Molecular Phylogenetics and Evolution 67 (2013) 246–254

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Characteristics of inteins in invertebrate iridoviruses and factors controlling insertion in their viral hosts ☆

Yves Bigot [a,*], Benoît Piégu [a], Sophie Casteret [a], Frédérick Gavory [b], Dennis K. Bideshi [c,f], Brian A. Federici [c,d,e]

[a] UMR INRA-CNRS 7247, PRC, Centre INRA de Nouzilly, 37380 Nouzilly, France
[b] CEA/Institut de Génomique GENOSCOPE, 2 rue Gaston Crémieux, CP 5706, 91057 Evry Cedex, France
[c] Department of Entomology, University of California, Riverside, CA 92521, USA
[d] Interdepartmental Graduate Programs in Cell, University of California, Riverside, CA 92521, USA
[e] Molecular and Developmental Biology, University of California, Riverside, CA 92521, USA
[f] California Baptist University, Department of Natural and Mathematical Sciences, 8432 Magnolia Avenue Riverside, CA 92504, USA

ARTICLE INFO

Article history:
Received 9 December 2012
Revised 16 January 2013
Accepted 29 January 2013
Available online 7 February 2013

Keywords:
Self-splicing
Transposable element
Host range
Nuclease
Hitchhiking
Horizontal transfer

ABSTRACT

Inteins are self-splicing proteins that occur in-frame within host-coded proteins. DNA elements coding for inteins insert specifically in highly conserved motifs of target genes. These mobile genetic elements have an uneven distribution and thus far have been found only in certain species of bacteria, archaea and fungi, a few viruses of algae and amoebozoa and in the entomopathogen, Chilo iridescent virus (CIV). Here, we report the discovery of seven new inteins parasitizing iridoviruses infecting metazoans: three within their δ DNA polymerase genes and four in genes coding for their large ribonucleotide reductase subunit. Analyses of coding sequences suggest that these inteins were acquired by ancestors shared by viruses currently classified as members of different families of viruses with large double-stranded (ds) DNA genomes and then were maintained by vertical transmission, or lost. Of significant interest is the finding that inteins present in the δ DNA polymerases of iridoviruses insert at a different location into the YGDTDS motif when compared to those found in other viruses and prokaryotes. In addition, our phylogenetic investigations suggest that inteins present in the δ DNA polymerases of these viruses might have an origin different from those found in prokaryotes. Finally, we use the sequence features of the intein insertion sites in host genes to discuss the high polymorphisms of inteins within and among viral species and the immunity of their genetic counterparts in the eukaryotic hosts of these viruses.

1. Introduction

Inteins are protein "introns" that specifically insert in-frame within the amino acid sequences of certain host proteins (Gogarten and Hilario, 2006; Barzel et al., 2011). Inteins have the ability to excise post-translationally and ligate the flanking portions of the host protein (exteins) with a peptide bond. Excision of the intein and subsequent ligation of the exteins are catalysed by the intein itself. Aside from their protein-splicing activity, which is catalysed by amino acid residues located at their N- and C-termini, many inteins contain a centrally-located site-specific DNA endonuclease domain that is thought to mediate the homing of the coding sequence into corresponding unoccupied integration sites. Most homing endonucleases (HEN) encountered in inteins belong to the LAGLIDADG family. Inteins are considered to be selfish elements as their site-specific insertion and protein-splicing activities averts deleterious effects on the activity of the host protein, and homing disseminates inteins by horizontal gene transfer.

From a mechanistic standpoint, the HEN mediates the mobility of its intein coding DNA segment by gene conversion between two heterozygous loci into two homozygous loci (Gogarten and Hilario, 2006; Barzel et al., 2011). Briefly, the HEN encoded by an intein DNA segment located at a specific site recognizes a DNA target sequence corresponding to an unoccupied allelic site. The HEN generates a double-stranded chromosomal break within the unoccupied target sequence that is repaired by the host DNA repair machinery using as a template the allele containing the intein gene. This gene conversion results in the replication of the intein gene into a previously unoccupied allele. The occupied allele is no longer a target for the HEN because the target site is eliminated by the insertion. Thus, inteins are mobile elements that occupy unique, specific sites in the genome. When copied into an empty

☆ The nucleotide sequences of the five Dpol and the five RNR genes of have been deposited in GenBank/EMBL/DDBJ database with the accession numbers FR851338–FR851347.

