A group I intron in the terminase gene of *Lactobacillus delbrueckii* subsp. *lactis* phage LL-H

Merja Mikkonen and Tapani Alatossava

An 837 nt long group IA intron was discovered in the *Lactobacillus delbrueckii* subsp. *lactis* virulent phage LL-H genome. The LL-H intron conforms well to the secondary structure that is common to all group I introns. The only exception is that the extreme 3' nucleotide of the intron is an A residue instead of the usual G; despite this the intron is efficiently spliced *in vivo*. This LL-H intron contains an ORF, ORF168, which shows homology with endonucleases encoded by ORFs contained in *Bacillus subtilis* phage introns. At present, the LL-H intron is the only one found in the phages of lactic acid bacteria and the first one to be found in a phage belonging to the most abundant taxonomic group, group B or Siphoviridae. The LL-H intron interrupts gene *terf*, the product of which (505 kDa, *Terf*) is significantly homologous to the large subunit of *B. subtilis* phage SPP1 terminase. The product of the upstream gene, *terS* of LL-H (15.9 kDa, *TerS*), shows homology to small subunits of *B. subtilis* terminases.

**Keywords**: group I intron, RNA structure, terminase, endonuclease, lactic acid bacteria, phage evolution

INTRODUCTION

Interrupted genes are found in all classes of organisms (Lewin, 1994). Most genes in higher eukaryotic genomes are interrupted, but in lower eukaryotes and especially in prokaryotes only a minority contains introns (Lewin, 1994). Introns can be divided into three general classes: nuclear introns, group I introns and group II introns (Lewin, 1994). Group I introns have a wide phylogenetic distribution (Lambowitz & Belfort, 1993). The majority of them have been discovered in fungal mitochondrial DNAs (Cech, 1988; Michel & Westhof, 1990). They also occur in plant mitochondria and chloroplasts, ciliate, slime mould, algal and fungal nuclei, and eubacterial and phage genomes (Cech, 1988; Michel & Westhof, 1990; Lambowitz & Belfort, 1993). Three introns have been found in the *Escherichia coli* phage T4 (Chu et al., 1984; Gott et al., 1986) and one in each of the following *Bacillus subtilis* phages SP01, (Goodrich-Blair et al., 1990) SP82, 2C, Ò (Goodrich-Blair & Shub, 1994) and 622 (Bechhofer et al., 1994). Because of the small number of characterized introns in eubacteria and their phages, it is impossible to draw firm conclusions about their origin and distribution. In addition to phage introns, only a few examples of group I introns have been discovered in the tRNA genes in cyano- and proteobacteria ("purple bacteria") (Kuhsel et al., 1990; Xu et al., 1990; Reinhold-Hurek & Shub, 1992; Biniszkwicz et al., 1994). For this reason, any new eubacterial group I intron will help to understand the role, distribution and origin of these genetic elements.

Group I introns catalyse their own splicing from mRNA via a series of transesterification reactions that are initiated by exogenous guanosine or one of its phosphorylated forms (Cech, 1990). Group I introns don’t generally share extensive sequence homology with each other, but they all have a common highly conserved secondary and tertiary structure (Cech; 1988; Michel & Westhof, 1990). The secondary structure contains both local and long-range pairing regions that generate short helices, and highly conserved sequence elements at defined locations (Cech, 1988; Michel & Westhof, 1990). Part of the secondary structure, together with other conserved nucleotides, fold into tertiary structure, forming the catalytic core of the intron, which is crucial for self-splicing (Cech, 1988, 1990; Michel & Westhof, 1990).
The size of group I introns can vary from about 200 to 3000 nt, mainly depending on whether or not they contain ORFs (Lambowitz & Belfort, 1993). Some of the ORFs overlap core structure elements, but most are located in the peripheral loops (Lambowitz & Belfort, 1993). All but one of the phage introns contain an ORF that is mostly looped out of the core structure (Chu et al., 1984; Gott et al., 1996; Goodrich-Blair et al., 1990; Lambowitz & Belfort, 1993; Bechhofer et al., 1994). Most of the intron-encoded proteins function either as maturases that promote the splicing of their host intron, or act as site-specific endonucleases that make the intron mobile (Lambowitz & Belfort, 1993).

