Nucleic Acids Research, 1994, Vol. 22, No. 7 1167–1171

Evidence for a group II intron in *Escherichia coli* inserted into a highly conserved reading frame associated with mobile DNA sequences

Volker Knoop* and Axel Brennicke
Institut für Genbiologische Forschung, Ihnestraβe 63, 14195 Berlin, Germany

Received January 6, 1994; Revised and Accepted March 7, 1994

ABSTRACT

The distribution of group II introns in the living world is an important aspect of the hypothesis which postulates their evolutionary relation to the nuclear spliceosome. As an alternative to the restricted experimental approaches towards their identification we devised a strategy to recognize group II introns in sequence data. By this approach we identified a locus on a plasmid in the bacterium *Escherichia coli*. Modelling of the derived RNA secondary structure reveals the presence of perfectly conserved domains V and VI as typical features of group II introns. An intron internal reading frame upstream of domain V is homologous to group II intron encoded maturases. A reading frame downstream of the predicted 3'- splice site is highly similar to a small polypeptide encoded in the central part of the *Agrobacterium tumefaciens* T-DNA. With the TBLASTN algorithm a set of plasmid-borne insertion sequences in Agrobacteria and Rhizobia and surprisingly also in a *Yersina pseudotuberculosis* strain was identified which contain this highly conserved reading frame.

INTRODUCTION

Group II introns are widespread in the organellar genomes of fungi and plants and are characterized by a typical secondary structure of six stem-loop domains (1). A few members of this intron class reassemble on the RNA level from independent transcripts, a process coined trans-splicing. This observation has supported the hypothesis of an evolutionary relationship between group II introns and the nuclear intron-spliceosomal apparatus (2). According to this theory nuclear introns represent descendants of group II introns that have invaded the eukaryotic nucleus through the bacterial endosymbiont. The recent discovery of group II introns in a cyanobacterium and a proteobacterium has given substantial support to this hypothesis (3). In this study group II introns could not be identified in the two *Escherichia coli* strains tested (3). The elegant experimental approach was PCR-based and relied on the presence of intron-borne open reading frames (maturases) that, however, are highly divergent and occur in only some members of this intron class. In the mitochondrial genomes of higher plants, for example, only one maturase-related reading frame has been found in more than 20 different group II introns identified so far.

The identification of group II introns is generally hampered by the limited primary sequence conservation which is essentially restricted to domain V. Modeling of the six domain structure is in some cases not straightforward, and current RNA-folding computer programs are of little help. The conserved features are insufficient to derive universally applicable experimental strategies since individual introns appear to lack certain otherwise well conserved traits characteristic of the group II intron class. As an alternative approach we have therefore integrated the limited primary sequence conservation of domain V from known plant mitochondrial group II introns and derived a domain V consensus sequence (GTI, Group Two Identifier). As a query input sequence it faithfully recognizes domain V structures in sequence data using alignment programs.

An extended database search with GTI has identified an *Escherichia coli* plasmid sequence among database entries encompassing known group II introns. The presence of a reading frame with similarity to a group II intron maturase upstream of the identified domain V and of an equally well conserved domain VI structure downstream of this site strongly indicate the presence of a group II intron. The prediction of the 3' splice site allows the identification of the presumptive intron's host gene which is highly similar to a small reading frame associated with mobile DNA sequences in bacteria.

RESULTS AND DISCUSSION

Identification of group II introns with GTI

The domain V consensus 5'-GAGCCGTRTGANRGGNRA-CBNBCACGTNCGGTTC-3' (GTI) has been derived initially as a tool for the analysis of novel plant mitochondrial sequence data (4). Domains V from plant mitochondrial group II introns in the genes *cox*2 (5), *nad*1 (6,7), *nad*4 (8) and *nad*5 (9) were aligned excluding obvious insertions and deletions. Consensus nucleotides were assigned using the IUPAC ambiguity code if 15 out of 20 sequences conform in a given position. An N was

