The Association of Group IIB Intron with Integrons in Hypersaline Environments

Sarah Sonbol
The American University in Cairo
Rania Siam (rsiam@aucegypt.edu)
The American University in Cairo https://orcid.org/0000-0002-2879-6368

Research

Keywords: Group II introns, integrons, CALINs, IS200/605, hypersaline, metagenomics, mobile genetic elements

Posted Date: September 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-69915/v1

Abstract

Background: Group II introns are mobile genetic elements used as efficient gene targeting tools. They function as both ribozymes and retroelements. Group IIC introns are the only class reported so far to be associated with integrons. In order to identify group II introns linked with integrons and CALINS (cluster of *att*C sites lacking a neighboring integron integrase) within halophiles, we mined for integrons in 28 assembled metagenomes from hypersaline environments and publically available 104 halophilic genomes using Integron Finder followed by blast search for group II intron reverse transcriptases (RT)s.

Results: We report the presence of different group II introns associated with integrons and integron-related sequences denoted by UHB.I1, UHB.I2, H.ha.1 and H.ha.F2. The first two were identified within putative integrons in the metagenome of Tanatar-5 hypersaline soda lake, belonging to IIC and IIB intron classes, respectively. Truncated introns H.ha.F1 and H.ha.F2 were also detected in a CALIN within the extreme halophile *Halorhodospira halochloris*, both belonging to group IIB introns. The intron-encoded proteins (IEP)s identified within group IIB introns belonged to different classes: CL1 class in UHB.I2 and bacterial class E in H.ha.Fa1 and H.ha.F2. A newly identified insertion sequence (IS*Hahl1*) of IS*200/605* superfamily was also identified adjacent to *H. halochloris* CALIN. Finally, an abundance of toxin-antitoxin (TA) systems was observed within the identified integrons.

Conclusion: So far, this is the first investigation of group II introns within integrons in halophilic genomes and metagenomes from hypersaline environments. We report the presence of group IIB introns associated with integrons or CALINs. This study provides the basis for understanding the role of group IIB introns in the evolution of halophiles and their potential biotechnological role.

Background

Group II introns are mobile genetic elements (MGE)s with catalytic RNA (ribozyme) and retroelements properties [1] [2], linked to non-Long terminal repeat elements [3] [4]. They are found in mitochondrial and chloroplast genomes of lower eukaryotes and plants, and in known bacterial and archaeal genomes [1] [3]. The transcribed ribozyme catalyzes the excision of the intron and its integration into new locations with the aid of an intron-encoded protein (IEP) [3]. Although the RNA sequence of the ribozyme is poorly conserved [5], it can be classified into 3 major groups (IIA, IIB and IIC) [6]. Group II introns classification is based on its conserved secondary and tertiary structure where it forms 6 double helical domains (DI-DVI) radiating from a central wheel [3] [6] (Fig. 1A). DI and DV form the minimal catalytic core of the ribozyme, while DIV encodes the intron open reading frame (ORF) [6]. Catalysis is promoted by the binding of Mg^{2+} ions to AGC triad [7] (CGC in case of group IIC introns [8]) and to an AY bulge, located in DV [7] (Fig. 1A). Amongst the 6 double helical domains, only DV and DVI are highly conserved [5].

Additionally, group II introns can be classified into subgroups, mitochondrial-like (ML), chloroplast-like class I (CL1), chloroplast-like class II (CL2) and bacterial classes A-E, based on their IEP [9]. Bacterial group II introns contain all previously mentioned subgroups, whereas organelles contain only CL and ML

subgroups [10]. The IEP can function as a reverse transcriptase (RT 0–7 subdomains), a maturase (X domain) which binds to the intron RNA to stabilize the secondary structure and assist RNA splicing, and a DNA endonuclease (En domain) [4] [6] [9]. A "YADD" motif necessary for the reverse transcription activity is highly conserved in all bacterial IEPs within RT5 domain [4] [7] (Fig. 1B). Each IEP subgroup can be associated with one RNA subclasses including mitochondrial (IIA1), chloroplast-like class I (IIB1), chloroplast-like class II (IIB2), bacterial class A (IIA/B), bacterial class B (IIB-like), bacterial class C (IIC), bacterial class D (IIB-like) and bacterial class E (IIA/B) [6]. Most bacterial IEPs are found within MGEs such as plasmids or insertion sequences (IS) [4].

Mobilization of group II introns occurs through an RNA intermediate leading to their duplication [11]. The ribozyme in its conserved secondary structure can catalyze its own splicing (excision) from a precursor transcript [12]. Intron splicing usually occurs via 2 sequential transesterification steps [6] starting with a nucleophilic attack of the hydroxyl group in a DVI conserved bulged adenosine (branching pathway) and ending with the formation of an intron lariat and the ligation of the 5' and 3' exons [1] [13]. A less efficient splicing mechanism may occur by water hydrolysis, without the aid of the bulged "A," resulting in a linear excised intron rather than a lariat [13]. The hydrolysis pathway seems to be more common in group IIC introns [12]. The excised intron transcript (RNA) remains associated with the IEP forming a ribonucleoparticle (RNP), which is then inserted (reverse splicing) into either an intronless allele (retrohoming) or to a non-cognate site (retrotransposition or ectopic transposition) with a lower frequency [1]. Reverse splicing into dsDNA requires cleavage of the sense strand, where the intron transcript gets inserted, followed by a cleavage in the antisense strand catalyzed by the En domain of IEP. En-independent retrohoming is connected to DNA replication since single stranded DNA (ssDNA) stretches are formed eliminating the need for a second strand cleavage [1]. Yet, reverse splicing into ds or ssDNA independent of DNA replication can also occur but less frequently [1]. Various studies have shown that putative intron boundaries have a consensus sequence of "GUGYG" at the 5' end and "AXX(X)XRAY" at the 3' end, including the bulged "A" in DVI [8]. To be inserted, the IEP recognizes specific nucleotides in the exons flanking the target site, followed by base pairing between short sequences in the DI loop of the intron RNA (Exon Binding Sites EBS) and sequences in the target site (Intron Binding Sites IBS) [1] [14]. In group II A and B, 5' exon is recognized mainly by 2 base pairing interactions; IBS1-EBS1 (6 bp) and IBS2-EBS2 (6 bp) [12]. In case of 3' exon, its first 1–3 nucleotides (δ') pair with (δ) position upstream of EBS1 in group IIA introns, while in group IIB, the first nucleotide of 3'exon (IBS3) pairs with (EBS3) position in DI double helix of the intron [1]. On the other hand, group IIC introns exhibit some variations in their target site recognition; both IBS1-EBS1 and IBS3-EBS3 interactions can be identified. However, pairing in IBS1-EBS1 is 3–4 bp rather than 6 bp. Up to this point there's no evidence for a IBS2-EBS2 interaction. It was identified that a stem-loop of a Rho-independent terminator or other inverted repeat structure such as an *att*C site is located upstream of IBS1 [1] [9].

*att*C sites are recombination sites usually found at the 3' end of an integron gene cassette, which can be recognized by the integron integrase (IntI), leading to integration or excision of integron gene cassettes [15]. An *att*C site is composed of 4 successive binding sites denoted by R", L", L' and R'. R" and R' are the only conserved domains with the consensus of 5'-RYYYACC-3' and 5'-GTTRRRY-3', respectively [15] [16].

The recombination reaction only involves the *att*C bottom strand (bs) which forms a stem loop structure, where R" and L" pair with R' and L' forming the R and L boxes, respectively [17] (Fig. 1C). Group IIC-*att*C introns form a specific lineage of group IIC introns, this was found to be inserted directly after or into the stem-loop motif of the *att*C site bs, in an opposite orientation to the gene cassettes transcription [9] [18]. Group IIC-*att*C introns can also integrate into *att*C sites within clusters of *att*C sites lacking a neighboring integron-integrase [18] (CALINs) [19]. The majority of these introns were inserted into a consensus sequence of TTGT/T (IBS1/IBS3) within an *att*C site [9] [18] (Fig. 1C). Moreover, despite *att*C sites preference, these introns were found to retain their ability to target other putative transcriptional terminators. This led to the suggestion that group IIC-*att*C introns might be involved in integron gene cassette formation by separately targeting an isolated *att*C site and a transcriptional terminator of any gene, followed by joining this *att*C site to that gene by homologous recombination [9]. Thus, presence of Group IIC-*att*C introns within gene cassette arrays may represent an intermediate step in the formation of some gene cassettes [9].

Members of group IIA and IIB introns have been successfully utilized as gene targeting vectors (targetrons) with high integration efficiency and target specificity [20]. On the other hand, group IIC introns have never been used in such applications, as their reverse splicing mechanism is not fully understood [20]. Furthermore, IEPs have a high potential to be used as RTs in different biotechnological applications that involve cDNA synthesis such as qRT-PCR and RNA sequencing (RNA-seq). Their high fidelity and lack of RNase H activity enables their reuse of RNA templates, making them superior to commercially available RTs [20].

