RESEARCH ARTICLE

INTERNATIONAL MICROBIOLOGY (2008) 11:41-48
DOI: 10.2436/20.1501.01.43 ISSN: 1139-6709 www.im.microbios.org

INTERNATIONAL MICROBIOLOGY

Massive presence of insertion sequences in the genome of SOPE, the primary endosymbiont of the rice weevil *Sitophilus oryzae*

Rosario Gil,[1]* Eugeni Belda,[1] María J. Gosalbes,[1] Luis Delaye,[1,2] Agnès Vallier,[3] Carole Vincent-Monégat,[3] Abdelaziz Heddi,[3] Francisco J. Silva,[1] Andrés Moya,[1] Amparo Latorre[1]

[1]Cavanilles Institute of Biodiversity and Evolutionary Biology and Department of Genetics, University of Valencia, Spain. [2]Faculty of Science, National Autonomous University of Mexico (UNAM), Mexico City, Mexico. [3]Functional Biology, Insects and Interactions (BF2I), IFR41, INRA, INSA-Lyon, Villeurbanne, France

Received 14 December 2007 · Accepted 10 February 2008

Summary. Bacteria that establish an obligate intracellular relationship with eukaryotic hosts undergo an evolutionary genomic reductive process. Recent studies have shown an increase in the number of mobile elements in the first stage of the adaptive process towards intracellular life, although these elements are absent in ancient endosymbionts. Here, the genome of SOPE, the obligate mutualistic endosymbiont of rice weevils, was used as a model to analyze the initial events that occur after symbiotic integration. During the first phases of the SOPE genome project, four different types of insertion sequence (IS) elements, belonging to well-characterized IS families from γ-proteobacteria, were identified. In the present study, these elements, which may represent more than 20% of the complete genome, were completely characterized; their relevance as a source of gene inactivation, chromosomal rearrangements, and as participants in the genome reductive process are discussed herein. [**Int Microbiol** 2008; 11(1): 41-48]

Key words: SOPE (*Sitophilus oryzae* primary endosymbiont) · *Sitophilus oryzae* (rice weevil) · insertion sequences (IS) · endosymbiosis

Introduction

The symbiotic associations of bacteria with eukaryotic cells include a large range of host interactions, from commensalism and mutualism to parasitism [12,21]. In some cases, the relationship is so tightly coupled that the bacteria are literally trapped inside specialized host cells, the bacteriocytes, and are vertically transmitted with each generation from the maternal host to the offspring. Bacterial endosymbioses are particularly widespread among insects, allowing the latter to grow on imbalanced food resources (such as plant sap, cereals, or blood) poor in essential nutrients, which are instead provided by the bacteria [2,10]. The bacterial transition to an obligate intracellular lifestyle triggers a cascade of changes that shape the structure and content of its genome, leading to a reduction in size and an increase in A+T content, among other features.

Intracellular symbionts can be regarded as primary (P)-endosymbionts, essential for host fitness and survival and unable to live outside the host cells, or secondary (S)-symbionts in that the bacteria are facultative and thus also able to survive outside the eukaryotic cell, including growth in lab-

***Corresponding author:** R. Gil
Institut Cavanilles de Biodiversitat i Biologia Evolutiva
Universitat de Valencia. Apartat 22085
46071 Valencia, Spain
Tel. +34-963543824. Fax +34-963543670
E-mail: rosario.gil@uv.es

oratory culture [5,9]. Molecular phylogenetic studies [2] have shown that P-endosymbionts generally have long evolutionary histories with their hosts, beginning with infection of the ancestor host by ancestor bacteria, followed by co-evolution of these partners. In contrast, S-symbionts appear to have established more recent host associations.

Our laboratories have been studying the evolution of endosymbiotic bacterial genomes. These studies have included attempts to determine the factors involved in the establishment of a mutualistic rather than pathogenic relationship between bacteria and their hosts. Our goal is to identify the changes that have occurred during bacterial adaptation to intracellular life and the processes responsible for them. Accordingly, genomes from bacteria with young and old endosymbiotic relationships with their eukaryotic hosts have been sequenced. All P-endosymbionts analyzed so far have genome sizes about eight to ten times smaller than those of their free-living relatives [21]. One of the most extreme cases is *Buchnera aphidicola*, a P-endosymbiont of the aphid *Cinara cedri*. The 422-kb genome was recently sequenced in our laboratory and may well represent a final stage in the endosymbiotic relationship [23]. Nonetheless, to fully understand the molecular and evolutionary events that govern the establishment of symbiosis and to uncover the early molecular processes involved in endosymbiont genome reduction, it is necessary to sequence and characterize the genome of bacteria at the initial stages of endosymbiosis. To this end, we chose the bacterium SOPE, the primary endosymbiont of the rice weevil *Sitophilus oryzae* (Insecta, Coleoptera: Dryophthoridae).

