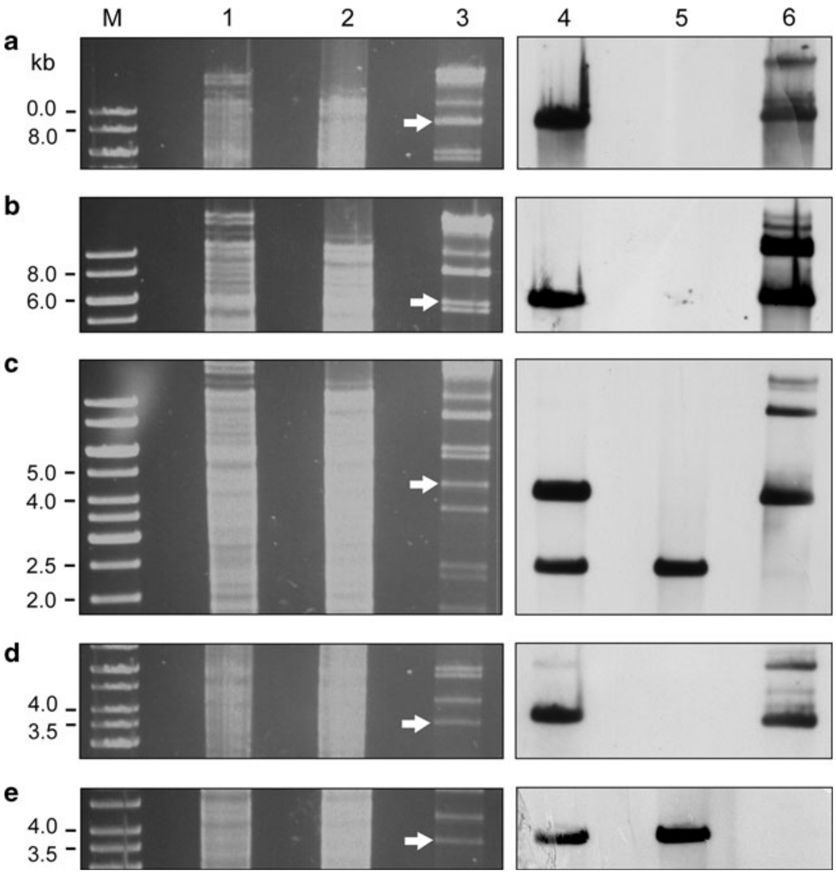

Fig. 1 Origin of the cloned *Pst*I restriction fragments by Southern analysis. Total DNA of the wild-type strain *M. luteus* A1, of the plasmid-deficient strain *M. luteus* A1–M1, and isolated pLMA1 DNA was digested using restriction endonuclease *Pst*I, separated on 1.0% agarose gels (*left panels*), and transferred onto nylon membranes. The DNA was hybridized with probes deduced and PCR-amplified from the cloned *Pst*I restriction fragments (see also Fig. 2). *Right panels* display corresponding sections of exposed X-ray films. Only relevant sections of gels/films are shown. *Arrows* point to pLMA1 restriction fragments with sizes of 8.6 kb (**a**), 6.2 kb (**b**), 4.5 kb (**c**), and 3.7 kb (**d**, **e**), respectively, that were selected for cloning and supposed to give a signal upon hybridization. *M*, DNA size standard; 1 and 4, *M. luteus* A1; 2 and 5, *M. luteus* A1–M1; 3 and 6, pLMA1

Analysis of the base composition revealed a GC content of 64.7, 71.3, 67.6, and 67.4% for the 8621, 6195, 4498, and 3681-bp insert, respectively. The GC content of the 3,721-bp chromosomal *M. luteus* A1 fragment was calculated to be 74%, which is identical to the value reported for *M. luteus* (Ohama et al. 1989) and matches nicely with the GC content of 73% of the chromosome of reference strain *M. luteus* NCTC 2665.

pLMA1 is rich in ORFs involved in transposition

The obtained sequence data of the cloned pLMA1 fragments were subjected to a gene prediction analysis. On the 8621, 6195, 4498, and 3681-bp insert a total of 12, 7, 5, and 5 ORFs were identified, respectively (Fig. 2), covering 76.3% of the sequence of all four pLMA1 fragments. Possible functions of the putative proteins, based on amino acid sequence similarities, are summarized in Supplementary Table S2. For 12 of the predicted ORFs, no function could be attributed.