Lactic acid bacteria, especially Lactobacillus and Lactococcus species, are widely employed in various food and feed industries and biotechnology processes. Lactic acid bacteria are Gram-positive, and belong more precisely to the phylogenetic group of low G + C-content Gram-positives to which B. subtilis also belongs (Olsen et al., 1994). Phylogenetically, lactic acid bacteria are quite distant from the Gram-negative E. coli (Olsen et al., 1994).

Phage LL-H is a virulent phage of Lactobacillus delbrueckii subsp. lactis that was first isolated in a Finnish dairy plant in 1972 (Alatossava & Pyhtila, 1980). With its small isometric head and long non-contractile tail, LL-H is a typical group B1 phage (Forsman & Alatossava, 1991). LL-H is the most studied representative of the DNA homology group, which contains most of the known Lactobacillus phages, both virulent and temperate (Mata et al., 1986; Sechaud et al., 1988). The LL-H genome is circularly permuted and terminally redundant double-stranded DNA (34.6 kb) (Forsman et al., 1986). So far about 60% of the LL-H genome sequence has been published (Vasala et al., 1993; Mikkonen & Alatossava, 1994; GenBank M96254).

In this paper we describe a group I intron present in the terminase gene of phage LL-H of Lb. delbrueckii subsp. lactis. This is the first time an intron has been discovered in a phage belonging to the taxonomic group B (or Siphoviridae), the most abundant phage group (Ackermann & DuBow, 1987).

**METHODS**

**Bacteria and phages.** Lb. delbrueckii subsp. lactis LKT was used as the propagation strain for phage LL-H. It was grown at 37 °C in MRS broth (Difco) that was supplemented with 20 mM CaCl₂. Phage propagation, concentration and purification were carried out as described previously (Forsman & Alatossava, 1991). E. coli JM101 was grown in 2 x TY broth and used as the host for M13 cloning vectors (Sambrook et al., 1989).

**Isolation and manipulation of phage LL-H DNA.** Phage particles were purified in a CsCl gradient as described by Alatossava & Pyhtila (1980). Phage and vector DNA extractions, restriction endonuclease digestions, agarose gel electrophoresis, DNA ligations and E. coli transformations were performed by standard methods (Sambrook et al., 1989).

**Phage LL-H RNA extraction.** Total RNA was isolated from phage LL-H-infected Lb. delbrueckii subsp. lactis LKT cells 60 min after infection and from non-infected LKT cells. Rifampicin (200 µg ml⁻¹) was used to inhibit RNA synthesis, and cells were collected by centrifugation. The cell pellet was suspended in 100 µl lysin buffer [10 mM MgCl₂, 250 mM NaCl, 50 mM sodium phosphate buffer (pH 5.5), 10% glucose (w/v)] and 5 µl [10 U µl⁻¹] purified LL-H lysin enzyme (kindly provided by Antti Vasala & Merja Valkkila, University of Oulu, Finland) was added. The samples were incubated for 10 min at room temperature, after which the total RNA was purified with the modified detergent/phenol method (Slater, 1984; Vasala et al., 1993).

**cDNA synthesis and PCR-amplification.** Total RNA isolated from LKT cells was used as a template to synthesize cDNA with 1 µl primer A (10 mM)[5'-AGTCTTAAGTGTCA-3'] using SuperScript Preamplification System (Gibco BRL) with modifications described by Beresford et al. (1993). The cDNA obtained with SuperScript reverse transcriptase was enzymically amplified by PCR and a GeneAmp-kit from Perkin Elmer Cetus, used as specified by the manufacturer. For amplification 5 µl primer A (10 mM) and 5 µl primer B (10 mM)[5'-AGTCTGAGGCTTCTCGTT-3'] were used. The amplification procedure used 30 cycles of denaturation at 92 °C for 1 min, annealing at 64 °C for 1 min and extension at 72 °C for 2 min. The programme included a pre-incubation at 92 °C for 2 min before the first cycle, and after the last cycle an incubation at 72 °C for 10 min that was followed by cooling to 4 °C, which was also the storing temperature. Similarly, 200 ng phage LL-H DNA was amplified with the same protocol and primers.