SOPE is a γ-proteobacterium that maintains a typical obligate mutualistic endosymbiosis with its host. The bacteria live inside bacteriocytes organized in an organ called the bacteriome, which surrounds the fore-midgut junction of the insect and occupies the apex of the female ovaries. These bacteria, which have not been cultured outside the host so far, provide at least amino acids and vitamins to the insect, which improve its fertility, development, and flying ability [14].

Symbiosis between SOPE and its host has been estimated to be less that 25 million years old [17], and is thus much younger than any P-endosymbiotic relationship thus far analyzed by genome sequencing of the bacterial component. It has been proposed that the presence of SOPE in cereal weevil lineages is due to its replacement of the ancestral endosymbiont *Candidatus* Nardonella in the family Dryophtoridae to which the rice weevil belongs [17]. Several other features are in accordance with a young relationship. For example, unlike other genomes from insect endosymbionts, the SOPE genome (about 3 Mb, as estimated by pulsed field gel electrophoresis), has not undergone a severe reduction in size [45] and shows a G+C content similar to that of free-living proteobacteria (54%) [13]. In addition, SOPE is closely related to *Sodalis glossinidius* [6,14], an S-symbiont found in the tsetse fly; the genome of the bacterium was recently sequenced [32]. In contrast to SOPE, *S. glossinidius* can be cultured in the laboratory and is found intra- and extracellularly within its host [5]. Therefore, a comparison of the genomes of these two bacteria will allow identification of the genes that are lost at the beginning of symbiogenesis and shed light on the process of bacterial adaptation to an endosymbiotic way of life.

At the beginning of the SOPE sequencing project (in progress in our laboratory), the SOPE genome was found to be highly colonized by several types of mobile elements, mainly those belonging to the group of insertion sequences (IS). IS elements are widespread in free-living bacteria, but they are completely absent in endosymbionts that share an extensive evolutionary history with their hosts [21]. In contrast, IS elements are quite abundant in strains and species that have only recently acquired an obligate intracellular way of life. Moreover, their relative abundance among bacteria that have recently evolved as specialized pathogens, e.g., the enteric bacteria *Shigella* and *Salmonella enterica* Typhi [15,35], is well-documented. Indeed, some studies have indicated that this is also a common trait among bacteria that have recently established mutualistic relationships with their hosts, with subsequent effects on the outcome of the symbiotic process [3,20]. The increase in the frequency of mobile elements is no doubt due to an increase in the replicative transposition of elements that were resident at the onset of symbiosis [8], thus representing a source of chromosomal rearrangements and gene inactivation [20]. Sequencing of the S-symbiont *S. glossinidius* genome showed that about a third of it is composed of inactivated genes at different degrees of disintegration and that it has accumulated many IS elements [32]. This same result was obtained with the genomes of two different *Wolbachia* species, indicating that the abundance of mobile elements is highly dependent on a bacterium's way of life. Thus, mobile elements represent 14% of the genome in the parasitic species *W. pipientis w*Mel (from *Drosophila melanogaster*) [36], but just 5.4% in the obligate endosymbiotic species *W. pipientis w*Bm (from the nematode *Brugia malayi*) [11]. Other facultative insect symbionts, such as *Candidatus* Hamiltonella defensa and *Candidatus* Arsenophonus arthropodicus, also contain mobile elements [8].

In the present work, we characterized the four most abundant IS elements found in the SOPE genome, as well as several of the genes interrupted by these elements. Our results provide new insights into the relevance of these mobile elements in the initial steps of genome evolution in obligate mutualistic symbionts.

Materials and methods

Insect rearing and strain selection. *Sitophilus oryzae* insects were reared on wheat grains at 27.5°C and 70% relative humidity. Rice weevils can be naturally infected with two types of bacteria, SOPE and *Wolbachia*. To obtain SOPE monosymbiotic insects, several *S. oryzae* strains were tested for the presence/absence of *Wolbachia* by PCR and fluorescence in situ hybridization (FISH), using specific *Wolbachia* primers and probes, as described previously [14]. The *S. oryzae* Bouriz (Vallée de l'Azergues, France) strain was found to harbor only SOPE and was therefore used in this work.

SOPE genomic DNA isolation. Bacterial DNA was extracted from *S. oryzae* bacteriomes by dissection of fourth-instar larvae (after larval beheading). The bacteriomes of about 200 larvae were homogenized in 1 ml of isolation buffer A (25 mM KCl, 250 mM sucrose, 35 mM Tris-HCl buffer, pH 7.5), and centrifuged in 5 ml of buffer A for 10 min at 400 ×*g* to eliminate cell nuclei and debris. The supernatant was then centrifuged for 5 min at 10,000 ×*g*, and the pellet was incubated for 1 h at 4°C in 195 μl of a DNase solution (Sigma-Aldrich, Saint Louis, MO, USA; 1 mg/ml) and 5 μl of 400 mM $MgCl_2$ to remove nuclear DNA contaminants. The reaction was stopped with 20 μl of 0.5 M EDTA. The bacterial solution was brought to a volume of 800 μl with buffer A and centrifuged for 10 min at 10,000 ×*g*. The pellet was resuspended in 186 μl of buffer B (1 M NaCl, 0.5% Brij-58, 0.2% deoxycholate, 0.2% *N*-lauroylsarcosine, 10 mM EDTA, 6 mM Tris-HCl buffer, pH 7.61), 10 μl of RNase A (Sigma-Aldrich; 10 mg/ml), and 4 μl of lysozyme (Roche, Indianapolis, IN, USA; 50 mg/ml), and incubated for 3 h at 37°C. Proteinase K (2 μl, 20 mg/ml, Roche) was added and the incubation was continued overnight at 37°C. Genomic DNA was purified by a standard phenol/chloroform protocol [27].