One remarkable and unexpected feature of the four pLMA1 fragments is the high proportion of genes involved in transposition; particularly, the 3,681-bp pLMA1 fragment is rich in such genes (Fig. 2). The function of the closest relative suggests ORF 3 of pP45-54 to be a DNA replication protein; other

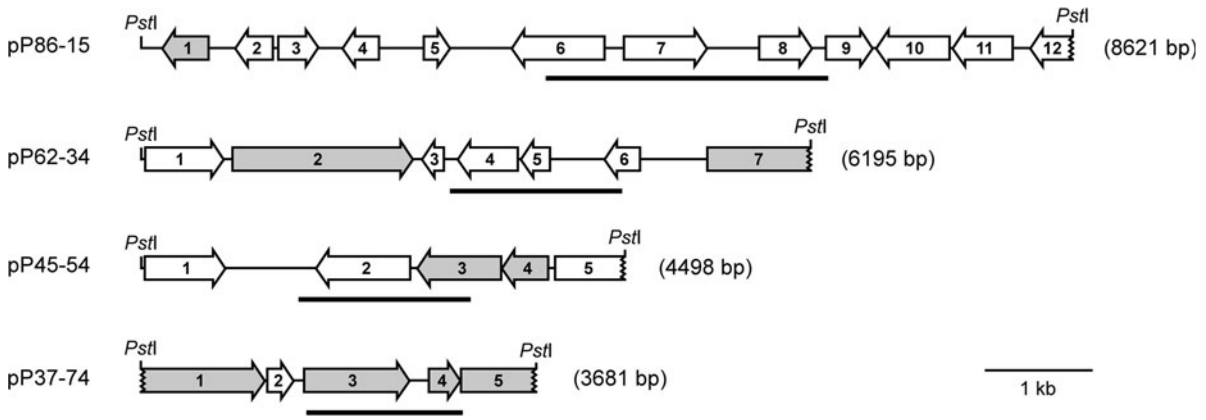

Fig. 2 Schematic representation of annotated ORFs on the sequenced *Pst*I restriction fragments of pLMA1. Designations of the cloning plasmids harboring the respective pLMA1 fragment are shown on the left, the sizes of the pLMA1 fragments are given in *brackets* on the right. Predicted ORFs are shown as *arrows*. *Arrows* missing the beginning or the *arrowhead* (serrated lines) indicate interrupted ORFs. The direction of the *arrows* corresponds to the transcriptional directions. ORFs encoding for proteins involved in transposition are *gray shaded*. Probes used in Southern analysis (see Fig. 1) correspond to *thick*, *black bars* with their respective localisation given below the pLMA1 fragments. Recognition sites of restriction endonuclease *Pst*I are indicated as such and by *short vertical lines*. The sequences of the cloned 8621, 6195, 4498, and 3681-bp pLMA1 fragments, and the 3,721-bp chromosomal fragment have been deposited in the EMBL nucleotide sequence database under accession numbers FN692038, FN692039, FN692040, FN692041, and FN692042, respectively

Table 1 Mapping results of the 454 reads

Name of plasmid	Insert size (bp)	Mapping reads		Depth coverage[b]		
		Absolute number	Relative[a] (%)	Min.	Max.	Mean
pP86-15	8621	7074	8.4	32	134	79
pP62-34	6195	3177	3.8	8	137	56
pP45-54	4498	3228	3.8	18	125	71
pP37-74	3681	3227	3.8	23	143	87
pP37-62	3721	4	0.0	0	1	0

[a] Absolute number of mapping reads divided by total number of reads (84,302)