In this study, we investigated group II introns associated with integrons and CALINs in 28 previously assembled metagenomes (1,236,831,758 nucleotides and 658,054 contigs) from different hypersaline environments and all publically available halophilic genomes (104 genomes). We identified -for the first time- group II introns belonging to different classes within integron gene cassette arrays in the metagenome from the hypersaline Tanatar-5 Soda lake, Russia (IIB/CL1 and IIC-*att*C) and within the genome of the extreme halophile *Halorhodospira halochloris* DSM 1059 (IIB/E). Tanatar-5 soda lake is an alkaline hypersaline lake with a pH of 9.9 and a salinity of 170 gl^{-1} [21] with a highly active microbial sulphide cycle [22]. *H. halochloris* is an obligate anaerobic phototroph that inhabits environments of highly saline, and alkaline conditions. The optimal growth conditions of *H. halochloris* requires the presence of sulphide, a pH of 8.1–9.1 [23] and salt concentration of 140–270 gl^{-1} [23]. Furthermore, we have identified a new IS element (IS*Hahl1*) of IS*200/605* superfamily adjacent to the CALIN sequence in *H. halochloris*, and submitted the sequence to the ISfinder database [24]. We investigate putative links between such mobile genetic elements, in the halophile genome and metagenome from a hypersaline environment, and whether an essential synchronized mobilization events occur enabling the adaptation of halophiles in salty environments.

Results

Different Intron encoded Protein (IEP) classes associated with hypersaline integrons and CALINS

We mined 658,054 contigs (1,236,831,758 bp) from 28 hypersaline aquatic metagenomes for integrons and identified CALINs, rather than full integrons, in most sites (unpublished results). Annotation of the identified gene cassettes revealed the presence of two-group II intron RT/maturases in 2 different contigs (LFIK01005867 and LFIK01005957) from Tanatar-5 hypersaline Soda Lake (TSL) in Kulunda steppe in Siberia, Russia. Here we refer to them as TSL1 and TSL2, respectively. The identified group II introns in TSL1 and TSL2 were previously referred to as uncultured halophilic bacterium introns 1 and 2 (UHB.I1 and UHB.I2), respectively. On the other hand, no group II RTs were found within integrons or CALINS of the examined 1,444,498 contigs (1,750,281,271 bp) from the 22 marine or 7 freshwater previously assembled metagenomes.

As TSL1 and TSL2 contigs, with group II introns, were identified from a hypersaline lake, it is expected that they belong to halophilic microorganisms. Thus, we examined publically available complete and partial 104 halophilic genomes to get a clearer picture of the group II introns associated with integrons in halophiles. Only 2 group II intron RT/maturases, in the same CALIN, in the genome of the extreme alkaliphilic and halophilic purple sulfur gammaproteobacterium *Halorhodospira halochloris* DSM 1059 [23] were detected. Apart from the identified CALIN, only one other group II intron RT existed was detected in *H. halochloris* (previously reported in NCBI nr database with the accession number WP_096410353.1). Fragmented introns identified within *H. halochloris* CALIN were denoted by H.ha.F1 and H.ha.F2.

To assign UHB.I1, UHBI2, H.ha.F1 and H.ha.F2 to specific introns classes, we constructed maximum likelihood phylogenetic tree with different classes of bacterial IEPs (Fig. 2). The phylogenetic tree revealed that UHB.I1 belongs to Group IIC-*att*C class, known to be associated with integrons [9] [18] [25]. On the other hand, UHB.I2 clustered with class CL1(IIB1), whereas H.ha.F1 and H.ha.F2 clustered with bacterial class E(IIB). Blastx analysis of the identified IEPs nucleotide sequences against group II intron database [26] [27] confirmed the results of our phylogenetic analysis. The closest hit to UHB.I1 was Ge.s.I1 of group IIC-*att*C class from *Geobacter sulfurreducens* with 53% identity and 67% similarity. Since the group II intron database is limited in number of sequences, we blasted the sequence against the vast NCBI nr database, better hits were obtained, as the best hit was a group II intron RT from a *Verrucomicrobia* bacterium (sequence ID: NBB81160.1) with 76% identity and 85% similarity. In case of UHB.I2, Sh.sp.I2 (CL1/IIB1), the closest hit was a *Shewanella* sp., with 53% identity and 69% similarity when blasted against group II introns database, whereas its closest hit on NCBI was a group II intron RT from *Legionella birminghamensis* (sequence ID: WP_054523790.1) with 56% identity and 70% similarity.

Multiple sequence alignment of UHB.I1, UHB.I2, H.ha.F1 and H.ha.F2, each with its closely related IEPs showed all required domains for IEPs but lacking the endonuclease domain (En^-) (Additional file 1 Fig. S1 and S2).

The aligned part of H.ha.F1 and H.ha.F2, which covers 60% of H.ha.F1 C-terminus, showed 95% identity to each other, with Ps.tu.I1 (E/IIB) from *Pseudoalteromonas tunicata* being their closest homolog. Both H.ha.F1 and H.ha.F2 showed 70% similarity to Ps.tu.I1 (E/IIB). H.ha.F1 and H.ha.F2 had also shown 63.1-64.3% similarities to IEPs from one uncultured archaeon ANME-1 (UA.I6, UA.I7 and UA.I8). In case of

H.ha.F1, an internal stop codon and a 79 bp-deletion were identified which most likely led to a frameshift and loss of RT3 and RT4 domains; whereas in H.ha.F2, the N-terminus, with domains RT0-4 necessary for the RT function, was absent (Additional file 1 Fig. S3).

Group II intron RT/maturases from Tanatar-5 hypersaline Soda Lake (TSL1) harbors a typical group IIC-*att*C intron within an array of gene cassettes

In order to identify group II introns to which the identified IEPs belong, sequences flanking these IEPs were further analyzed. In case of TSL1, a complete group IIC-*att*C intron (UHB.I1) was detected within an array of gene cassettes inserted in opposite orientation to the adjacent gene cassettes. As the array of gene cassettes detected within the TSL1 contig (9835bp) are at the periphery of the contig, it is not clear whether it is a part of a complete integron or a CALIN lacking the adjacent *int*I gene (Fig. 3A, Additional file 1 Fig. S4 and S5 and Additional file 2 Table S1).

Based on alignment with closely related introns, UHB.I1 boundaries were identical to the sequence 5'-GTGTG…AT-3'. The predicted intron structure, done by MFOLD [28] showed the essential sequences in DI and DV, putatively involved in intron mobility such as EBS1 and ESB3 (Fig. 4).

Examination of UHB.I1 flanking exons was done to determine IBS1 and IBS3 target sites in the 5' and 3' exon, respectively (Fig. 5). Aligning the intron flanking exons with 10 exons from homologous introns showed high conservation only in the 5' exon (Fig. 5). Examination of the 5' exon showed a predicted stem-loop structure (Additional file 1 Fig. S6A). A putative *att*C site bs was detected, but with an atypical AAT triad within the R box, instead of the conserved AAC triad (Additional file 1 Fig. S6B). Additionally, the intron was inserted within the conserved target sequence of Group IIC-*att*C: TTGT/T (Additional file 1 Fig. S6B).

The gene cassette at which the intron is inserted has 3 other ORFs, 2 encode for conserved hypothetical proteins, while the first ORF encodes for a putative serine hydrolase (betalactamase transpeptidase). Two other gene cassettes within TSL1 encode for type II toxin-antitoxin (TA) systems. Other ORFs within the array either encode for conserved hypothetical proteins or show no similarities with proteins in nr database (Fig. 3A and Additional file 2 Table S1).

Being at the periphery of the TSL1 contig, the 5' region of the detected gene cassette array seems to be missing. Thus, it is not clear whether it is a CALIN or part of a full integron with essential integron components at the 5' region such as *int*I gene, *att*I and P_C promoter. Since previous studies showed that internal promoters within the oppositely inserted introns can drive the expression of gene cassette ORFs at the 3' end of the array (those present after the intron) [3], we searched for the presence of putative promoters within UHB.I1. Two putative promoters were detected (Additional file 1 Fig. S4 and Additional file 2 Table S1). It seems that these putative promoters are responsible for the expression of just one downstream ORF encoding for a hypothetical protein, since the TA operon in the next gene cassette had its own predicted promoter (Additional file 2 Table S1).

TSL2 and a CALIN within *Halorhodospira halochloris* genome harbor group IIB introns

Following the same steps described for the identification of UHB.I1 in TSL1, we examined the sequences surrounding the detected group II intron RT in TSL2 and within *H. halochloris* CALIN. Unexpectedly, we identified group IIB introns associated with gene cassette arrays in TSL2 and in the genome of *H. halochloris*. In TSL2 (9772bp contig), a full group IIB1 intron was detected, with its IEP belonging to CL1 class (Fig. 3B, Additional file 1 Fig. S4 and S5, and additional file 2 Table S1). Unfortunately, the array was at the periphery of the contig, as with the TSL1 contig. Thus, the 5' region of the integron or the CALIN was missed and the identified ORF in the first gene cassettes was relatively short (144 bp) with no start codon (Fig. 3B and Additional file 2 Table S1).