*To whom correspondence should be addressed

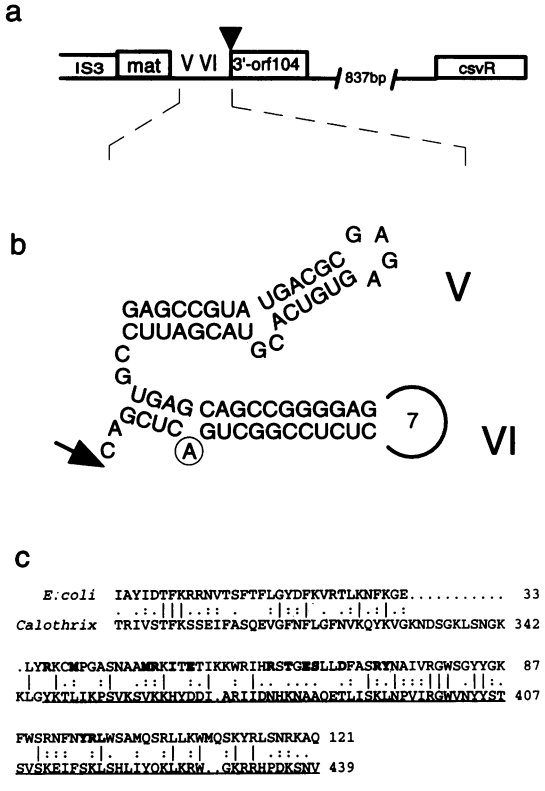

Figure 1. A group II intron structure identified upstream of the *csv*R gene in an enterotoxic *E.coli* strain of serotype O167:H5. **a**, IS3 denotes a stretch of 205 nucleotides highly similar to an *E.coli* IS3 element; mat denotes a reading frame of 121 amino acids with homology to group II intron-encoded maturases which is aminoterminally disrupted by the IS3 sequence. The gene into which the intron sequence is inserted is identified as orf104 (see text). **b**, Secondary structure modelling of domains V and VI shows the typical features of group II introns. The splice site is indicated by an arrow and the bulging A residue for lariat formation is circled. **c**, Amino acid alignment of the *E.coli* maturase reading frame with the maturase of the recently identified *Calothrix* intron (3) produced by the GAP program of the UWGCG package with default settings. Numbering for the *Calothrix* sequence is taken from the database entry. Underlining indicates the extension of domain X and bold letters indicate positions where the *E.coli* sequence corresponds better to the domain X consensus than the *Calothrix* sequence.

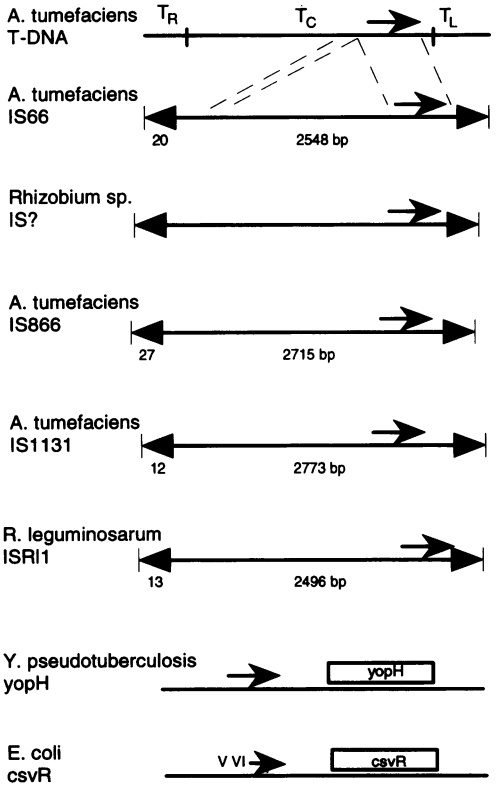

Figure 2. Location of orf104 homologues (arrows) in a subset of the loci compiled in table 1. A part of the central region of the *Agrobacterium tumefaciens* T-DNA (T_C) is apparently derived from the IS66 element of which a central region of about 1.5 kB is deleted (18). This deletion has left the region encoding orf104 intact. T_L and T_R are the left and right parts of the T-DNA, respectively, which are transferred into the DNA of the plant hosts. The orf104 homologues are in all cases encoded near one of the IS termini with the reading frames pointing outward. Our interpretation of the Rhizobium plasmid sequence is somehow in conflict with the description of this locus in the database entry (X74068). The length of the IS sequences and their flanking inverted repeats are indicated. The distance between orf104 and the inverted repeat sequence can be as short as 13 nucleotides in the case of ISRI1. IS elements may similarly be the reason for the presence of orf104 homologues in the two human pathogens.