[b] *Min.* minimum, *Max.* maximum

closely related proteins are annotated as transposase helpers. A copy of this ORF (99% nucleotide sequence identity) was found on the chromosome of *M. luteus* NCTC 2665. Such ORF is probably also present on the chromosome of *M. luteus* A1, as suggested by above Southern analysis (Fig. 1c). ORF 1 of pP86-15, ORFs 2 and 7 of pP62-34, ORF 4 of pP45-54, and ORFs 1, 3, 4, and 5 of pP37-74 (gray shaded in Fig. 2) share significant similarities with known transposases (Supplementary Table S2). Interestingly, nucleotide sequence analysis revealed similarities to putative transposases MC9, MC13, MC19, MC34, and MC35 of pSD10, a 50-kb circular cryptic plasmid from a marine *Micrococcus* strain (Zhong et al. 2002). A region starting 43 bp upstream of ORF 7 (pP62-34) and extending through the ORF is similar to a 1,007 bp block consisting of the first 894 nucleotides of MC13 and 113 nucleotides of its upstream sequence; both share 84% nucleotide sequence identity. It is not surprising that such a similar region is again found for MC19, as MC13 and MC19 are identical except a 1-bp difference (Zhong et al. 2002). The first 1,067 nucleotides of ORF 1 of the pP37-74 insert display 88% sequence identity to nucleotides 84–1,150 of MC9. ORFs 4 and 5 of the same insert located adjacent to each other and overlapping by 4 bp, exhibit a genetic organization identical to MC34 and MC35 of pSD10. Moreover, a 1,056-bp region, consisting of ORFs 4 and 5 (pP37-74) and including 44 bp of upstream sequence of ORF 4, shares 89% nucleotide sequence identity with a region starting 44 bp upstream of MC34 and covering MC34 as well as the first 713 nucleotides of MC35. It is to be expected that the high degree of nucleotide sequence identity of the three truncated putative transposase genes (ORFs 1 and 5 of pP37-74, ORF 7 of pP62-34) continues throughout the missing coding sequences of these genes.

High similarities of above transposase genes and adjacent non-coding regions and their presence in different *Micrococcus* strains points to horizontal gene transfer as well as exchange of genetic information between the plasmid and the host chromosome. Furthermore, the high number of ORFs on pLMA1 involved in transposition, which probably increases upon further sequencing (the total sequence of 22,995 bp obtained so far represents only one-fifth of the entire length of the plasmid), indicates that pLMA1 is presumably very flexible with respect to its genetic composition.

A number of transposases are associated with insertion sequence (IS) elements, which differ in size and sequence composition and are flanked by short inverted repeats (IR), usually between 10 and 40 bp in size (Mahillon and Chandler 1998). Based on IR sequences reported for pSD10 (Zhong et al. 2002), we checked the pLMA1 fragments and identified a 24-bp region upstream of ORF 4 of pP37-74 (5′-GGACTGACGCACGTGTAGGTGACA-3′) differing in six (underlined) positions from the IR (5′-GGACTGGTGTACACATAGGT-GACA-3′) found upstream of MC35. Despite the lack of the corresponding IR, which is probably located downstream of ORF 5 (pP37-74) within yet unknown pLMA1 sequence, this finding suggests ORFs 4 and 5 (pP37-74) being part of an IS element. Although the cloned pLMA1 fragments are rich in genes encoding putative transposases, further IRs were not identified. However, for some bacterial genomes and plasmids IS elements lacking IRs have been reported (Burland et al. 1998).

IS elements and transposons are frequently associated with antibiotic resistance genes (Varaldo et al. 2009). *M. luteus* A1, hosting pLMA1, is resistant to a number of antibiotics, particularly to macrolides, such as erythromycin, for which plasmid curing experiments suggested a link to pLMA1 (Dib et al. 2010a). Resistance to erythromycin is either conferred by a 23S rRNA methylase-mediated target site modification or by an efflux system; the latter resulting in low-level resistance (Varaldo et al. 2009). Since the minimal inhibitory concentration of erythromycin for *M. luteus* A1 wild-type reached a level > 256 μg ml^{-1} (Dib et al. 2010a), a methylase-mediated resistance mechanism is likely to be encoded. As we were not able to identify an ORF coding for a 23S rRNA methylase, such function is presumably located on the yet unknown pLMA1 sequence.