The secondary structure of the intron showed a typical IIB intron with essential sequences required for intron folding and base pairing with target site, except for IBS3-EBS3. EBS3 base exists within a bulge at the folded structure [1]; however, the anticipated bulge was absent (Fig. 6). The intron boundaries were different from the known consensus sequence 5'-GUGYG..AY-3', as the boundaries in this case were 5'-UUGCG..GU-3'. Unlike group IIC-*att*C introns, UHB.I2 was inserted in the same orientation of the gene cassettes in the array. Several promoters were predicted within UHB.I2 that could serve as promoters for downstream ORFs in the array (Additional file 1 Fig. S4 and Additional file 2 Table S1). Although upstream stem-loop structures were only reported within group IIC introns, we detected UHB.I2 intron immediately after an *att*C site in the array (Additional file 1 Fig. S6C). Examination of UHB.I2 flanking exons with homologous introns showed poor conservation for both exons except for the first 2 nucleotides in 3' exon (Fig. 7).

H. halochloris introns identified within its CALIN (H.ha.F1 and H.ha.F2) were both fragmented at their 5' end, and we only identified their 3' end of the intron (DV and DVI) and part of the IEP ORFs (Additional file 1 Fig. S4 and S5). Folding of DV and DVI, depicting the 3' part of a group IIB intron were predicted in both intron RNAs (additional file 1 Fig. S7). Here again, putative promoters were predicted within H.ha.F1 and H.ha.F2 (Additional file 1 Fig. S4 and Additional file 2 Table S1). In all cases putative promoters directly upstream of all identified introns were detected (Additional file 2 Table S1).

Gene cassette arrays with identified group II introns are all associated with type II toxin-antitoxin (TA) systems

Following the identification of group II introns within integrons and CALINs, we analyzed other genetic components within these integrons. All detected ORFs within these integrons were BLASTed and annotated. We found that the three examined arrays contain type II TA systems of various types (Fig. 3 and Additional file 2 Table S1). Two TA gene cassettes within TSL1 array were detected. In case of *H. halochloris* CALIN, most of the ORFs identified within the gene cassettes belonged to toxins and antitoxins of type II TA systems giving rise to five TA systems within the CALIN. Three of the five TA systems were of the same type (BrnT/A family). Both H.ha.F1 and H.ha.F2 were inserted within gene cassettes with TA operons. However, in the gene cassette at which H.ha.F2 is inserted, a frameshift within

the *HicA* toxin gene was found, casting doubt on its possible expression. In case of TSL2, the TA system identified was just downstream of the last *att*C site in the array.

An insertional sequence element (IS*200/605)* lies directly downstream of *H. halochloris* CALIN

The relatively small length of TSL1 and TSL2 contigs limited our ability to search for *int*I genes or other MGEs close to the identified gene cassette arrays. This was not an obstacle in case of *H. halochloris* due to the availability of its full genome sequence. Examination of *H. halochloris* genome revealed the presence of just one CALIN with absence of *int*I genes in the whole genome. This CALIN contained 10 gene cassettes, with 6 ORFs in one gene cassette (detailed annotations in Additional file 2 Table S1). Directly, downstream of the last *att*C in the identified CALIN, we found a new insertion sequence (IS) (Fig. 3C), that we submitted to the ISfinder database [24] under the name IS*Hahl1*. It belonged to the complex IS*200/605* family that has no inverted repeats. Instead, palindromic hairpin structures were identified at both ends. Such structures are known to be involved in transposition [29]. The hairpin structures were compared to that of IS*CARN6* (the closest homologue in ISfinder database); and showed 66% identity to IS*Hahl1* (Additional file 1 Fig. S8).

Two ORFs of opposite orientations were identified within IS*Hahl1*; *tnp*A and *tnp*B. The former (80% identity to IS*CARN6* TnpA) encodes for a putative HUH enzymes superfamily transposase, whereas the latter (56% identity to IS*CARN6* TnpB) encodes for an accessory protein that is speculated to be involved in negative regulation of transposition [29]. The configuration of the two ORFs is characteristic of IS*605* group within the IS*200/605* family [29]. A second IS*605* group IS element was identified about 50 kb upstream, with 57% identity to IS*Hahl*1 and 68% identity to *Escherichia coli* IS*Ec46,* and inserted in the opposite orientation to IS*Hahl1*. However, both *tnp*A and *tnp*B of the second identified IS*200/605* element had frameshifts, most probably rendering them nonfunctional.

To determine if the studied genetic elements in *H. halochloris* are transcribed from leading or lagging strands, we searched for the origin of replication (*Ori*C). GammaBOriS tool [30] results showed that the most probable *Ori*C position lies in position 2,834,560-bp-*H.haochloris*-genome and between2787842-2789091 bp. Based on this position, the top strand of the gene cassettes in the identified CALIN seems to be transcribed on the leading strand. This also means that H.ha.F1, H.ha.F2, IS*Hahl1* and the mutated IS*200/605* family member are on the leading strand, while the *att*C sites' bottom strands in the CALIN are on the lagging strand.

Discussion

Identification of integron-associated group II introns sequences from a hypersaline metagenome and in *H. halochloris*

Presence of group II introns has been reported in different bacterial, archaeal and organeller genomes [1]; however, their association with integrons has been limited to IIC-*att*C subclass [9] [18] [25]. To date, none of these integron-associated-introns have been found in halophiles. Here, we have analyzed group II

introns associated with integrons and CALINs in publically 104 available halophile genomes and previously assembled 28 hypersaline metagenomes (a total of 658,054 contigs corresponding to 1,236,831,758 bp) in an attempt to understand the role of specific mobile genetic elements in environmental adaptation of halophiles. We have detected integron-associated-group II introns, class IIC-*att*C and class IIB in the metagenome of the hypersaline Tanatar-5 Soda lake, in Russia and in the genome of the extreme halophile *Halorhodospira halochloris.* Intriguingly, we did not find any group II introns associated with integrons in the remaining analyzed metagenomes. However, we cannot rule out the probability of detecting integron-associated-group II introns in other hypersaline metagenomes. Our findings infer an adaptation role for these integrons in hypersaline alkaline environments. Group II introns have high biotechnological potential, where few members belonging to IIA and IIB classes, have already been commercialized as targetrons [20].

Our newly detected group IIC-*att*C intron, UHB.I1, from the metagenome of the hypersaline Tanatar-5 lake in Russia, is inserted in opposite orientation to the transcription of the adjacent gene cassettes, which is typical of group IIC-*att*C introns [9] [18] [25]. On the other hand, UHB.I2, isolated from the same metagenome, belonged to group IIB1 rather than group IIC and its IEP clustered with CL1 class. This intron was in the same orientation of the gene cassettes transcription, just downstream an *att*C site. Unlike other group II introns, group IIC introns possess a stem-loop structure upstream of the insertion site [1] [9]. *att*C bs seems to serve the function of the upstream stem-loop, in group IIC-*att*C, as known IIC-*att*C introns are inserted within putative *att*C bottom strands [9]. Although in case of group IIB introns, no role of upstream secondary structures has ever been reported, it is intriguing to speculate a role of the secondary structure in the identification of target site, as the *att*C top strand can also form a non-recombinogenic hairpin.

Upon examination of the flanking exons of both TSL introns, there were no clear subclass conservation in flanking exons. However, 5’ exon, rather than 3’ exon, was highly conserved among group IIC-*att*C introns (Fig. 5). In addition, in UHB.I1 a +4G residue was detected, which was previously found to be required for 3’ exon recognition in *Sinorhizobium meliloti* RmInt1 intron (IIB3/D class) [31] (Fig. 5). On the other hand, although no sequence conservation was found in exons flanking UHB.I2, it showed an AT rich 3’ exon (Fig. 7). The same observation was found with *Lactococcus lactis* Ll.LtrB intron (group IIA), where reverse splicing was inhibited by increasing the exon’s GC content [14]. Further experiments should be performed to determine the role of the UHB.I1 AT rich 3’ exon in reverse splicing. UHB.I2 intron seems to fold into nearly typical group IIB intron secondary structure yet the bulge containing the EBS3 site in the DI coordination loop, was missing (Fig. 3B). It is likely that IBS3 on the target site interacts with an alternative EBS3 site or position.

Identification of putatively essential ssDNA stretches for group II intron mobilization and *att*C sites recombination in *H. halochloris*

All our identified IEPs lacked an endonuclease domain (En^-), which is in more than half the bacterial group II introns IEPs [1] [6]. Since En^- IEPs are incapable of a second strand cleavage, they depend on the

host replication machinery for insertion into new target sites [1]. In the case of group IIC introns, the formation of secondary structures is crucial for insertion [9].