**DNA sequencing.** The nucleotide sequence of phage LL-H was determined independently in both directions by the method of Sanger et al. (1977). Phage LL-H DNA was sequenced from cloned fragments with Sequence version 2.0 (USB). The cycle sequencing technique modified from the method of Murray (1989) and TAQuence version 2.0 (USB) were used to sequence the PCR-amplified cDNA and to confirm the LL-H nucleotide sequence with direct sequencing from the phage LL-H DNA. The sequencing reaction products were stored at 4 °C. They were separated on a standard 6% (w/v) acrylamide/7 M urea gel (Sambrook et al., 1989).

**Sequence analysis.** Computer analyses were performed with the Genetics Computer Group (GCG) sequence analysis software package version 8-UNIX (Devereux et al., 1984), LKB DNASIS (V7.0) and LKB PROSIS (V6.02). Databases were searched with the nucleotide sequence and for the ORFs with both the nucleotide and the amino acid sequences. Database searches were performed with programs FASTA (Pearson & Lipman 1988) and BLAST (Altschul et al., 1990).

**RESULTS AND DISCUSSION**

The phage LL-H genome contains a group IA intron that is efficiently spliced in vivo

During the phage LL-H genome sequencing project the ORF168 was found to share significant amino acid homology with several intron-encoded ORFs of B. subtilis phages (Goodrich-Blair et al., 1990; Goodrich-Blair & Shub, 1994). Analysis of the nucleotide sequence confirmed that the phage LL-H genome could contain a group I intron. The ORF168 and the putative intron were located in a 3 kb PsiI-SalI fragment (Figs 1 and 2; GenBank L37351) that lies between the par site (Forsman & Alatossava, 1991) and the gene g37 that encodes a minor structural protein gp61 of LL-H (GenBank
Group I intron in phage LL-H terminase gene

Fig. 1. Organization of the phage LL-H genome region containing the LL-H intron and terminase genes. In the phage LL-H restriction map (Forsman & Alatossava, 1991) the genome is shown linearized at its packaging site (pac). The previously characterized structural-protein-encoding genes (Mikkonen & Alatossava, 1994) and the lysin gene (lysA) of phage LL-H (M96254) are marked with arrows beneath the map. The 3.1 kb SalI-PstI fragment is shown in more detail below the map. The horizontal arrows show the orientation, position and size of putative genes as derived from the nucleotide sequence presented in this study. At the beginning of an arrow, an open circle denotes an RBS (Mikkonen et al., 1994). The vertical arrowheads indicate the intron splice sites. P, PstI; S, SalI.

The secondary structure predicted for the phage LL-H intron conforms well to that of all group I introns (Burke et al., 1987) and is shown in Fig. 3. The secondary structure forms the intron core and is essential for the splicing (Cech, 1988, 1990; Michel & Westhof, 1990). It consists of pairing elements from P1 to P10 and of four conserved sequence elements that always occur in the same order, 5'-PQRS-3' (Burke et al., 1987; Cech, 1988). In the LL-H intron, the sequences of these elements match the consensus well (Cech, 1988). All conserved pairing elements are present in the LL-H intron, except that it does not contain the P2 element, which is found in only about two-thirds of group I introns (Michel & Westhof, 1990). For example, the T4 phage intron in the gene sunY has no P2 (Shub et al., 1988).

Between stems P7 and P3 of the LL-H intron, two extra stem–loop structures, P7.1 and P7.2, could be predicted (Fig. 3). This feature is characteristic for introns belonging to the various IA subgroups (Cech, 1988; Michel & Westhof, 1990). In subgroups IA2 and IA3, there are two stem–loops but in IA1 there is only one (Michel & Westhof, 1990). Based on this, the LL-H intron seems to belong to subgroup IA2 or IA3 (Michel & Westhof, 1990). The other phage introns so far described belong to subgroup IA2 (Michel & Westhof, 1990; Bechhofer et al., 1994; Goodrich-Blair & Shub, 1994).