used for positions in which all 4 nucleotides are found, and ambiguous positions 13, 19, 20, 22 and 24 were arbitrarily assigned as R, C, B, B and A respectively. Group II introns are easily identified with alignment programs (FASTA, GAP, BESTFIT) of the UWGCG program package (10) using GTI as query sequence. For example, GTI identifies the group II introns of the completely sequenced *Marchantia polymorpha* mitochondrial genome (11). In a database search with the FASTA program (12) GTI identified besides the known group II introns a hitherto overlooked domain V structure upstream of the *cox*1 gene in pea (13) in the corresponding database entry (X14409). Preliminary data confirm the assumption that a functional group II intron is present at this location (unpublished observations).

A group II intron structure in *Escherichia coli*

Besides the previously known organellar group II introns the extended database search with GTI revealed an *Escherichia coli* sequence entry describing the *csv*R gene, a plasmid-borne gene involved in the virulence of a pathogenic *E.coli* strain (14). The identified domain V sequence is located 1 kb upstream of the *csv*R startcodon (Fig. 1a). The derived secondary structure shows all relevant domain V features (1) including two base-paired regions of 8 and 6 bp, respectively, a bulging CG dinucleotide and a purine rich terminal loop of four nucleotides (Fig. 1b). Only one mismatch (A−C) is observed in the proximal stem. This structure thus conforms better to the conserved domain V features than those of some other known functionally characterized group II introns. In addition to the well conserved domain V, the sequences beginning 2 nucleotides downstream can readily be folded into a perfect domain VI secondary structure with the bulging A residue for lariat formation in the distal part of its stem. This secondary structure terminates with the conserved RAY motif indicative of the 3′ end of group II introns and thus allows prediction of the 3′ splice site.

A maturase encoding reading frame

A protein sequence deduced from the region upstream of the identified domain V has similarity to the recently identified 'domain X' (15) of maturase proteins encoded within domain

Table 1. Homologues of orf104 identified with the TBLASTN program of the NCBI Blast e-mail server with default settings.

accession number	species	ref.	description of locus	location of orf 104	assumed frameshifts
X00493	*Agrobacterium tumefaciens*	17	T-DNA	14882-14571	
M10204	*Agrobacterium tumefaciens*	18	IS66	2239-2550	+1,-2,+1
X74068	Rhizobium sp.	unpubl.	symb. plasmid DNA	2381-2696	-1
M82888	*Agrobacterium tumefaciens*	22	IS1131	2161-2479	-2
M25805	*Agrobacterium tumefaciens*	23	IS866	2307-2615	
L19650	*Rhizobium leguminosarum*	unpubl.	ISRI1	318-26	+1
Y00551	*Yersinia pseudotuberculosis*	19	*yop*H	(306)-521	
M19352	*Agrobacterium tumefaciens*	24	*pin*F1/*pin*F2 genes	(1668) - (19904)	
Z22524	*Homo sapiens*	unpubl.	none (PCR product)	130-end	
K03313	*Convolvulus arvensis*	21	flank of Ri-plasmid insertion	(473) - (342)	
X51418	*Agrobacterium rhizogenes*	25	*vir*a gene	end - (4034)	
X60106	*Escherichia coli*	14	*csv*R gene	[728]-893	-1

Minor frameshifts are in some cases assumed to extend the orf104 reading frames as listed. Parentheses indicate the absence of conserved amino- or carboxytermini and square brackets the presence of the intervening sequence in *E.coli*.