* Corresponding author. Fax: +33 (0) 2 47 42 77 43.
E-mail address: yves.bigot@tours.inra.fr (Y. Bigot).

1055-7903/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2013.01.017

allele, they are retained by the donor allele. Inteins therefore appear to be encoded by mobile genetic elements (MGEs).

To be stably maintained in genomes, inteins are required to be inserted in-frame within a host protein coding sequence and to have highly conserved integration targets to be able to disseminate by horizontal gene transfer among their potential hosts. In agreement with these viability constraints, inteins were previously found predominantly in a restricted subset of about 50 genes, including DNA and RNA polymerases, ribonucleotide reductases (RNRs), H + ATPases, proteases, and metabolic enzymes (Swithers et al., 2009). Moreover, DNA segments coding inteins are known to integrate in the middle of DNA sites of about 24-bp that encode highly conserved amino acid motifs involved in the catalytic activities of the host protein. For example, some of the inteins found in genes coding for the δ DNA polymerase B (DpolB) are inserted in the middle of the most conserved motif involved in polymerase activity, that is, YGD/TDS (so-called pol-c insertion site in InBase; Perler, 2002), and those found in the ribonucleotide reductases are located in the middle of the conserved motif, SNL/CSE (so-called RIR1 insertion site in InBase; Perler, 2002).

Our knowledge about the occurrence of inteins in organisms and viruses is very limited, and to date, they have primarily been identified in bacteria and archaea (see http://blocks.fhcrc.org/~pietro/inteins and http://www.neb.com/neb/inteins.html). Inteins also occur in phage genomes that contain genes orthologous to those of their host, as exemplified in databases for RNRs of the *Vibrio* ICP1phage (Acc. No. ADX89383.1), the *Brochothrix* A9 phage (Acc. No. ADJ53162.1) or the *Bacillus* SPBc2 phage (Acc. No. NP_046713.1). More recently, several inteins have been found in unicellular eukaryotes such as amoebae, fungi and green algae (Goodwin et al., 2006). Similar to what has been found in prokaryotes, inteins are now known to occur in certain viruses of amoeba and green algae, mostly viruses of the *Mimiviridae* (Claverie et al., 2009; Arslan et al., 2011) and *Phycodnaviridae* (Wilson et al., 2009) families that contain DpolB or RNRs genes in their genome (Nagasaki et al., 2005; Ogata et al., 2005; Wilson et al., 2009). This finding suggested that phages and viruses are potential "vectors" of inteins across species and thus are responsible for the limited distribution of inteins among prokaryotes and unicellular eukaryotes (Culley et al., 2009). It is possible that inteins could have been acquired by unicellular eukaryotes and their viruses from bacteria and archaea as a result from the feeding of amoeba, fungi and green algae on these prokaryotes, similar to what has been previously proposed for some bacterial insertion elements (*IS*; Filée et al., 2006).