pLMA1 is laced with short repetitive sequences

To verify that repetitive sequences are responsible for the failure of the 454 read assembly, we looked for repeats within the nucleotide sequences of the cloned pLMA1 fragments. Indeed, a huge number of repetitive sequences (mostly ranging from 6 to 12 nucleotides) were found. Such repetitions occur at high frequency and are spread quite evenly over the pLMA1 fragments; partially they overlap each other. Conspicuously arranged repeats were seen on the 6,195-bp *Pst*I fragment; we identified five direct repeats (putative iterons) starting at nucleotide positions 2510, 2571, 2632, 2693, and 2754, respectively. Iterons 1–4 are perfect direct repetitions sharing a 29-bp consensus sequence (5′-GGAAGCCCCGCGCACGCAGGGATGAGCCC-3′); only iteron 5 (5′-GGAAGCTCCGCCCGCGCAGGGATGAGCCA-3′) differs at four (underlined) positions. All five iterons are regularly spaced by 32 bp.

Such iteron sequences are characteristic for replication origins of linear plasmids such as pCLP of *Mycobacterium celatum* (Le Dantec et al. 2001) and many *Streptomyces* linear plasmids such as pSLA2 (Chang et al. 1996) and pRL2 (Zhang et al. 2008). Iterons, as the initiation sites of plasmid DNA replication, usually are located adjacent to *rep* genes: *rep*1 encoding a DNA-binding protein and *rep*2 a DNA helicase (Chang et al. 1996; Zhang et al. 2009); though, other recently described *rep* genes are clearly discriminable from the above (Zhang et al. 2006). For pLMA1, ORFs adjacent to the iterative sequences do not shown any similarity to known *rep* loci. However, iterons and genes involved in plasmid replication may be separated by nonessential genes, as for the *Streptomyces* linear plasmid pRL2 (Zhang et al. 2009), or by noncoding sequences as for pSHK1 (Zhang et al. 2009). Indeed, studies on the structure of replication loci of *Streptomyces* linear plasmids in general revealed an unexpected variety of components and their positions (Zhang et al. 2009).

However, the role of the putative iteron sequences of pLMA1 as well as other functions located outside of the sequenced fragments, possibly involved in plasmid replication, needs to be elucidated.

The arrangement of repetitive sequences is exemplarily shown for the 3,681-bp pLMA1 fragment (pP37-74) in comparison to the 3,721-bp chromosomal fragment of *M. luteus* A1 (pP37-62)