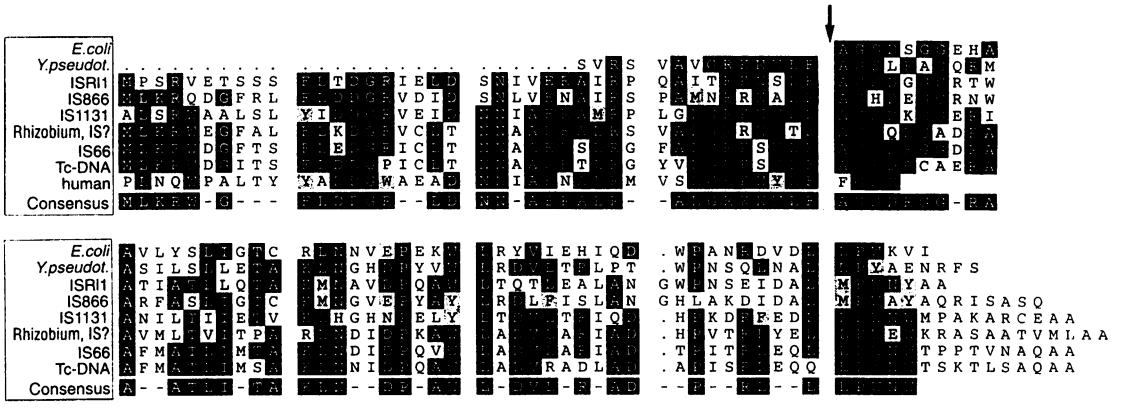

Figure 3. Alignment of the orf104 homologues encoded in the loci depicted in figure 2. The orf104 homologue identified in human DNA (Z22524) was included in the alignment, which was assembled in the one-letter-code manually with the help of the PRETTYBOX program. The orf104 reading frame in IS1131 is aminoterminally extended for three amino acids starting with a methionine. Identical residues are highlighted with black, conserved residues with grey boxes, respectively. The arrow identifies the 3'-splice site of the group II intron in *E.coli*. No single frameshift could align a conserved aminoterminal region for the *Y.pseudotuberculosis* sequence. The human DNA sequence entry ends after the last amino acid shown. Five of eight carboxyterminal sequences end with two alanine residues.

IV of some group II introns (Fig. 1c). Domain X of about 100 amino acids, located downstream of the reverse transcriptase (RTase) domains, is little conserved. It is, however, present in all known group II intron maturases and thus believed to be essential for splicing while the not generally present RTase-like domains are responsible for intron mobility (15). The *E.coli* maturase sequence is most similar to the recently identified intron-borne maturase in the cyanobacterial *Calothrix* intron (3) with 22% identical and 48% similar residues on the amino acid level. Both sequences contain the single highly conserved stretch of amino acids within the postulated domain X (Fig. 1c). The alignment is (except for the RGWxxYY stretch) somewhat arbitrary since the *Calothrix* sequence otherwise corresponds only in single amino acid positions to the domain X consensus and the *E.coli* sequence corresponds at certain positions better to other maturases previously compiled (Fig. 1c). The location 82 nucleotides upstream of domain V classify the *E.coli* maturase with the maturases lacking a zinc-finger-like domain at their carboxytermini. The gap introduced in the alignment is compatible with the assumed aminoterminal extension of domain X (Fig. 1c).

The *E.coli* maturase reading frame displays a 6:1 ratio of positively to negatively charged amino acid residues within domain X consistent with its predicted role in RNA binding. The location of the maturase-like reading frame upstream of domain V makes the otherwise low similarity to other group II intron maturases significant. Additional upstream sequence information is required to deduce the entire maturase reading frame and the 5'-splice site. The maturase reading frame is apparently interrupted by an insertion sequence of the IS3 type and it remains to be determined whether the IS3 element has led to extended rearrangements of this locus.

The intron's host gene

We then addressed the question into which gene the newly identified *E.coli* group II intron sequence is inserted. One of the reading frames downstream of the assumed splice site continues for 53 amino acids. Database searches with the BLAST Network Server at the NCBI (16) identified the hypothetical polypeptide 14 of 104 amino acids encoded in the central region (Tc) of the *Agrobacterium tumefaciens* T-DNA (17). Between the *E.coli* and *A.tumefaciens* amino acid sequences 42% of the residues are identical and 62% are similar. The probability of a random similarity is calculated to be less than 10^{-6}. We further on refer to these homologous loci as orf104.