Based on GammaBOriS [30] identification of the origin of replication in *H. halochloris*, H.ha.F1 and H.ha.2 are inserted within the leading strand rather than the lagging strand; a documented yet rare phenomena [1]. Furthermore, despite the above mentioned reliance of En^- IEPs group II introns on host replication machinery for complete retrohoming and retrotransposition, a possible minor retrohoming pathway independent of DNA replication can exist, at which introns can reverse splice into double stranded (ds) or transiently ssDNA target sites [1].

In *att*C recombination, replication is not only important for the formation of the folded bs, but also for the resolution of recombination products [32] [33]. However, the presence of single stranded proteins (SSP) hampers the formation of a fully folded *att*C bs in absence of integron integrase (IntI) [34] [35]. In the absence of IntI, an equilibrium between the opened *att*C bs and a partially structured *att*C bs which forms a complex with SSPs exists [35]. We did not detect *intI* genes in the genome of *H. halochloris,* despite the presence of a CALIN. Therefore, the role of these gene cassettes in the absence of *int*I in the genome of *H. halochloris* raises a question of whether they function just as reservoirs for horizontal transfer of gene cassettes or they have an unidentified role. The identified introns within *H. halochloris* CALIN, H.ha.F1 and H.ha.F2 are both 5’ truncated introns and only the 3’ end was identified, and important RT domains within their IEP ORFs were also absent most probably leading to non-functional IEPs. It is already documented that truncated introns are more common than full-length introns in bacterial genomes [10]. Yet, a putatively functional IEP ORF (80% identical to H.ha.1 IEP) was detected, about 6.5 kb upstream of the CALIN (Acc.no WP_096410353.1). Perhaps both H.ha.F1 and H.ha.F2 were formed as a result of incomplete reverse transcription due to replication slippage caused by the presence of hairpin structures. Manually and with the aid of MFOLD [28], we have detected an *att*C-like structure upstream of H.ha.F1 (Additional file 1 Fig. S9A) and a putative *att*C site upstream of H.ha.F2, showing a nearly typical *att*C site bs secondary structure (Additional file 1 Fig. S9B). Again, the presence of these secondary structures before group IIB introns further suggests their possible role in recognition of target sites.

Clustering of MGEs requiring ssDNA in hypersaline group II introns

Coexistence of group II introns, integrons and IS elements may have a combined role in increasing genomic plasticity in extreme hypersaline environments. In *H. halochloris* CALIN, we have identified directly downstream of the last gene cassette, at which H.ha.F2 is inserted, a new IS element “IS*Hahl1*”. IS*Hahl*1 belongs to IS*605* group of IS*200/605* family were t*np*A and *tnp*B are transcribed in opposite directions.

Insertion sequences belonging to IS*200/605* family are distinguished from other IS elements by their transposition mechanism; 1- utilizing obligatory ssDNA intermediates, 2- absence of nucleotides loss or gain, 3- requiring transposase “TnpA” belonging to the “HUH” superfamily of enzymes rather than the “DDE” family of classical IS elements [29] [36] and 4- the presence of hairpin structures at both ends [29].

Transposition is strand specific and follows a “peel and paste” mechanism at which an excised circular single stranded intermediate integrates into a single stranded target site [29]. For transposition to take place, both ends need to be single stranded at the same time. Thus, a link between IS*200/605* family members’ transposition and replication was reported, with more frequent transposition into the lagging strand [29]. Unexpectedly, the IS active “top” strand that carries the target sequence was found on the leading strand, yet *tnp*A gene was transcribed on the lagging strand. In some cases, presence of IS*200/605* elements on the leading strand was attributed to genomic rearrangements [29]. In fact, it was suggested that identical IS*605* elements in *H. pylori* had caused rearrangement within its genome [37]. Although in the *H. halochloris* genome we have identified a poorly homologous IS*506* element on the opposite strand (57% identity to IS*Hahl1*), it is unlikely that homologous recombination can occur between the 2 elements. The rationale behind our mining for similar IS element was to inspect the possibility of mobilization of the adjacent CALIN sequence. Yet the difference between the sequences of the 2 identified IS*605* group element and the large distance between them (¬50kb) confines this possibility. Even though previous studies reported a link between IS*200/605* transposition and replication, high transposition frequencies were reported with DNA repair mechanisms when large ssDNA stretches become available [38].

It is interesting to note the clustering of different genetic elements (*att*C sites, group II introns and IS*200/605*) that require single stranded and secondary structures for function. These elements have been linked to replication as one of the main sources for ssDNA [1] [29] [32] [39]. Further experimental studies should be performed to delineate the interaction between the gene cassettes, group II introns and IS*200/605* elements from hypersaline environments.

Abundance of Toxin-Antitoxin (TA) systems in hypersaline integron-associated structures

Finally, our analysis showed abundance of TA systems belonging to different classes in all identified arrays. In both TSL1 and TSL2, a TA system was detected directly downstream of the last gene cassette in the array (Fig. 2A and B). The abundance of TA systems as gene cassettes within integrons has already been observed in different studies [15]. It is hypothesized that TA systems could have a role in maintaining the integrity of these integrons by preventing deletions of existing arrays [15] [40]. Nonetheless, the accumulation of 6 different TA systems within the identified *H. halochloris* CALIN is intriguing. In fact, both H.ha.F1 and H.ha.F2 truncated introns were inserted into gene cassettes composed of a TA operon, although in case of H.ha.F2, a frameshift due to a one nucleotide deletion in the HicA family toxin ORF is observed. In addition, 3 TA systems of the BrnT/A family were detected within the CALIN. The claimed hypothesis that TA systems are important for the integrity and maintenance of the adjacent chromosomal structures indicates that adjacent gene cassettes and even secondary structures have unraveled essential roles. Moreover, the large number of expressed TA systems in a genome was found to have a role in increasing the population of persisters that can survive under different stress conditions [40]. It is therefore not surprising that the detected TA systems in the metagenome and genome from hypersaline environments would support the adaptation and growth of the persistent halophiles. ParE toxins of TA systems, which were identified in TSL1, TSL2 and *H.*

halochloris CALIN, were shown to induce DNA damage, which in turn induces an SOS response, activating DNA repair mechanisms where ssDNA stretches are formed allowing different transpositions and recombination events to take place [40]. Similarly the identified TA cassettes from hypersaline environments can increase mobilization of different MGEs such as integron gene cassettes, prophages and transposons [40].

Conclusions

Integrons and integron-like sequences have been particularly associated with Group IIC-*att*C introns. In this study we identified a Group IIC-*att*C from the hypersaline Tanatar-5 Soda lake metagenome in Russia and named it “UHB.I1”. We have also detected different classes of group IIB introns within gene cassette arrays in the same metagenome and in a CALIN in the extreme halophile *H. halochloris*. These findings could help decipher the role of group II introns associated with integrons or integron-associated sequences in hypersaline environments. A new insertion sequence IS*Hahl1*, belonging to IS*200/605* elements was also identified adjacent to *H. halochloris* CALIN. The clustering of different MGEs, particularly those requiring single-stranded secondary structures for their function, suggests interplay between these different elements and cellular processes such as replication, transcription and horizontal gene transfer of prokaryotes residing in hypersaline environments. The abundance of toxin-antitoxin systems in all our studied gene cassette arrays, either as gene cassettes or right after the last *att*C site, strengthens their potential role in maintaining the integrity of the adjacent arrays, enhancing the mobility of adjacent mobile elements and increasing the persistence of the cells to adapt to their hypersaline and alkaline environments.

Methods

Analyzed samples

We analyzed publicly available metagenomic assemblies from different hypersaline environments (28 assemblies of a total of 1,236,831,758 bp and 658,054 contigs) in addition to completely or partially sequenced genomes of halophilic bacteria (24 complete and 33 partial with a total size of 202.81 Mb) and archaea (25 complete and 22 partial with a total size of 166.02 Mb). Table 1 shows all analyzed assemblies, whereas a list of halophilic bacteria and archaea was obtained from the Halodom database [41] in November 2019: “halodom.bio.auth.gr” (Additional file 2 Table S2 and S3). The analyzed metagenomic assemblies were all available already assembled hypersaline metagenomes on NCBI or from our lab. For comparative reasons, metagenomic assemblies from 22 marine and 7 freshwater environments (1,750,281,271 bp and 1,444,498 contigs) were subjected to the same analysis (Additional file 2 Table S4). The marine assembled metagenomes were selected from different geographical locations, different depths if applicable with a tendency towards choosing those with smaller number of contigs for easier processing. In case of freshwater assemblies, we used all publicly available assembled metagenomes on NCBI.