Library construction and sequencing. Shotgun sequence libraries were prepared as described [33]. SOPE genomic DNA was sonicated, and random fragments of ~1.5 kb were cloned into the TOPO XL plasmid (Invitrogen, Carlsbad, CA, USA). Genomic DNA was digested with *Bam*HI, and 4- to 6-kb fragments were selected for cloning in pUC18. The recombinant plasmids were purified and sequenced using universal primers. Dye terminator cycle sequence analysis was carried out using sequencing kits (Applied Biosystems, Foster City, CA, USA) at the sequencing facility of the University of Valencia. All trace data were analyzed using the Staden Package Software Program [31] for trimming of vector sequences, data assembly, editing, and finishing processes.

Identification and analysis of IS elements. The quality of the shotgun libraries was assessed by the sequencing and analysis of about 500 clones using BLASTX searches [http://www.ncbi.nlm.nih.gov/blast/Blast.cgi] [1]. The results revealed that nearly a third of the obtained sequences matched those of transposases found in typical IS elements from γ-proteobacteria. Therefore, to characterize the highly abundant IS elements, a local database was created containing all such sequences. After alignments and BLASTX searches were done the results were compared with the *ISfinder* dataset [http://www-is.biotoul.fr/] [28]. Inverted and direct repeats were identified with the programs *Palindrome* and *Etandem* included in the EMBOSS package [26]. The sizes of the direct repeats generated after the insertion of each IS element type were visually inspected by comparison of the flanking nucleotides of complete IS element copies. Finally, the protein-coding genes were annotated. The consensus sequence of each type of IS element was used to design sets of sequencing primers (Table 1), which enabled the complete sequencing of each copy plus a few hundred nucleotides of their flanking regions.

None of the identified IS presented internal *Bam*HI sites. To identify clones from the *Bam*HI library harboring internal IS elements, dot-blot hybridization analysis was carried out using as probes PCR products amplified with internal primers for ISsope*1* and ISsope*2* (Table 1). These probes were labeled with digoxigenin (Roche). All procedures were done according to the manufacturer's instructions. The identified clones containing IS elements were sequenced using the corresponding specific sets of primers. Additionally, when an ISsope*3* or ISsope*4* element was reached by direct sequencing of the two ends of a clone, the corresponding IS was sequenced with a set of primers specific for this element (Table 1). Several clones harbored two or more copies of the same element, which was detected by the mixture of sequences once the electropherogram readings had arrived at the flanking sequences. To resolve this sequencing ambiguity, those clones were PCR-amplified using selected primers recognized only by each IS copy and the resulting products were sequenced.

The IS sequences obtained were aligned with ClustalW of MEGA3 software [18]. The p-distances (number of differences/alignment lengths) between copies of the same type of element were estimated with the pairwise deletion option in the MEGA3 program. Due to the variable number of direct repeats in the left inverted repeat (IRL) of ISsope*3* (Table 2), the first 80 nucleotides of the alignment were removed for distance estimation. The part of the nucleotide alignments encoding the transposase genes in ISsope*1*, ISsope*2*, and ISsope*4,* and the two coding genes in ISsope*3,* were isolated and translated. Genes with stop codons or frameshifts were considered defective. In the absence of any further information, non-synonymous substitutions that changed the encoded amino acids as well as small insertion multiples of three nucleotides were considered to yield functional products. Small increases in gene length due to the loss of the consensus stop codon and the use of a downstream-proximal one were also considered functional.

Identification of ORFs interrupted by IS elements in the contigs from the SOPE sequencing genome project. A contig can be defined as a set of overlapping DNA segments derived from a single genetic source, as well as the consensus DNA sequence reconstructed from such segments. Since contigs represent fragmentary data and as such are not likely to represent a complete gene, a heuristic search was applied in order to identify putative coding genes interrupted by IS elements. Accordingly, the positions of putative coding regions were mapped through BLASTX searches of contigs using a database of protein sequences from 260 selected complete bacterial genomes. Only the best-hits were selected for that purpose. The locations of the detected inverted repeats from the different identified ISs in the contigs were then mapped using Perl script (available upon request). Finally, the compatibility between sequences having the best-hit length and the coordinates of the IS in the contig was analyzed.