The 5′ splicing site of group I introns lies in the pairing region P1 formed by the 3′ end of the 5′ exon and the 5′ end of the intron sequence (Cech, 1988, 1990; Michel & Westhof, 1990). The splicing site is nearly always preceded by a U that pairs with a G (Cech, 1988, 1990; Michel & Westhof, 1990). The exon–intron boundary of the LL-H intron was located before the U of the last U-G pair (nt 1424 and 1432 in Fig. 3).
**Fig. 2.** Nucleotide sequence of the 3116 bp PstI-Sall fragment (this study, GenBank L37351) and 96 bp from the 5' end of genome region encoding structural proteins of phage LL-H (Mikkonen & Alatossava, 1994). It contains the LL-H intron and the putative terminase genes \( \text{terS} \) and \( \text{terL} \), which consists of two exons \( \text{terL1} \) and \( \text{terL2} \). The proteins encoded by the LL-H sequence are marked beneath the nucleotide sequence in standard one letter amino acid code. Aligned to the amino acid sequences are examples of homologous amino acid sequences found in databases.
Fig. 3. Proposed secondary structure of the LL-H intron. Arrows indicate the 5' and 3' splice-site sequences, determined by sequencing of the PCR-amplified cDNA synthesized from mRNA. Lowercase letters denote the exon sequences and the capital letters denote the intron sequence. Conserved secondary structure elements of group I introns are labelled with numbered circles, also marked are the conserved sequence elements P, Q, R and S (Burke et al., 1987; Cech, 1988). Lines have been added to make the figure less crowded; no nucleotides have been omitted except where specifically marked (loops L6b and L7.1). The start codon (AUG) of ORF168 is marked in the L6b and the stop codon (UAA) is boxed. The formation of the P9.0 and P10 is shown in the insert in the right lower corner (the situation depicted is just before the ligation step). The secondary structure was determined manually with the help of GCG programs RNAFOLD, SQUIGGLES, STEMLOOP and REPEAT. The numbering of nucleotides is the same as in Fig. 2.

It was evident from the sequence of the ligated exons that the 3' splicing site of the LL-H intron was located before nt 2261 (Figs 2 and 3). The extreme 3' nucleotide of the LL-H intron was an A, which was confirmed by direct PCR sequencing from the LL-H genomic DNA. In all the previously reported group I introns, the nucleotide preceding the 3' splicing site has always been a G residue (Cech, 1988; Michel & Westhof, 1990). However, it has been shown by mutational analysis of the Tetrahymena thermophila intron that if the 3' terminal G residue has been changed into an A, the 3' splicing is slightly delayed but still remains accurate (Michel et al., 1989). The four intron nucleotides (3'-GCUU-5') immediately preceding the G residue pairing with the last U of the 5' exon can pair with the first four nucleotides of the 3' exon (5'-CGAA-3') forming P10 of LL-H intron (Fig. 3). The intron sequences of P1 and P10 form the internal guide sequence that pairs with the exons and aligns them for ligation (Michel et al., 1989). In the LL-H intron P9.0 could be formed by two A-U pairs (nt 2205-2206 and 2258-2259 in Fig. 3). Most often the P9.0 pairing contains two base pairs, though it is not present in all group I introns (Burke, 1989; Michel et al., 1989; Michel & Westhof, 1990). If neither P9.0 nor P10 is present, the 3' splicing is disrupted, but the presence of even one of them restores it (Michel et al., 1989). Because the LL-H intron contains optimize the alignments are indicated by dots in the nucleotide sequence and by dashes in the amino acid sequence. Identical amino acids are in boxed capital letters and functionally similar amino acids are in capital letters. The putative amino acid sequence motifs are underlined and marked in figure. Termination codons are marked with asterisks. Vertical lines indicate the 5' and 3' splice-site sequences of the 837 nt intron. The putative RBSs are underlined in the nucleotide sequence and the PstI and Sall recognition sites are boxed. The division of amino acids into groups of functional similarity was: 1, small, non-aliphatic, non-charged (Ala, Ser, Thr, Pro, Gly); 2, polar, not-positively charged (Asn, Asp, Glu, Gln); 3, positively charged (His, Arg, Lys); 4, aliphatic and methionine (Met, Leu, Ile, Val); 5, aromatic, non-charged (Phe, Tyr, Trp) (LKB PROSIS, V6.02).
both the P10 and P9.0 pairings, it may help the LL-H intron to splice correctly and efficiently in vivo, as shown above by PCR amplification and sequencing, even though the 3' terminal residue of the LL-H intron is an A instead of the common G.