Table 1. Analyzed metagenomic assemblies from different hypersaline environments

Site	Description	Assembly Accession number or reference	Total assembled sequence length	Number of contigs
GR	Grendel Spring, Yellowstone National Park, Wyoming, USA	GCA_900244995.1	33631634	11151
GNM1	Guerrero Negro mat, Mexico 0-1mm depth	GCA_000206585.1, [42], [43]	8530607	11351
GNM2	Guerrero Negro mat, Mexico 1-2mm depth	GCA_000206565.1, [42], [43]	7390978	10551
GNM3	Guerrero Negro mat, Mexico 2-3mm depth	GCA_000206545.1, [42], [43]	8209846	11423
GNMt4	Guerrero Negro mat, Mexico 3-4mm depth	GCA_000206525.1, [42], [43]	8130049	11724
GNM5	Guerrero Negro mat, Mexico 4-5mm depth	GCA_000206505.1, [42], [43]	9689398	14128
GNM6	Guerrero Negro mat, Mexico 5-6mm depth	GCA_000206485.1 ,[42], [43]	8291075	11380
GNM7	Guerrero Negro mat, Mexico 6-10mm depth	GCA_000206465.1, [42], [43]	9759240	13649
GNM8	Guerrero Negro mat, Mexico 10-22mm depth	GCA_000206445.1, [42], [43]	7914434	11356
GNM9	Guerrero Negro mat, Mexico 22-34mm depth	GCA_000206425.1, [42], [43]	8308787	11596
GNM10	Guerrero Negro mat, Mexico 34-49mm depth	GCA_000206405.1, [42], [43]	7132956	10297
ATII SDM	Atlantis II Deep Brine Sediment, Red Sea	[44], [45], [46]	40413330	41726
DD SDM	Discovery Deep Brine Sediment, Red Sea	[44], [45], [46]	52421642	51829
Th	Thetis Mediterranean deep-sea hypersaline lakes	GCA_001684355.1	13102297	10347
ATII INF	Atlantis II Deep Brine interface, Red Sea	[46], [47]	16014945	24317
DD INF	Discovery Deep Brine interface , Red Sea	[46], [47]	11647401	18413
KD UINF	Kebrit Deep Upper interface, Red Sea	[46], [47]	42652688	45750
KD LINF	Kebrit Deep Lower interface, Red Sea	[46], [47]	50280352	74666

ATII LCL	Atlantis II Deep Brine, Lower convective layer, Red Sea	[46], [47]	46518597	43555
ATII UCL	Atlantis II Deep Brine,Upper convective layer, Red Sea	[46], [47]	21343827	29592
DD BR	Discovery Deep Brine , Red Sea	[46], [47]	12244355	18850
KD BR	Kebrit Deep Brine, Red Sea	[46], [47]	35162057	74666
TSL	brine of Lake Tanatar-5 (Soda Lake), Russia: Kulunda steppe	GCA_001564335.1	193970398	19350
TTCSL	brine of Tanatar trona crystallizer (Soda Lake), Russia: Kulunda steppe	GCA_001563815.1	106596264	9426
PSL	brine of Picturesque Lake (Soda Lake), Russia: Kulunda steppe	GCA_001564315.1	251189393	25098
Ty	Lake Tyrrell, Victoria, Australia	GCA_000347535.1, [48], [49]	62549170	15008
Na	Namib Desert Hosabes playa, Namibia	GCA_001543535.1	10867082	11304
BSL	brine of Lake Bitter-1 (Soda Lake), Russia: Kulunda steppe	GCA_001563825.1	152868956	15551

Identification of integrons and CALINs

Integron finder version 2.0 [19] was used to search for complete integrons, Integron integrase genes (*intI*) and CALINs in hypersaline metagenomic assemblies and genomes of different halophiles. We used the option “local detection” on the command line with all contigs and an 8 kb distance threshold between successive identified *attC* sites to ensure the detection of all potential *attC* sites. At least 2 *attC* sites should be detected within the 8 kb threshold to be reported as a positive result. A search for integron cassette promoters (P_C) and primary recombination sites (*attI*) for known integron classes (1, 2 and 3) was also performed.

Identification of group II introns

Identified sequences were further inspected by running BLAST search of all identified ORFs within gene cassettes against NCBI nr BLAST database. ORFs identified as group II RT/maturase were further analyzed by blastx against group II intron database (http://webapps2.ucalgary.ca/~groupii/) [26] [27] and their amino acid sequences were aligned with close hits in order to identify IEP different domains that were defined in group II intron database (http://webapps2.ucalgary.ca/~groupii/html/static/orfalignment.php) [26] [27]. Identification of intron boundaries was done by the MFOLD webserver, which folds the introns RNA structure [28] based on the

known secondary structures of group II intron classes, that showed the high similarity to our newly identified introns. First, for each identified Group II intron RT, the region downstream of the ORF was aligned with 3-6 sequences from close hits obtained by blast using MUSCLE [50] [51]. This was done to identify the most conserved DV in addition to DVI and the 3' boundary of the intron. This was followed by searching for the basal stem of DIV by looking for a sequence complementary to the sequence just upstream DV within the ORF or within 200bp upstream of the ORF start codon. Identification of the 5' domains (DI, DII and DIII) was mainly done by searching for a putative 5' boundary following the consensus sequence GUGYG and folding into a structure similar to the consensus structure of the identified group II intron class. Even with the low sequence conservation in upstream domains, multiple sequence alignment with close introns helped in determination of the final folding structure. Moreover, exon binding sequences (EBS1, 2 and 3) and sequences involved in tertiary structures such as α-α', β-β', δ-δ', □-□' and γ-γ' Watson-Crick base pairs, ζ-ζ' and η-η' tetraloop-receptor interactions and κ-κ' and λ-λ' non Watson-Crick interactions [7] were determined manually whenever applicable. The final secondary structure was then depicted using Pseudoviewer3 [52].

Sequence logos of intron boundaries and 5' and 3' exons of each identified intron with its closest homologues (obtained by Blastx against group II intron database) were illustrated using WebLogo ver. 2.8.2 [53].

Detection of introns upstream regions, including rho independent terminators or hairpin structures, were done using ARNold webserver [54] [55] and MFOLD [28], respectively.

Insertion sequences identification

ISfinder [24] was used to search for insertion sequences within contigs or genomes at which integrons or CALINS were identified.

ORFs annotation and promoter predictions

All predicted ORFs within identified gene cassettes were manually curated and annotated based on Blastx results against NCBI nr database. Search for promoters for gene cassettes and within group II introns was done using bprom tool [56].

Phylogenetic analysis

34 bacterial IEPs from different classes were aligned to the 4 identified IEPs in this study using MUSCLE [51], along with Mitochondrial IEP from Liverwart *Marchantia polymorpha* which was used as an outgroup. Molecular phylogenetic analysis was done with MEGA7 [57] using the Maximum Likelihood with LG substitution model. The tree was drawn to scale, with branch lengths depicting the number of substitutions per site. Statistical support of the tree was done by bootstrap analyses with 1,000 samplings.

Determination of *H. halochloris* leading and lagging strands

GammaBOriS tool specifically designed for identification of origin of replication (*Ori*C) sequences in gammaproteobacterial chromosomes [30] was used for identification of probable *H. halochloris Ori*C. Based on the approach used by Mao et al [58]. The position of the replication termination site was roughly calculated as half of the genome DNA sequence starting from the identified *Ori*C. The leading and lagging strands of each half was then determined based on the knowledge that the leading strands encodes for a much larger number of genes than the lagging strand [58].

List Of Abbreviations

bs Bottom strand

CALIN Clusters of *att*C sites lacking a neighboring integron-integrase

cDNA Complementary DNA

CL Chloroplast-like

dsDNA Double-stranded DNA

EBS Exon binding site

En Endonuclease

IBS Intron binding site

IEP Intron encoded Protein

intI Integron integrase

IS Insertion Sequence

MGE Mobile genetic element

ML Mitochondrial-like

ORF Open reading frame

qRT-PCR Quantitative real time reverse transcriptase ploymerase chain reaction

RNA-seq RNA sequencing

RNP Ribonucleoprotein

RT Reverse transcriptase

ssDNA Single-stranded DNA

TA Toxin-Antitoxin

Declarations

Availability of Data and Materials

Accession numbers to all used publically available assembled metagenomes are included in methods section (Table 1) and Additional file 2 Table S4. Assembled metagenomes from Red Sea brine pools and from hydrothermal vents are available from the corresponding author on reasonable request.

List of halophilic bacteria and archaea was obtained from the Halodom database [41] in November 2019: (halodom.bio.auth.gr) and accession numbers of analyzed genomes are included in Additional file 2 Tables S2 and S3.

Group II introns sequences and their corresponding IEPs were obtained from group II intron database (http://webapps2.ucalgary.ca/~groupii/html/static/intro.php) [26] [27].

Identified IS during this study is submitted at ISfinder (https://isfinder.biotoul.fr/) [24] under the name IS*Hahl*1.

All data generated during this study are included in this published article and its additional files.

Competing Interests

The authors declare that they have no competing interests.

Funding

SS was funded by a Youssef Jameel PhD fellowship. RS was partially funded by the American University in Cairo faculty Research Grant.