To gain insight into the potential function of this polypeptide we used the TBLASTN program which translates available nucleic acid sequences in all six reading frames while screening

with a protein query sequence. Surprisingly, this strategy revealed 13 database entries encoding additional homologous polypeptide sequences. Probabilities for random similarity are calculated to lie in between 7×10^{-5} and 2.2×10^{-25} for these homologies. Since two pairs of sequence entries describe identical loci the total number is reduced to 11 (Table 1). The location of orf104 in the respective loci is depicted in figure 2. Four entries are identified as insertion sequences on *Agrobacterium tumefaciens* (IS66, IS1131 and IS 866) and *Rhizobium leguminosarum* (ISRl1) plasmids. Entry X74068 (Rochepeau *et al.*, unpublished) which is described as symbiotic plasmid DNA in a Rhizobium species is 64% similar to IS66 over its entire length and therefore presumably represents a related insertion sequence. This assumption is corroborated by the conservation of orf104 (see below).

The *A.tumefaciens* Tc-DNA appears to be at least partially derived from the insertion of IS-elements into the ancestor T-DNA (18). In this respect it is interesting to note that the homology of IS66 with the Tc-DNA is essentially restricted to the region encoding orf104 (Fig.2). Homologues of orf104 are thus shared by a family of insertion sequences of similar sizes (2.5−2.8 kB) carried on soil bacterial plasmids which encode functions for interaction with their plant hosts. The homologies of orf104 to sequence entries M19352 and X51418 which describe other loci on the Ti- and Ri-plasmids are most likely due to the former insertion of IS elements closely related to those shown in figure 2. The assumption of a functional importance for orf104 is corroborated by pairwise sequence alignments of the IS elements which show higher sequence conservation in orf104 than over their entire length. Moreover, nucleotide exchanges within orf104 are strongly biased towards 3rd codon positions.

Most of the conserved reading frame orf104 is also present upstream of a likewise plasmid-borne gene (*yop*H) associated with virulence of the respective *Yersinia pseudotuberculosis* strain (19). This arrangement is analogous to the *E.coli csv*R-carrying plasmid, and in both cases the presence of orf104 might represent the insertion of an IS-element related to the IS-family in plant-associated bacteria as depicted in figure 2. This assumption is at least for Yersinia corroborated by the description of a locus in *Yersinia enterolitica* (20) with 98% identity on the nucleotide level to the *yop*H gene. Homology between these two genes breaks off upstream of the coding regions, a finding compatible with the assumption that an IS element is inserted upstream of *yop*H.

The alignment of the orf104 homologues (Fig. 3) shows that the group II intron sequence in *E.coli* is inserted into the best conserved stretch of amino acids. The amino acid translations are derived from continuous reading frames in the T_C-DNA and in IS866 while reading frame shifts (Table 1) extend the polypeptide homologies in the other cases. Whether these reading frame shifts reflect sequencing errors or *in vivo* deletion/insertion mutations remains to be analyzed. The latter possibility does not exclude a functional role of orf104 since the IS elements were generally identified by comparison of IS-containing and IS-less loci and not by functional analysis. It is thus at present unclear whether all examples represent intact mobile sequences or degenerate remnants thereof. Moreover, translational frame-shifting has been shown to occur during the expression of IS-encoded genes (for review see 26).

In addition to the bacterial gene loci the TBLASTN search revealed high similarity of orf104 to two eukaryotic sequence entries. The human DNA sequence entry Z22524 (Borodin *et al.*, unpublished) potentially encodes the first 44 amino acids of an orf104 homologue. Whether the rest of the reading frame is conserved in human DNA remains open, since the homology is located at one end of this small database entry. The first 570 nucleotides of sequence entry K03313 were reported to be plant (*Convolvulus arvensis*) nuclear DNA flanking a T-DNA insertion from *Agrobacterium rhizogenes* (21). 190 bp of this region, however, are 70% similar to the 3'-region of the possible Rhizobium IS-element mentioned above (X74068) and contain the central region of the highly conserved orf104 reading frame. This observation suggests a revision of the border between the plant nuclear DNA and the inserted T-DNA.

Concluding remarks

The conserved location of homologues of the *E.coli* intron's host gene (orf104) on plasmids involved in pathogenicity and particularly the involvement of these plasmids in sequence transfers crossing species borders allows speculations on the origin of the *E.coli* intron sequence. Consequently, the possibility of horizontal sequence transfer has to be taken into account for theories on the evolutionary radiation of group II introns.