Results and Discussion

In one of the first steps in the SOPE genome sequencing project, approximately 500 clones were sequenced and analyzed by BLASTX. The results confirmed the quality of the obtained genomic DNA and libraries and showed that nearly a third of the obtained sequences matched those of transposases found in typical IS elements from γ-proteobacteria. Bacterial IS are characterized by the presence of at least one gene containing a transposase, which is responsible for IS transposition, flanked by two inverted short sequences. Up to four different IS types were identified and found to be either complete or truncated, with three of them being widely present (Fig. 1, Table 2). All of the elements showed clear homology with known γ-proteobacterial IS families. The two most abundant IS elements, ISsope*1* (IS*5* family) and ISsope*2*

Table 1. List of oligonucleotides used as probes or specific primers to sequence the complete insertion sequences (IS) elements and their flanking regions

Target IS	Oligonucleotide	Sequence (5′-3′)	Start position	Application	PCR product size (bp)
ISsope*1*	ISt1_216F	CCGCTTCACTACACCGATA	216	IS sequencing	
ISsope*1*	ISt1_479R	GACTTTCCATTCGCCTTC	479	IS sequencing	
ISsope*1*	ISt1_650F	TGACAGTGCTTACGATAC	650	IS sequencing	
ISsope*1*	IS1_18f	CGAACTTTTAGGTGACTGG	18	PCR amplification of hybridization probe	387
ISsope*1*	IS1_404r	GGCGTTTTTATGCTGATG	404	PCR amplification of hybridization probe	
ISsope*2*	ISt2_237R	TTTAGGCTGATTTTTATCGT	237	IS sequencing	
ISsope*2*	ISt2_643F	ATCGAAGGCCAGAAAGAG	643	IS sequencing	
ISsope*2*	ISt2_809R	GGATACACCGYGTTAATAG	809	IS sequencing	
ISsope*2*	ISt2_1144F	GACGACRCAGTRAAAAAG	1144	IS sequencing	
ISsope*2*	ISt2_476F	CGCTGGTCTCAAAGGTYA	476	IS sequencing	
ISsope*2*	IS2_92f	AGGCTCTGGCTAACGAAC	92	PCR amplification of hybridization probe	478
ISsope*2*	IS2_569r	TCAAGATAAACAATGGGATAGA	569	PCR amplification of hybridization probe	
ISsope*3*	ISt3_199R	CCGTTCTCGCTTTCTTYTT	199	IS sequencing	
ISsope*3*	ISt3_112F	ATCGCTCATCTTCTGTTCC	112	IS sequencing	
ISsope*3*	ISt3_924F	GAGAAAACCGAAAGACAAGG	924	IS sequencing	
ISsope*3*	ISt3_1024R	CGGCCAGCGAGTAGAACC	1024	IS sequencing	
ISsope*3*	ISt3_1667F	GCGTATCAATGCTGGTTC	1667	IS sequencing	
ISsope*3*	ISt3_1831R	CCCAGCTCCCCATAACTC	1831	IS sequencing	
ISsope*3*	ISt3_2107F	AATATAGCGTGCGTTACTGG	2107	IS sequencing	
ISsope*3*	ISt3_527R	GAGTAACCGAGGGCATCAC	527	IS sequencing	
ISsope*3*	ISt3_1485R	TTTGCTCTTTTGGATGGA	1485	IS sequencing	
ISsope*4*	pIS4_302R	YACATCAGCCTCAACCAG	302	IS sequencing	
ISsope*4*	pIS4_212F	CGGTAAATCCTGCTCCAT	212	IS sequencing	
ISsope*4*	pIS4_1337R	TACAACATCACCAGCAAAAA	1337	IS sequencing	
ISsope*4*	pIS4_1676R	GCATTTTTACTTTACTTATTC	1676	IS sequencing	
ISsope*4*	pIS4_1541F	TGAAGCAGTAAGATAAATGG	1541	IS sequencing	

(IS*256* family), consisted only of a transposase gene and almost perfect inverted repeats (IRs). The presence of these highly abundant elements in both SOPE and SZPE was recently reported [28]. However, in contrast to those findings, our data indicated that ISsope*1* (which probably corresponds with IS*903* in the previous study) is even more abundant than ISsope*2*. A third element, ISsope*4* (ISL*3* family), with almost perfect IRs and a transposase gene, also contains a putative ORF of 333 nucleotides in the complementary strand that gave no homology with any sequence in the databases. A member of this family (IS*1096*) also has an additional ORF (*tnp*R), but it is located in the same strand and is related to ORFs from *Agrobacterium rhizogenes* and *Rhizobium* plasmids [18]. The ORF detected in ISsope*4* might be the remnants of an earlier, not yet described gene encoding a regulatory protein. Finally, ISsope*3* (IS*21* family) contains two ORFs in the same coding strand, separated by 11 nucleotides between the stop and start codons. IS*21* family members exhibit two consecutive ORFs: a long upstream frame (*ist*A) and a shorter downstream frame (*ist*B*)*. Both IstA and IstB