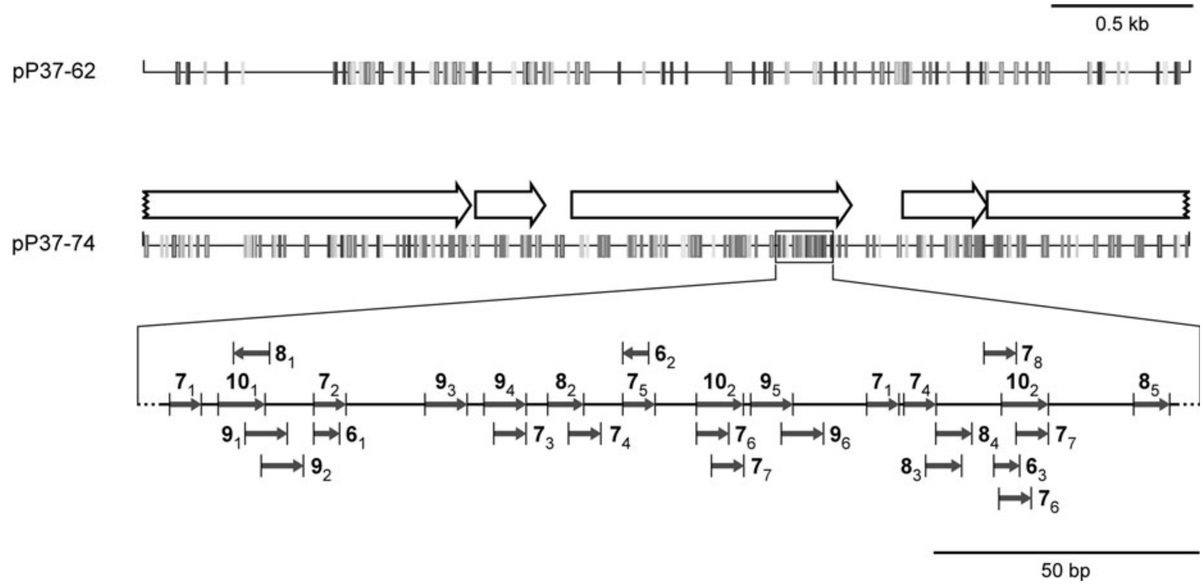

Fig. 3 Schematic map indicating identified repetitive sequences. The chromosomal fragment of *M. luteus* A1 (pP37-62) and a fragment of pLMA1 (pP37-74) are shown as *horizontal*, *black lines*. Annotated ORFs of the pLMA1 fragment are denoted as *open arrows*. *Arrows* missing the beginning or the *arrowhead* (*serrated lines*) show interrupted ORFs. *Vertical lines* in different *gray scales* indicate repetitive sequences. In the enlarged section shown below, such sequences are exhibited in more detail. Length and orientation of the repetitive sequences are marked by *gray arrows*. Repeat lengths are earmarked by *bold numbers*. Identical repeats are designated by the same subscript. The 0.5-kb size bar is for pP37-62 and pP37-74, the 50-bp bar is for the enlarged section only

Table 2 Frequency of selected sequence repeats on pP37-74

Sequence	Length (nt)	Frequency observed	Frequency expected	Ratio[a]
CCCGCC	6	6	6.9×10^{0}	0.9×10^{0}
GCGGTG	6	10	1.9×10^{0}	5.2×10^{0}
TCGACG	6	8	1.2×10^{0}	6.5×10^{0}
CGGCTGG	7	4	7.0×10^{-1}	5.7×10^{0}
CGCGAAG	7	3	4.6×10^{-1}	6.5×10^{0}
GAACCGG	7	4	4.6×10^{-1}	8.7×10^{0}
CACCACCG	8	3	2.1×10^{-1}	1.4×10^{1}
TCGAGCAG	8	3	6.8×10^{-2}	4.4×10^{1}
AGCTCGGCA	9	2	2.5×10^{-2}	8.1×10^{1}
GCGAACCTG	9	3	2.5×10^{-2}	1.2×10^{2}
GGGGCCGTGT	10	2	1.0×10^{-2}	1.9×10^{2}
CTGCCACGAG	10	2	8.8×10^{-3}	2.3×10^{2}
GATGCCGGCCC	11	2	5.7×10^{-3}	3.5×10^{2}
GTGATCAACGC	11	2	6.4×10^{-4}	3.1×10^{3}
CGCGCCGGCACG	12	2	4.3×10^{-3}	4.6×10^{2}

[a] Frequency observed divided by frequency expected

(Fig. 3). Though the chromosome also has sequence repeats, their number is lower and they are irregularly arranged (Fig. 3). To appraise the observed frequency of the single identified repetitive sequences, we calculated their expected frequency and compared both values (Table 2). This calculation, based on length and nucleotide composition of the DNA fragment, clearly shows an over-representation of the sequence repeats; deviation from the expectation becomes higher the longer the repeat is (Table 2).

Such dispersed repetitions along with the short read lengths (see above) can be considered the major reason for the failure of the 454 read assembly. Additives for high-GC DNA as well as an optimized sequencing system (Titanium Series, Roche), becoming only quite recently available, may allow for longer reads of up to 400–500 bases. However, despite such improvements, assembly of reads of pLMA1 still remains uncertain. As Sanger/pyrosequencing hybrid approaches were already successfully used for the sequencing of bacterial genomes, we decided to apply conventional Sanger sequencing for pLMA1 which currently is in progress.