The LL-H intron contains an ORF of 504 nt that potentially encodes a protein of 168 amino acids (Fig. 2), the calculated molecular mass of which is 20.0 kDa. ORF168 mostly lies outside the secondary structure essential for group I intron splicing. It begins in the large peripheral loop of P6 (L6b) and its 3' end forms the 3' side of stem P6 and the stop codon is located in the very conserved joining section J6/7 (Fig. 3). The initiation codon of ORF168 and its putative RBS (Mikkonen et al., 1994) are located in a region separate from the secondary structure required for splicing (Fig. 3).

All but one of the previously described phage introns contain an ORF that is typically located in one of the peripheral loops (Shub et al., 1988; Goodrich-Blair et al., 1990; Goodrich-Blair & Shub, 1994; Bechhofer et al., 1994). The amino acid sequence translated from LL-H ORF168 was used as a template to search through various databases. Statistically significant amino acid alignments were obtained with the intron-encoded ORFs from the B. subtilis phages SP01, SP82, 2C and Oe (Goodrich-Blair et al., 1990; Goodrich-Blair & Shub, 1994), and from a free-standing ORF of B. subtilis phage SPPI (Pedre et al., 1994). All homologies obtained to the ORF168 product were quite similar and concentrated on the amino-terminal part of each protein (e.g. SP01, 42.1% identity and 61.1% similarity in the first 95 amino acids; Fig. 2). The intron-encoded proteins of the B. subtilis phages SP01 and SP82, are known to be site-specific DNA endonucleases (Goodrich-Blair & Shub, 1994). The intron-encoded endonucleases contain a recently described HI-NL-H motif (Goodrich-Blair & Shub, 1994; Shub et al. 1994), which also was found in the product of ORF168 (Fig. 2). The same motif is found in a bacterial restriction endonuclease and in the Zn-finger-like domains of several group II intron ORFs (Goodrich-Blair & Shub, 1994; Shub et al., 1994). The similarity of the LL-H intron-encoded ORF168 to the B. subtilis phage intron-ORFs suggests that its product is also an endonuclease. The high similarity of the amino acid sequences in their amino-terminal part may reflect structures that are needed for the formation of the active site of the endonuclease, while the highly variable carboxyl termini could be involved in DNA target site recognition (Goodrich-Blair & Shub, 1994).

Phage LL-H belongs to the taxonomic group B, Siphoviridae, phages with long non-contractile tails (Ackermann & DuBow, 1987; Forsman & Alatossava, 1991). The Siphoviridae family is the most abundant of phage groups; over 50% of all phages reported up to 1987 belong to this group (Ackermann & DuBow, 1987). All the previously reported group I introns have been found in phages belonging to taxonomic group A, Myoviridae, phages with contractile tails (Ackermann & DuBow, 1987), and having large and complex genomes (Chu et al., 1984; Gott et al., 1986; Ackermann & DuBow, 1987; Mosig & Eiserling, 1988; Stewart, 1988; Goodrich-Blair et al., 1990; Goodrich-Blair & Shub, 1994; Bechhofer et al., 1994). Compared to these group A phages, LL-H has a much smaller genome, 34.6 kb (Forsman & Alatossava, 1991), being only about one-quarter of that of the previously described intron-containing phages. With the present discovery of the LL-H intron, group I introns have now been found in the two distinct phage taxonomic groups A and B, to which about 80% of all known phages belong (Ackermann & DuBow, 1987). These discoveries could mean that bacterial introns might be much more common than previously suspected.