Authors' contribution

SS and RS contributed in analysis design and manuscript writing. SS analyzed and interpreted results. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to Dr. Nahla Hussein and Dr. Laila Ziko for their valuable comments on the manuscript.

References

1. Toro N, Jiménez-Zurdo JI, García-Rodríguez FM. Bacterial group II introns: not just splicing. FEMS Microbiol Rev. 2007;31:342–58.

2. Dai L, Zimmerly S. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. Nucleic Acids Res. 2002;30:1091–102.
3. Léon G, Quiroga C, Centrón D, Roy PH. Diversity and strength of internal outward-oriented promoters in group IIC-attC introns. Nucleic Acids Res. 2010;38:8196–207.
4. Zimmerly S, Hausner G, Wu X. Phylogenetic relationships among group II intron ORFs. Nucleic Acids Res. 2001;29:1238–50.
5. Dai L, Zimmerly S. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. Nucleic Acids Res. 2002;30:1091–102.
6. Lambowitz AM, Zimmerly S. Mobile Group II Introns. Annu Rev Genet. 2004;38:1–35.
7. Lambowitz AM, Zimmerly S. Group II Introns: Mobile Ribozymes that Invade DNA. Cold Spring Harb Perspect Biol. 2011;3:a003616.
8. Martínez-Abarca F, Toro N. Group II introns in the bacterial world. Mol Microbiol. 2000;38:917–26.
9. Léon G, Roy PH. Potential role of group IIC-attC introns in integron cassette formation. J Bacteriol. 2009;191:6040–51.
10. Toro N, Martínez-Abarca F. Comprehensive Phylogenetic Analysis of Bacterial Group II Intron-Encoded ORFs Lacking the DNA Endonuclease Domain Reveals New Varieties. PLoS ONE. 2013;8. doi:10.1371/journal.pone.0055102.
11. Cerveau N, Leclercq S, Bouchon D, Cordaux R. Evolutionary Dynamics and Genomic Impact of Prokaryote Transposable Elements. In: Pontarotti P, editor. Evolutionary Biology – Concepts, Biodiversity, Macroevolution and Genome Evolution. Berlin, Heidelberg: Springer; 2011. p. 291–312. doi:10.1007/978-3-642-20763-1_17.
12. Toor N, Robart AR, Christianson J, Zimmerly S. Self-splicing of a group IIC intron: 5' exon recognition and alternative 5' splicing events implicate the stem–loop motif of a transcriptional terminator. Nucleic Acids Res. 2006;34:6461–71.
13. Qin PZ, Pyle AM. The architectural organization and mechanistic function of group II intron structural elements. Curr Opin Struct Biol. 1998;8:301–8.
14. Mohr G, Smith D, Belfort M, Lambowitz AM. Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. Genes Dev. 2000;14:559–73.
15. Cambray G, Guerout A-M, Mazel D. Integrons. Annu Rev Genet. 2010;44:141–66.
16. Gillings MR. Integrons: Past, Present, and Future. Microbiol Mol Biol Rev. 2014;78:257–77.
17. MacDonald D, Demarre G, Bouvier M, Mazel D, Gopaul DN. Structural basis for broad DNA-specificity in integron recombination. Nature. 2006;440:1157–62.
18. Quiroga C, Centrón D. Using genomic data to determine the diversity and distribution of target site motifs recognized by class C-attC group II introns. J Mol Evol. 2009;68:539–49.
19. Cury J, Jové T, Touchon M, Néron B, Rocha EP. Identification and analysis of integrons and cassette arrays in bacterial genomes. Nucleic Acids Res. 2016;44:4539–50.

20. Enyeart PJ, Mohr G, Ellington AD, Lambowitz AM. Biotechnological applications of mobile group II introns and their reverse transcriptases: gene targeting, RNA-seq, and non-coding RNA analysis. Mob DNA. 2014;5:1–19.
21. Vavourakis CD, Ghai R, Rodriguez-Valera F, Sorokin DY, Tringe SG, Hugenholtz P, et al. Metagenomic insights into the uncultured diversity and physiology of microbes in four hypersaline soda lake brines. Front Microbiol. 2016;7:211.
22. Sorokin DY. The Microbial Sulfur Cycle at Extremely Haloalkaline Conditions of Soda Lakes. Front Microbiol. 2011;2. doi:10.3389/fmicb.2011.00044.
23. Garrity G, Brenner DJ, Krieg NR, Staley JR, editors. Bergey's Manual® of Systematic Bacteriology: Volume 2: The Proteobacteria, Part B: The Gammaproteobacteria. 2nd edition. Springer US; 2005. https://www.springer.com/gp/book/9780387241449. Accessed 31 May 2019.
24. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006;34 Database issue:D32-36.
25. Quiroga C, Roy PH, Centrón D. The S.ma.I2 class C group II intron inserts at integron attC sites. Microbiology,. 2008;154:1341–53.
26. Candales MA, Duong A, Hood KS, Li T, Neufeld RAE, Sun R, et al. Database for bacterial group II introns. Nucleic Acids Res. 2012;40:D187–90.
27. Dai L, Toor N, Olson R, Keeping A, Zimmerly S. Database for mobile group II introns. Nucleic Acids Res. 2003;31:424–6.
28. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31:3406–15.
29. He S, Corneloup A, Guynet C, Lavatine L, Caumont-Sarcos A, Siguier P, et al. The IS200/IS605 Family and "Peel and Paste" Single-strand Transposition Mechanism. Microbiol Spectr. 2015;3.
30. Sperlea T, Muth L, Martin R, Weigel C, Waldminghaus T, Heider D. gammaBOriS: Identification and Taxonomic Classification of Origins of Replication in Gammaproteobacteria using Motif-based Machine Learning. Sci Rep. 2020;10:6727.
31. Martínez-Abarca F, Barrientos-Durán A, Fernández-López M, Toro N. The RmInt1 group II intron has two different retrohoming pathways for mobility using predominantly the nascent lagging strand at DNA replication forks for priming. Nucleic Acids Res. 2004;32:2880–8.
32. Loot C, Nivina A, Cury J, Escudero JA, Ducos-Galand M, Bikard D, et al. Differences in Integron Cassette Excision Dynamics Shape a Trade-Off between Evolvability and Genetic Capacitance. mBio. 2017;8:e02296-16.
33. Loot C, Ducos-Galand M, Escudero JA, Bouvier M, Mazel D. Replicative resolution of integron cassette insertion. Nucleic Acids Res. 2012;40:8361–70.
34. Loot C, Parissi V, Escudero JA, Amarir-Bouhram J, Bikard D, Mazel D. The Integron Integrase Efficiently Prevents the Melting Effect of *Escherichia coli* Single-Stranded DNA-Binding Protein on Folded attC Sites. J Bacteriol. 2014;196:762–71.

35. Grieb MS, Nivina A, Cheeseman BL, Hartmann A, Mazel D, Schlierf M. Dynamic stepwise opening of integron attC DNA hairpins by SSB prevents toxicity and ensures functionality. Nucleic Acids Res. 2017;45:10555–63.
36. Chandler M, de la Cruz F, Dyda F, Hickman AB, Moncalian G, Ton-Hoang B. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. Nat Rev Microbiol. 2013;11:525–38.
37. Akopyants NS, Clifton SW, Kersulyte D, Crabtree JE, Youree BE, Reece CA, et al. Analyses of the cag pathogenicity island of *Helicobacter pylori*. Mol Microbiol. 1998;28:37–53.
38. Pasternak C, Ton-Hoang B, Coste G, Bailone A, Chandler M, Sommer S. Irradiation-Induced *Deinococcus radiodurans* Genome Fragmentation Triggers Transposition of a Single Resident Insertion Sequence. PLOS Genet. 2010;6:e1000799.
39. Trinh TQ, Sinden RR. Preferential DNA secondary structure mutagenesis in the lagging strand of replication in E. coli. Nature. 1991;352:544–7.
40. Gerdes K, editor. Prokaryotic Toxin-Antitoxins. Berlin Heidelberg: Springer-Verlag; 2013. doi:10.1007/978-3-642-33253-1.
41. Loukas A, Kappas I, Abatzopoulos TJ. HaloDom: a new database of halophiles across all life domains. J Biol Res-Thessalon. 2018;25:2.
42. Harris JK, Caporaso JG, Walker JJ, Spear JR, Gold NJ, Robertson CE, et al. Phylogenetic stratigraphy in the Guerrero Negro hypersaline microbial mat. ISME J. 2013;7:50–60.
43. Kunin V, Raes J, Harris JK, Spear JR, Walker JJ, Ivanova N, et al. Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. Mol Syst Biol. 2008;4:198.
44. Siam R, Mustafa GA, Sharaf H, Moustafa A, Ramadan AR, Antunes A, et al. Unique Prokaryotic Consortia in Geochemically Distinct Sediments from Red Sea Atlantis II and Discovery Deep Brine Pools. PLoS ONE. 2012;7:e42872.
45. Adel M, Elbehery AHA, Aziz SK, Aziz RK, Grossart H-P, Siam R. Viruses- to -mobile genetic elements skew in the deep Atlantis II brine pool sediments. Sci Rep. 2016;6:32704.
46. Ziko L, Adel M, Malash MN, Siam R. Insights into Red Sea Brine Pool Specialized Metabolism Gene Clusters Encoding Potential Metabolites for Biotechnological Applications and Extremophile Survival. Mar Drugs. 2019;17.
47. Abdallah RZ, Adel M, Ouf A, Sayed A, Ghazy MA, Alam I, et al. Aerobic methanotrophic communities at the Red Sea brine-seawater interface. Front Microbiol. 2014;5. doi:10.3389/fmicb.2014.00487.
48. Podell S, Ugalde JA, Narasingarao P, Banfield JF, Heidelberg KB, Allen EE. Assembly-driven community genomics of a hypersaline microbial ecosystem. PloS One. 2013;8:e61692.
49. Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, et al. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J. 2012;6:81–93.

50. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47:W636–41.
51. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
52. Byun Y, Han K. PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots. Bioinformatics. 2009;25:1435–7.
53. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: A Sequence Logo Generator. Genome Res. 2004;14:1188–90.
54. Gautheret D, Lambert A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. J Mol Biol. 2001;313:1003–11.
55. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res. 2001;29:4724–35.
56. Solovyev V, Salamov A. Automatic annotation of microbial genomes and metagenomic sequences. In: Metagenomics and its applications in agriculture, biomedicine and environmental studies. Nova Science Publishers; 2011. p. 61–78.
57. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016;33:1870–4.
58. Mao X, Zhang H, Yin Y, Xu Y. The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. Nucleic Acids Res. 2012;40:8210–8.

Figures

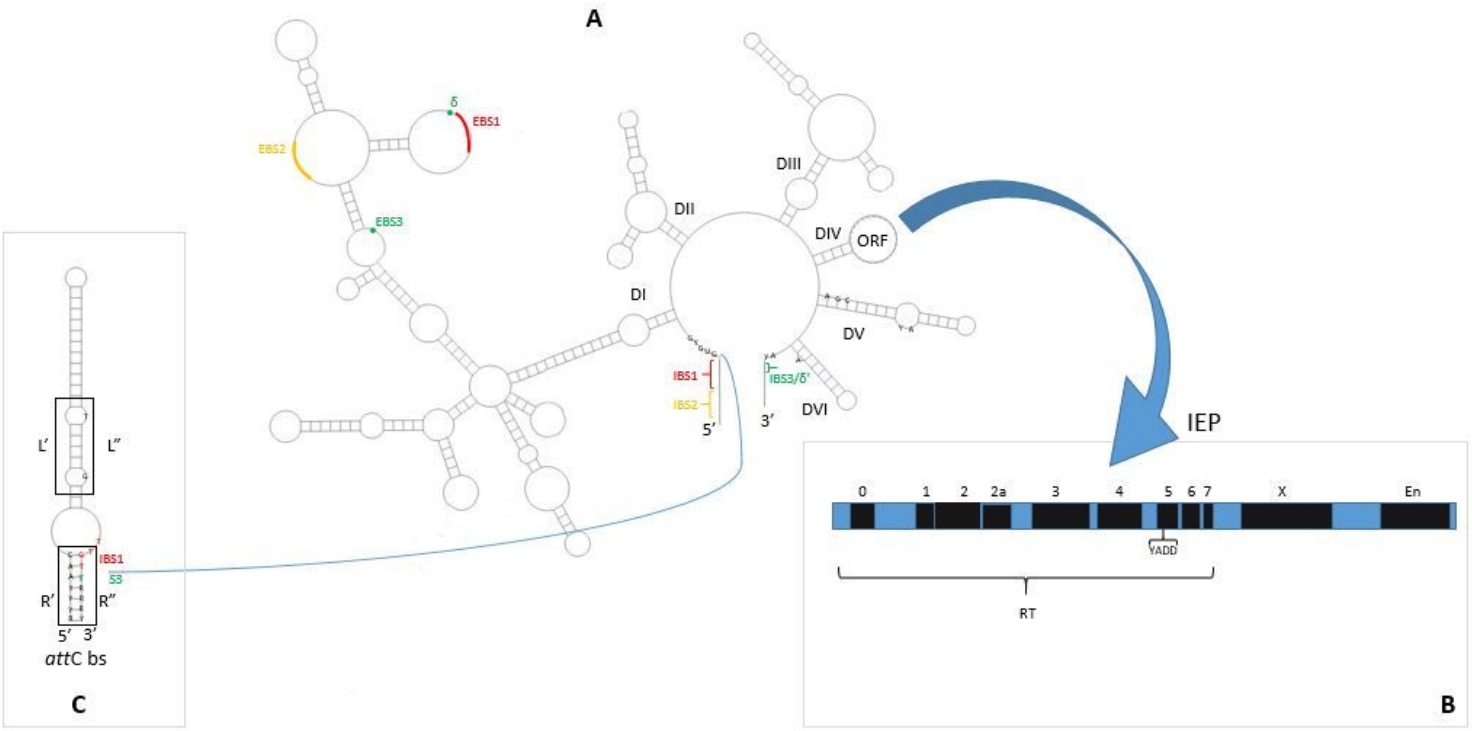

Figure 1

General secondary structure of group II intron RNA, attC site and domains of IEP. Group II intron is composed of 6 domains (DI-DVI) at which DI and DV form its catalytic core (A). Intron encoded protein (IEP) is encoded by ORF in DIV (A). The main domains of an IEP ORF (RT: reverse transcriptase, X: maturase and En: endonuclease are depicted in the schematic diagram of IEP (B). Recognition of target site occurs mainly via base-pairing between short sequences at 5' exon (Intron binding sites IBS1 and 2) with exon binding sites (EBS1 and 2) on the intron and either IBS3 or δ' on exon 3' (based on intron class) with EBS3 or δ on the intron (A). In case of group IIC, IBS2 is replaced by a hairpin structure such as attC bottom strand (bs) in group IIC-attC (C) at which the intron is inserted at the R" sequence into the consensus sequence TTGT/T (IBS1/IBS3).