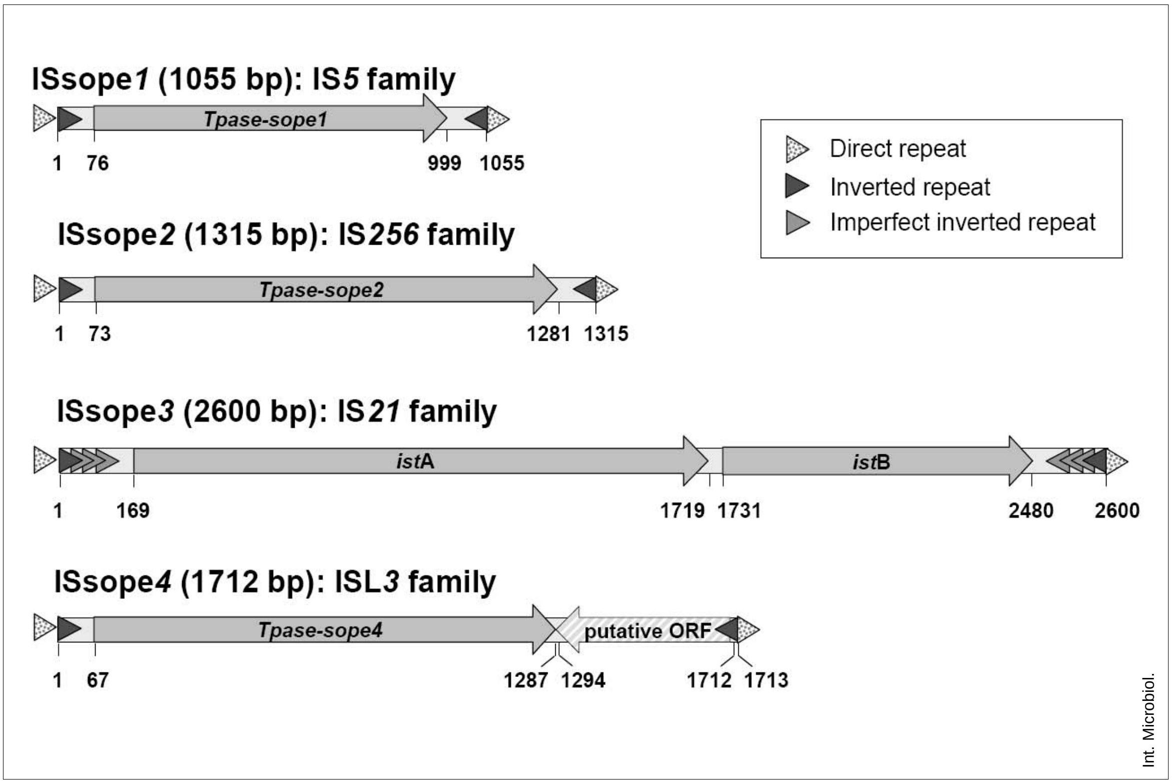

Fig. 1. The four most abundant types of insertion elements found in SOPE. Each one belongs to a well-characterized IS family from γ-proteobacteria. The size of the identified ORFs, as well as the inverted and direct repeats flanking the IS are indicated.

are required for integration [25]. Unlike the other three elements, ISsope*3* is flanked by a complex IR, which includes a variable number of direct repeats. Several members of the IS*21* family also carry multiple repeat sequences that surround part of the terminal IR. It has been postulated that these sequences represent transposase binding sites [18]. In many cases, the short direct repeats generated by duplication of the target genome sequence during the transposition event were identified. The number of nucleotides duplicated as a result of transposition of the different elements agrees with the number identified in the different IS families.

It was also observed that none of the IS contained an internal *Bam*HI site. Therefore, in order to get more information about these elements, a library of long *Bam*HI inserts (4 to 6 kb) was prepared. These were detected based on the presence of ISs, using direct sequencing of both ends and hybridization with probes specific for the different elements. The complete IS sequence together with the flanking sequences was determined as described in Materials and methods. Several complete or almost complete copies of the characterized IS elements, normal or defective, were then compared. All of the copies of each IS type were very similar to each other, with average values of 0.011–0.018 differences/total length (Table 2). For ISsope*1* and ISsope*2*, some of the pair-wise distances were 0, indicating that either the transposition events occurred very recently or that there has been gene conversion in the SOPE genome. These identical copies were flanked by different nucleotide sequences, corroborating that they did not arise through readings of the same copy. For each type of element, a consensus sequence representing the most frequent nucleotide for each position was obtained (accession nos. AM921789–921792). It should be noted that, despite the fact that the number of differences was very small, no copy of IS identical to the consensus sequence was found for any of the four elements.