Summarizing, we have presented evidence for the presence of a group II intron in *Escherichia coli*. This finding suggests that investigation of group II intron structure and function may be possible employing the powerful methods of molecular biology in *E.coli*. The identified intron is inserted into the best conserved region of a reading frame associated with mobile DNA sequences (IS elements) on plasmids of (pathogenic) bacteria. The conservation of this reading frame (orf104) and its conserved location associated with mobile DNA sequences in a wide range of bacteria are indicative for its functional importance.

ACKNOWLEDGEMENTS

We are grateful to all computer people for creating possibilities to extract more biological information out of sequence data, especially to the people of the UWGCG and the NCBI Blast server and Sebastian Kloska and Sabine Sünkel.

REFERENCES

1. Michel, F., Umesono, K. and Ozeki, H. (1989) *Gene* 82, 5−30.
2. Sharp, P.A. (1991) *Science* 254, 663.
3. Ferat, J.-L. and Michel, F. (1993) *Nature* 364, 358−361.
4. Knoop, V. (1991) PhD thesis, Freie Universität Berlin.
5. Fox, T.D. and Leaver, C.J. (1981) *Cell* 26, 315−323.
6. Chapdelaine, Y. and Bonen, L. (1991) *Cell* 65, 465−472.
7. Wissinger, B., Schuster, W. and Brennicke, A. (1991) *Cell* 65, 473−482.
8. Lamattina, L. and Grienenberger, J.-M. (1991) *Nucl. Acids Res.* 19, 3275−3282.
9. Knoop, V., Schuster, W., Wissinger, B. and Brennicke, A. (1991) *EMBO J.* 10, 3483−3493.
10. Devereux, J., Haeberli, P. and Smithies, O. (1984) *Nucl. Acids Res.* 12, 387−395.
11. Oda, K., Yamato, K., Ohta, E., Nakamura, Y., Takemura, M., Nozato, N., Akashi, K., Kanegae, T., Ogura, Y., Kohchi, T. and Ohyama, K. (1992) *J. Mol. Biol.* 223, 1−7.
12. Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444−2448.
13. Kemmerer, E.C. and Wu, R. (1990) *Gene* 89, 157−162
14. de Haan, L.A.M., Willshaw, G.A., van der Zeijst, B.A.M. and Gaastra, W. (1991) *FEMS Microbiol. Letters* 83, 341−346.
15. Mohr, G., Perlman, P.S. and Lambowitz, A.M. (1993) *Nucl. Acids Res.* 21, 4991−4997.

16. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol.* 215, 403–410.
17. Barker, R.F., Idler, K.B., Thompson, D.V. and Kemp, J.D. (1983) *Plant Mol. Biol.* 2, 335–350.
18. Machida, Y., Sakurai, M., Kiyokawa, S., Ubasawa, A., Suzuki, Y. and Ikeda, J. (1984) *Proc. Natl. Acad. Sci. USA* 81, 7495–7499.
19. Bölin, I. and Wolf-Watz, H. (1988) *Mol. Mircrobiol.* 2, 237–245.
20. Michiels, T. and Cornelis, G. (1988) *Microb. Pathog.* 5, 449–459.
21. Slightom, J.L., Durand-Tardif, M., Jouanin, L. and Tepfer, D. (1986) *J. Biol. Chem.* 261, 108–121.
22. Wabiko, H. (1992) *Gene* 114, 229–233.
23. Bonnard, G., Vincent, F. and Otten, L. (1989) *Plasmid* 22, 70–81.
24. Kanemoto, R.H., Powell, A.T., Akiyoshi, D.E., Regier, D.A., Kerstetter, R.A., Nester, E.W., Hawes, M.C. and Gordon, M.P. (1989) *J. Bact.* 171, 2506–2512.
25. Endoh, H., Aoyama, T., Hirayama, T. and Oka, A. (1990) *FEBS Lett.* 271, 28–32.
26. Chandler, M. and Fayet, O. (1993) *Mol. Microbiol.* 7, 497–503.