Acknowledgments We gratefully acknowledge the support by PICT-Agencia Nacional de Promoción Científica y Tecnológica. Julián R. Dib was a recipient of a CONICET and DAAD fellowship.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3): 403–410

Armstrong K, Bauer WR (1982) Preferential site-dependent cleavage by restriction endonuclease *Pst*I. Nucleic Acids Res 10(3):993–1007

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2004) Genbank: update. Nucleic Acids Res 32(Database issue):D23–D26

Burland V, Shao Y, Perna NT, Plunkett G, Sofia HJ, Blattner FR (1998) The complete DNA sequence and analysis of the large virulence plasmid of *Escherichia coli* O157:H7. Nucleic Acids Res 26(18):4196–4204

Chang PC, Kim ES, Cohen SN (1996) *Streptomyces* linear plasmids that contain a phage-like, centrally located, replication origin. Mol Microbiol 22(5):789–800

Dib J, Motok J, Zenoff VF, Ordonez O, Farias ME (2008) Occurrence of resistance to antibiotics, UV-B, and arsenic in bacteria isolated from extreme environments in high-altitude (above 4400 m) Andean wetlands. Curr Microbiol 56(5):510–517

Dib JR, Wagenknecht M, Hill RT, Farias ME, Meinhardt F (2010a) First report of linear megaplasmids in the genus *Micrococcus*. Plasmid 63(1):40–45

Dib JR, Wagenknecht M, Hill RT, Farias ME, Meinhardt F (2010b) Novel linear megaplasmid from *Brevibacterium* sp. isolated from extreme environment. J Basic Microbiol 50(3):280–284

Goldberg SM, Johnson J, Busam D, Feldblyum T, Ferriera S, Friedman R et al (2006) A sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci USA 103(30):11240–11245

Le Dantec C, Winter N, Gicquel B, Vincent V, Picardeau M (2001) Genomic sequence and transcriptional analysis of a 23-kilobase mycobacterial linear plasmid: evidence for horizontal transfer and identification of plasmid maintenance systems. J Bacteriol 183(7):2157–2164

Mahillon J, Chandler M (1998) Insertion sequences. Microbiol Mol Biol Rev 62(3):725–774

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057): 376–380

Meinhardt F, Klassen R (2007) Microbial linear plasmids. Springer-Verlag, Berlin

Ohama T, Muto A, Osawa S (1989) Spectinomycin operon of *Micrococcus luteus*: evolutionary implications of organization and novel codon usage. J Mol Evol 29(5):381–395

Pearson WR (1994) Using the FASTA program to search protein and DNA sequence databases. Methods Mol Biol 25:365–389

Sharma D, Issac B, Raghava GP, Ramaswamy R (2004) Spectral repeat finder (SRF): identification of repetitive sequences using Fourier transformation. Bioinformatics 20(9):1405–1412

Varaldo PE, Montanari MP, Giovanetti E (2009) Genetic elements responsible for erythromycin resistance in streptococci. Antimicrob Agents Chemother 53(2):343–353

Wagenknecht M, Meinhardt F (2010) Copy number determination, expression analysis of genes potentially involved in replication, and stability assays of pAL1—the linear megaplasmid of *Arthrobacter nitroguajacolicus* Rü61a. Microbiol Res. doi:10.1016/j.micres.2009.12.005

Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B, Stein N (2006) 454 Sequencing put to the test using the complex genome of barley. BMC Genomics 7:275

Yang CC, Huang CH, Li CY, Tsay YG, Lee SC, Chen CW (2002) The terminal proteins of linear streptomyces chromosomes and plasmids: a novel class of replication priming proteins. Mol Microbiol 43(2):297–305

Zhang R, Yang Y, Fang P, Jiang C, Xu L, Zhu Y et al (2006) Diversity of telomere palindromic sequences and replication genes among *Streptomyces* linear plasmids. Appl Environ Microbiol 72(9):5728–5733

Zhang R, Xia H, Guo P, Qin Z (2009) Variation in the replication loci of *Streptomyces* linear plasmids. FEMS Microbiol Lett 290(2):209–216

Zhong Z, Caspi R, Mincer T, Helinski D, Knauf V, Boardman K et al (2002) A 50-kb plasmid rich in mobile gene sequences isolated from a marine *Micrococcus*. Plasmid 47(1):1–9