A relevant percentage of the copies of IS elements contained defective transposase genes (24, 34, and 50% for ISsope*1*, ISsope*2*, and ISsope*3*, respectively). Defective transposases present different types of mutations, including small and large indels and point mutations causing premature stop codons. Since bacterial IS transposition is often restricted to *cis* activity [22], it is likely that those elements with non-functional transposases are inactive and are therefore subject to the gradual erosion postulated to take place in all non-essential regions of endosymbiont genomes, according to the genome reduction model proposed by Silva et al. [29].

Table 2. Structural and evolutionary characteristics of the four most abundant IS elements in the SOPE genome. Analyzed sequences are complete IS elements, except for those elements in which a maximum of 50 nucleotides at the 3′ end were not identified

	ISsope*1*	ISsope*2*	ISsope*3*	ISsope*4*
IS family	IS*5*	IS*256*	IS*21*	ISL*3*
Structural features				
Consensus sequence length (bp)	1055	1315	2600[b]	1712
G+C (%)	51.9	53.7	54.5	48.9
Number of open reading frames (ORFs)	1	1	2	2
Inverted repeats (bp)	18	30	33 (left), 34 (right) + (23 x 3 or 4)[c]	24
Direct repeats derived from transposition (bp)	9	8	5	8
Number of complete sequences analyzed[a]	54	35	8	3
IS elements with intact ORFs (%)	41 (76)	23 (66)	4 (50)	3 (100)
IS with ORF1 (transposase) with loss-of-function mutations (%)	13 (24)	12 (34)	2 (50)	0 (0)
IS with ORF2 with loss-of-function mutations (%)			1 (25)	
p-distance (number of differences/total length)				
Minimum	0	0	0.008	0.014
Maximum	0.024	0.026	0.014	0.023
Mean	0.013	0.011	0.012	0.018
Inverted repeats detected in the SOPE sequencing project	410	193	78	19

[a] Sequences were complete or almost complete.
[b] The actual length depends on the number of direct repeats in the remaining IR, with 2600 bp the most frequent length for this IS element.
[c] Terminal IR sizes are 34 (left) and 33 (right). Repeat units vary in size (from 21 to 25 bp); total left and right IR extend up to 126 and 112 bp, respectively.

It is well-known that although IS elements, as well as other transposable elements, are widespread in bacteria, their capacity for transposition within the genome is generally tightly regulated [22], such that most bacterial genomes contain only a few copies of a limited number of IS types [34]. However, recent data have shown that ISs are relatively abundant in bacteria that have recently evolved host-restricted lifestyles, as is the case for *S. glossinidius*, the S-symbiont of tsetse flies, and the two grain weevil P-endosymbionts, SZPE (from *S. zeamais*) and SOPE (from *S. oryzae*) [24,32]. Phylogenetic studies showed that these grain weevil symbionts replaced *Candidatus* Nardonella, the ancient symbiont present in other members of the Dryophtoridae weevils [17], and form a sister clade of *S. glossinidius*, that still keeps its ability to return to free-living conditions. Therefore, *S. glossinidius* and SOPE provide a useful model of an initial state along the evolutionary path from free-living bacteria to ancient endosymbionts.

Comparative genome analyses have provided evidence that genome shrinkage in endosymbionts consists of two separate stages. In the first, massive gene loss most likely occurs soon after the establishment of obligate symbiosis, probably by means of large deletion events that caused the elimination of series of contiguous genes [19]. In the second, genome reduction proceeds through a process of gradual pseudogenization and gene loss scattered throughout the genome [29]. After the establishment of symbiosis, the high abundance of very similar mobile elements in direct orientation that can serve as the substrate for unequal recombination would lead to a loss of the region between two elements, thus promoting genome size reduction during the above-described first stage. Furthermore, the proliferation of such elements would also be involved in the inactivation of genes that have been rendered unnecessary in the newly protected and nutrient-rich environment of the host. In the present study, any ORF interrupted by the presence of one of the four classes of IS was identified (see Materials and methods). At least one inverted repeat from one of the four IS reported here was found to be present in 476 contigs from the SOPE sequencing project (in progress), including 215 contigs in which a putative protein-coding gene was interrupted by an IS element. The functional classification of the interrupted genes is shown in Fig. 2. Most