To date, no inteins have been described in metazoans. A decade ago, one intein was, however, characterized in the RNR of the Chilo iridescent virus (Pietrokovski, 1998; IIV6/CIV (Jakob et al., 2001); family *Iridoviridae*), a virus with a broad host spectrum among invertebrates (Chinchar et al., 2009; Nalçacioglu et al., 2009). Mining databases also revealed the occurrence of inteins in the RNR genes of two other invertebrate iridoviruses, IIV9/WIV (*Wiseana* iridescent virus, Acc. No. AAY24449.1, data recently confirmed by Wong et al., (2011)), and IIV16/CzIV (*Costelytra zealandica* iridescent virus, Acc. No. AAY24450.1). In the present study, our analysis of preliminary versions of five invertebrate iridovirus genomes, specifically, those of two IIV22 isolates, IIV25, IIV30, and IIV31 (Williams and Cory, 1994), identified novel inteins in DpolB and RNRs genes. The three DNA segments coding inteins found in the DpolB gene of IIV22, IIV30 and IIV31 are the first to be characterized in this host gene in metazoan viruses. Their N-terminal end and insertion site within the host DpolB show significant differences that could be specific for this kind of iridovirus intein. The four inteins found in RNRs of both IIV22 isolates, IIV25 and IIV30, have features similar to those of the inteins previously found in IIV9 and IIV16 RNR genes. The relative abundance of inteins in iridoviruses suggests that they may also occur in some genes of metazoan invertebrate species. However, analysis of DNA sequence conservation at level of the regions that encode the YGD/TDS and SNL/CSE motifs in DpolB and RNR genes, respectively, suggests that selective constraints on the sequence of these sites might be responsible for limiting the host range of these two new types of inteins.

2. Materials and methods

2.1. Sources of iridovirus genomic DNA

Iridovirus types IIV22 (Cameron, 1990), IIV22 Aberystwyth (IIV22A; Williams, 1994), IIV25 Aberystwyth and IIV30 strains were kindly supplied by Professor Trevor Williams (Instituto de Ecologia AC, Xalapa, Mexico) and Professor Primitivo Caballero (Universidad Publica de Navarra, Pamplona, Spain). The four strains were amplified by infecting third instar larvae of *Spodoptera frugiperda* (order Lepidoptera, family Noctuidae). Seven days after infection, larvae were frozen at −80 °C. IIV31 was recovered from a natural populations of *Armadillidium vulgare* (Crustacea, Isopoda) infected with this virus-type breeding in ivy beds on the campus of the University California at Riverside (Bigot et al., 2000). The five viruses and their genomic DNA (gDNA) were purified as described by Bigot et al. (2000).

2.2. DNA sequencing and contig assemblies

IIV gDNA libraries were made as described by Dudley et al. (2012). The sequencing of each DNA segment was performed using the Roche-454 procedure. Sequences were assembled into contigs using the Newbler assembler package from Roche.

2.3. Phylogenetic analyses

Sequences used to calculate phylogeny were first determined using BLAST results from databases (Dereeper et al., 2010) and verified from the information available from InBase (Perler, 2002) at http://www.neb.com/neb/inteins.html. Alignments were performed with ClustalW, T-Coffee, Contralign, Mafft, MSAprobes, Muscle, ProbCons and POA (for review, Dessimoz and Gil, 2010). They were then qualified with the evaluation mode of M-Coffee. T-Coffee and Maftt alignments were retained for the DpolB and the RIR1 intein data sets, respectively (Supplementary material and methods, Supplementary material online). Alignments were manually inspected and conserved blocks were revised using SeaviewV4 (Gouy et al., 2010). Phylogenetic trees based on maximum likelihood were calculated with PhyML at http://www.phylogeny.fr/ (Guindon et al., 2010). Parameters used were WAG (substitution matrix), 0 (proportion of invariable sites), 7 in a, and 5 in f (number of relative substitution rate categories) and F (substitution model). The protein substitution model, the proportion of invariable sites, the number of relative substitution, number of rate categories and substitution model for ML trees were selected and evaluated by ProtTEST 3 (Darriba et al., 2011). The best model was chosen on the basis of the Akaike Information Criterion (AIC, Akaike, 1973). Trees were drawn with TreeDyn. Sequence alignments used for calculations are provided as Supplementary data 1 and 2. Complementary phylogenetic analyses are supplied in Supplementary material and methods.

2.4. Analyses of intein insertion sites

Insertion site sequences in a window of 24-bp were extracted from databases using NCBI facilities (Supplementary Tables S2