**The LL-H intron interrupts a terminase gene**

To understand the possible biological role of the LL-H intron one should know about the interrupted gene and how intron splicing could affect its function. The nucleotide sequence of the 3·1 kb Psrl-SalI fragment containing the LL-H intron was analysed for possible coding regions. In the main coding strand three coding regions, two upstream of and one downstream of the LL-H intron, were detected (Figs 1 and 2). The amino acid sequences translated from the three coding regions (terS, terL1 and terL2 in Figs 1 and 2) were used as templates in database searches. Statistically significant amino acid alignments were obtained with terminase enzyme subunits of B. subtilis phages SPPI, SF6 and p15 (Chai et al., 1992, 1994). Like LL-H they are pac-type phages, and the products of g1 (gp1) and g2 (gp2), encoding the small and large subunits of the SPPI-family of terminases (Chai et al., 1992, 1994), are essential for the pac-site cleavage. The small terminase subunits (gp1) of all three B. subtilis phages are about equally homologous to TerS of phage LL-H. The large terminase subunit (gp2) has been characterized only from phage SPPI (Chai et al., 1992). The amino-terminal part of TerS is 43.8% identical and 62.0% similar to the first 121 amino acids of gp1 of SPPI. The TerL1 is 30.4% identical and 55.1% similar to the 133 amino-terminus of gp2 of SPPI and TerL2 is 23.8% identical and 45.0% similar to the 289 carboxy-terminal amino acids of gp2 of SPPI. terS potentially encodes a protein of 142 amino acids, the calculated molecular mass of which is 15.9 kDa. Removal of the LL-H intron and exon (terL1 and terL2) ligation form the combined ORF terL, which potentially encodes the TerL protein of 433 amino acids with calculated molecular mass 50.5 kDa. Additional evidence for the biological activity of the LL-H intron is that the formation of TerL, homologous counterpart of gp2 of phage SPPI, is possible only after the LL-H intron self-splices terS mRNA.

Phage terminase genes characteristically overlap by a few nucleotides and the DNA interaction sites are either within, or close to and upstream of, genes encoding the terminase proteins (Black, 1989; Chai et al., 1992; Powell et al., 1990; Feiss, 1986). Phage LL-H terminase genes terS and terL overlap by eight nucleotides and the LL-H pac site has been located about 1·2 kb upstream from terS.
Group I intron in phage LL-H terminase gene

(Forsman & Alatossava, 1991). Phage terminases are hetero-oligomers consisting of small and large subunits (Black, 1989). The small subunit is important for the process recognition and DNA binding, and it also appears to bind and hydrolyse ATP (Black, 1989). A putative NTP binding motif (AXXXXXGKL, motif A of Walker et al., 1982), DNA binding motif (helix-turn-helix) and NTP hydrolysis motif (DE, motif B of Walker et al., 1982) have been identified in the amino acid sequence of gp1 of phages SPPI, SP6 and ρ15 (Chai et al., 1992, 1994). These motifs can also be found in TerS of phage LL-H (Fig. 2). The large subunit appears to bind to the prohead and may be involved in the cutting of phage DNA molecules (Black, 1989; Feiss, 1986). TerS and TerL could represent the small and large subunits of LL-H terminase, respectively, and belong to the SPPI-family of terminases (Chai et al., 1992, 1994).

ACKNOWLEDGEMENTS

This study was supported by grants from the Ministry of Agriculture, the Academy of Finland (SA 1051317) and CEC BIOTECH program (G-project BIO2-CT94-3055). We thank Anne Karjalainen for excellent technical assistance. The purified LL-H lysis enzyme was kindly provided by our colleagues Antti Vasala PhIL. lic. and Merja Välikkilä MSc, Department of Genetics, University of Oulu.

REFERENCES


\textit{Received 20 February 1995; revised 15 May 1995; accepted 5 June 1995.}