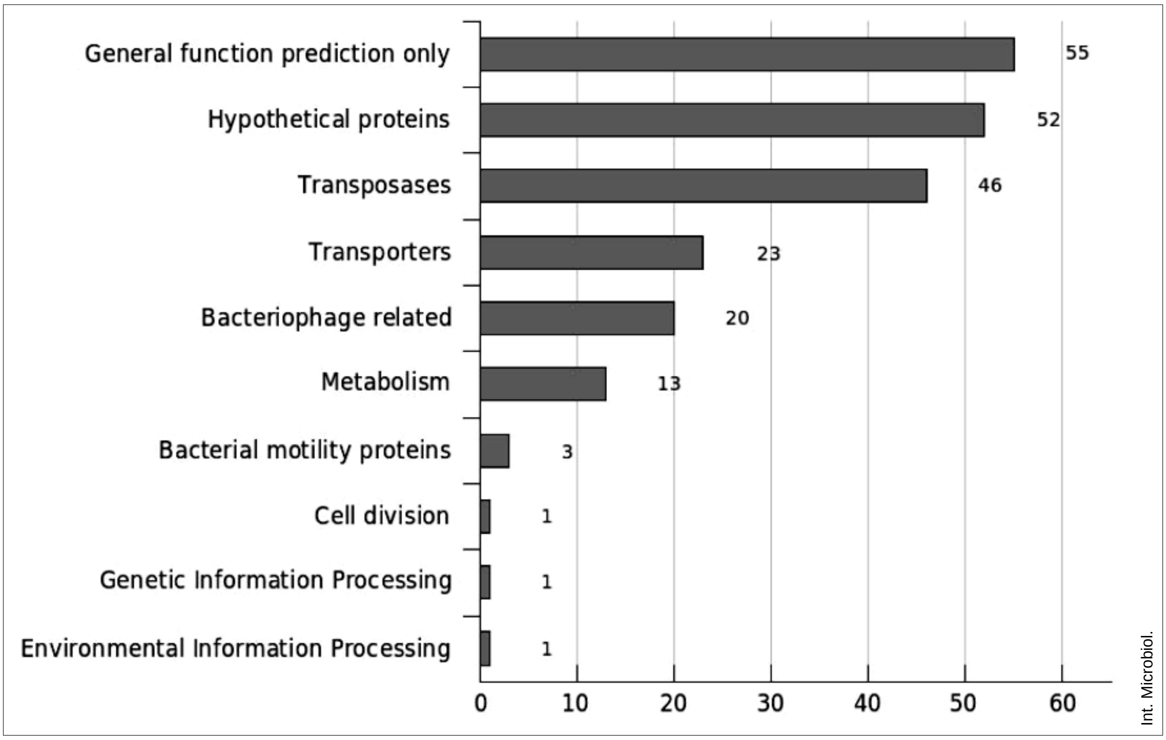

Fig. 2. Functional classification of the identified ORFs interrupted by IS elements analyzed in this work.

encoded truncated proteins belong to the categories "hypothetical proteins" or "general function prediction only". There were also many transposases as well as several genes involved in transport. Multiple copies of a gene encoding a dam methylase were found near or interrupted by IS elements in SOPE (data not shown). Most of these copies carried inactivating mutations, suggesting that massive proliferation of IS elements is, at least in part, due to a defective control mechanism regulating their transposition [22]. In addition, the proliferation of these elements is probably favored by the small effective population sizes of obligate symbionts, which decreases the effectiveness of natural selection and thus allows the accumulation of slightly deleterious IS insertions [24].

Our results are in concordance with the hypothesis of Wagner et al. [34], who proposed that, after an IS enters the genome, its copy number expands rapidly through transposition. Consequently, there is a low degree of diversity among the different copies of an IS present in a given genome, and that of SOPE is no exception. Eventually, due to the deleterious effect of their accumulation, IS elements become extinct through natural selection. In the case of bacteria that have recently established a symbiotic relationship with a eukaryotic host, many of their genes become non-essential, allowing the accumulation of ISs without a detrimental effect on the fitness of the association. Yet, these elements ultimately become extinct as well, since there are no traces of ISs in endosymbiotic bacteria with long relationships to their hosts and massive genome reduction [30]. In free-living bacteria, IS elements may be reintroduced through horizontal gene transfer (HGT). This process cannot take place in bacteria that have acquired an obligate intracellular way of life, since HGT is no longer possible due to bacterial isolation as well as the loss of genes involved in DNA uptake [30] and DNA recombination, in early stages of the endosymbiotic association [7].

The characterization and analysis of the four most frequent IS types found in SOPE has allowed us to propose an evolutionary scenario for the first stage of symbiotic integration of mutualistic bacteria. The complete sequence of the SOPE genome, and its comparison with those from strict pathogenic bacteria and ancient mutualistic endosymbionts will provide new clues into the genes that are among the first to be lost during the reductive process, the possible role played by IS elements in their disappearance, and their relevance in the outcome of the symbiotic process.

Acknowledgements. Financial support was provided by projects BFU2006-06003 from the Spanish Ministry of Education and Science (MEC), to AL; the French ANR (EndoSymArt) to AH; and GV/2007/050 from Generalitat Valenciana, Spain, to RG. RG is a recipient of a grant from the Ramón y Cajal Program of the MEC, Spain. EB is a recipient of a fellowship from the MEC, Spain.

References

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402
2. Baumann P (2005) Biology bacteriocyte-associated endosymbionts of plant sap-sucking insects. Annu Rev Microbiol 59:155-189
3. Bordenstein SR, Reznikoff WS (2005) Mobile DNA in obligate intracellular bacteria. Nat Rev Microbiol 3:688-699
4. Charles H, Condemine G, Nardon C, Nardon P (1997) Genome size characterization of the endocellular symbiotic bacteria of the weevil *Sitophius oryzae*, using pulse field gel electrophoresis. Insect Biochem Molec Biol 27:345-350
5. Dale C, Maudlin I (1999) *Sodalis* gen. nov. and *Sodalis glossinidius* sp. nov., a microaerophilic secondary endosymbiont of the tsetse fly *Glossina morsitans morsitans*. Int J Syst Bacteriol 49:267-275
6. Dale C, Plague GR, Wang B, Ochman H, Moran NA (2002) Type III secretion systems and the evolution of mutualistic endosymbiosis. Proc Natl Acad Sci USA 99:12397-12402
7. Dale C, Wang B, Moran NA, Ochman H (2003) Loss of DNA recombinational repair enzymes in the initial stages of genome degeneration. Mol Biol Evol 20:1188-1194
8. Dale C, Moran NA (2006) Molecular interactions between bacterial symbionts and their hosts. Cell 126:453-465
9. Darby AC, Chandler SM, Welburn SC, Douglas AE (2005) Aphid-symbiotic bacteria cultured in insect cell lines. Appl Environ Microbiol 71:4833-4839
10. Douglas AE (1998) Nutritional interactions in insect-microbial symbioses: aphids and their symbiotic bacteria *Buchnera*. Annu Rev Entomol 43:17-37
11. Foster J, Ganatra M, Kamal I, et al. (2005) The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. PLoS Biol 3:e121
12. Guerrero R, Berlanga M (2006) Life's unity and flexibility: the ecological link. Int Microbiol 9:225-235
13. Heddi A, Charles H, Khatchadourian C, Bonnot G, Nardon P (1998) Molecular characterization of the principal symbiotic bacteria of the weevil *Sitophilus oryzae*: a peculiar G + C content of an endocytobiotic DNA. J Mol Evol 47:52-61
14. Heddi A, Grenier AM, Khatchadourian C, Charles H, Nardon P (1999) Four intracellular genomes direct weevil biology: nuclear, mitochondrial, principal endosymbiont, and *Wolbachia*. Proc Natl Acad Sci USA 96:6814-6819
15. Jin Q, Yuan Z, Xu J, et al. (2002) Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. Nucleic Acids Res 30:4432-4441
16. Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform 5:150-163
17. Lefevre C, Charles H, Vallier A, Delobel B, Farrell B, Heddi A (2004) Endosymbiont Phylogenesis in the Dryophthoridae weevils: Evidence for bacterial replacement. Mol Biol Evol 21:965-973
18. Mahillon J, Chandler M (1998) Insertion sequences. Microbiol Mol Biol Rev 62:725-774
19. Moran NA, Mira A (2001) The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. Genome Biol 2:research0054.1–0054.12
20. Moran NA, Plague GR (2004) Genomic changes following host restriction in bacteria. Curr Opin Genet Dev 14:627-633
21. Moya A, Peretó J, Gil R, Latorre A (2008) Learning how to live together: genomic insights into prokaryote-animal symbioses. Nat Rev Genet 9:218-229
22. Nagy Z, Chandler M (2004) Regulation of transposition in bacteria. Res Microbiol 155:387-398
23. Pérez-Brocal V, Gil R, Ramos S, et al. (2006) A small microbial genome: The end of a long symbiotic relationship? Science 314:312-313
24. Plague GR, Dunbar HE, Tran PL, Moran NA (2007) Extensive proliferation of transposable elements in heritable bacterial symbionts. J Bacteriol doi:10.1128/JB.01082-07
25. Reimmann C, Haas D (1990) The *istA* gene of insertion sequence IS*21* is essential for cleavage at the inner 3′ ends of tandemly repeated IS*21* elements in vitro. EMBO J 9:4055-4063
26. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet 16:276-277
27. Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
28. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M (2006) ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res 34:D32-36
29. Silva FJ, Latorre A, Moya A (2001) Genome size reduction through multiple events of gene disintegration in *Buchnera* APS. Trends Genet 17:615-618
30. Silva FJ, Latorre A, Moya A (2003) Why are the genomes of endosymbiotic bacteria so stable? Trends Genet 19:176-180
31. Staden R, Beal KF, Bonfield JK (2000) The Staden package, 1998. Methods Mol Biol 132:115-130
32. Toh H, Weiss BL, Perkin SA, Yamashita A, Oshima K, Hattori M, Aksoy S (2006) Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. Genome Res 16:149-156
33. van Ham RCHJ, Kamerbeek J, Palacios C, et al. (2003) Reductive genome evolution in *Buchnera aphidicola*. Proc Natl Acad Sci USA 100:581-586
34. Wagner A, Lewis C, Bichsel M (2007) A survey of bacterial insertion sequences using IScan. Nucleic Acids Res 35:5284-5293
35. Wei J, Goldberg MB, Burland V, et al. (2003) Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. Infect Immun 71:2775-2786
36. Wu M, Sun LV, Vamathevan J, et al. (2004) Phylogenomics of the reproductive parasite *Wolbachia pipientis w*Mel: a streamlined genome overrun by mobile genetic elements. PLoS Biol 2:E69