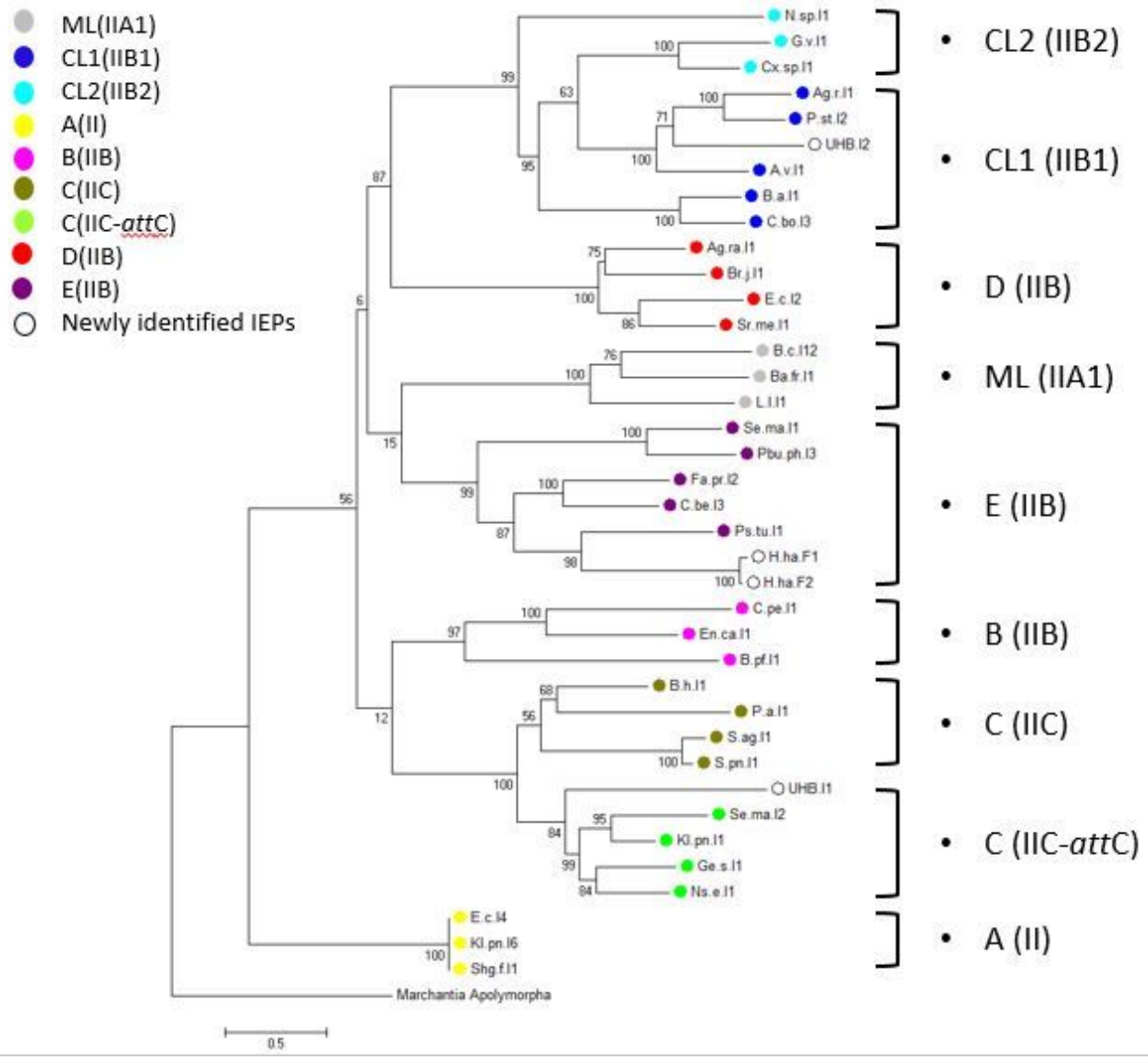

Figure 2

Phylogenetic tree of identified putative IEPs with IEPs from different bacterial groups. UHB.I1 clusters with group IIC-attC, UHB.I2 with group IIB (Chloroplast-like1 class) and both H.ha.F1 and H.ha.2 with group IIB (bacterial class E). IEPs abbreviations are based on their introns nomenclature in group II introns database [26] [27]. Mitochondrial IEP from Liverwart Marchantia polymorpha is used as an outgroup. Bootstrap values are indicated as percentages of 1000 replicates.

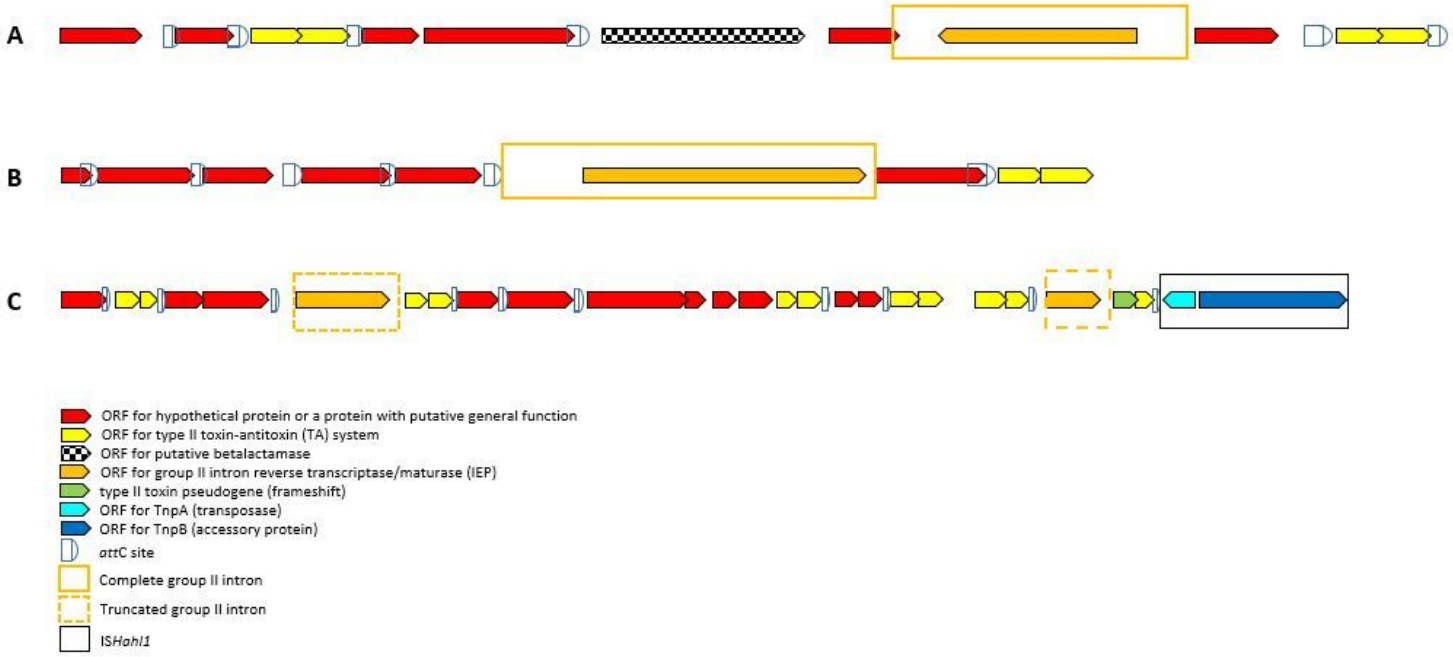

Figure 3

Schematic representation of identified gene cassette arrays where group II introns are inserted. A: TSL1, B: TSL2, C: H. halochloris CALIN. Arrow heads of different ORFs show the direction of their transcription. Colored legend show different genetic elements depicted.

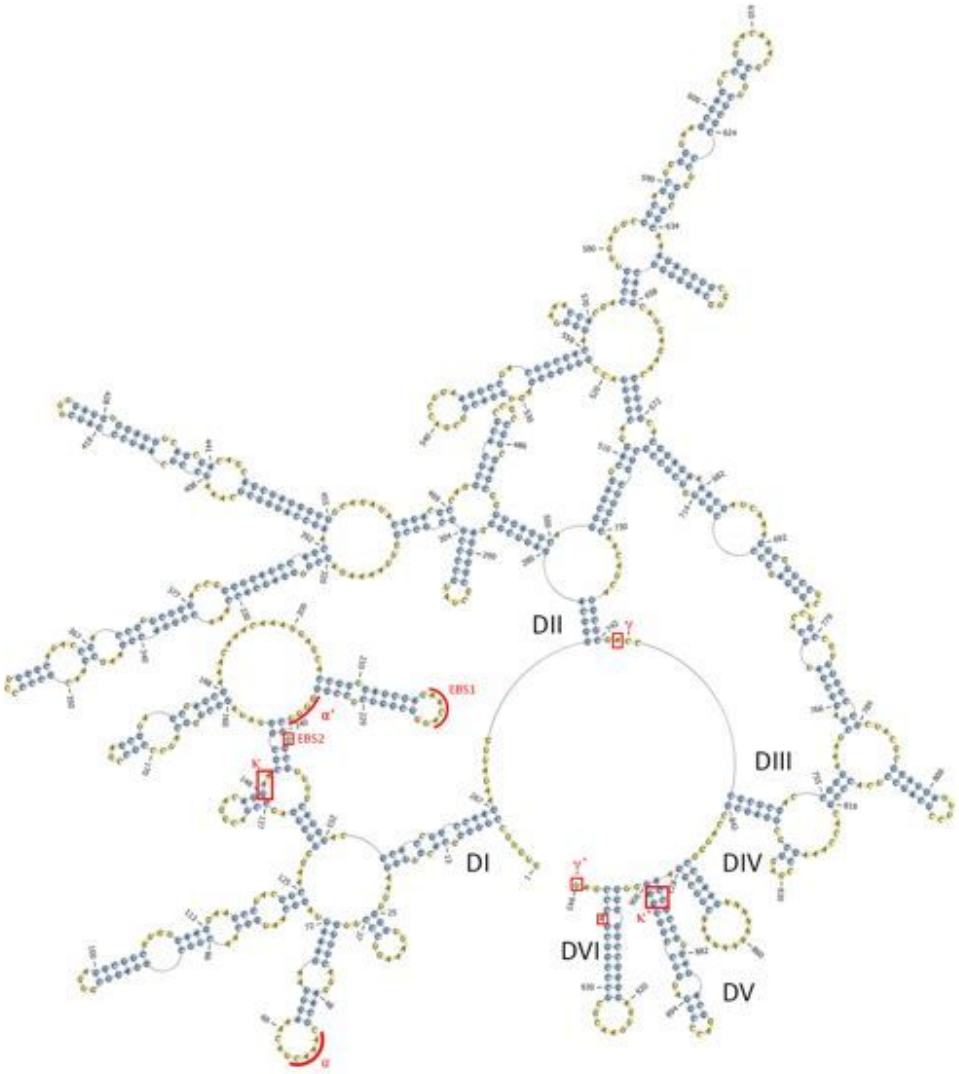

Figure 4

Secondary structure of group II intron UHB.I1. UHB.I1 identified in TSL1 contig shows necessary sequences required for intron splicing and reverse splicing. Important sequences are shown within red rectangles or curved lines. EBS1 and EBS3 are important for base-pairing with target site in flanking exons, whereas other identified sequences are necessary for intron folding (Watson-Crick γ-γ’ and non-Watson-Crick κ-κ’ internal base-pairing).

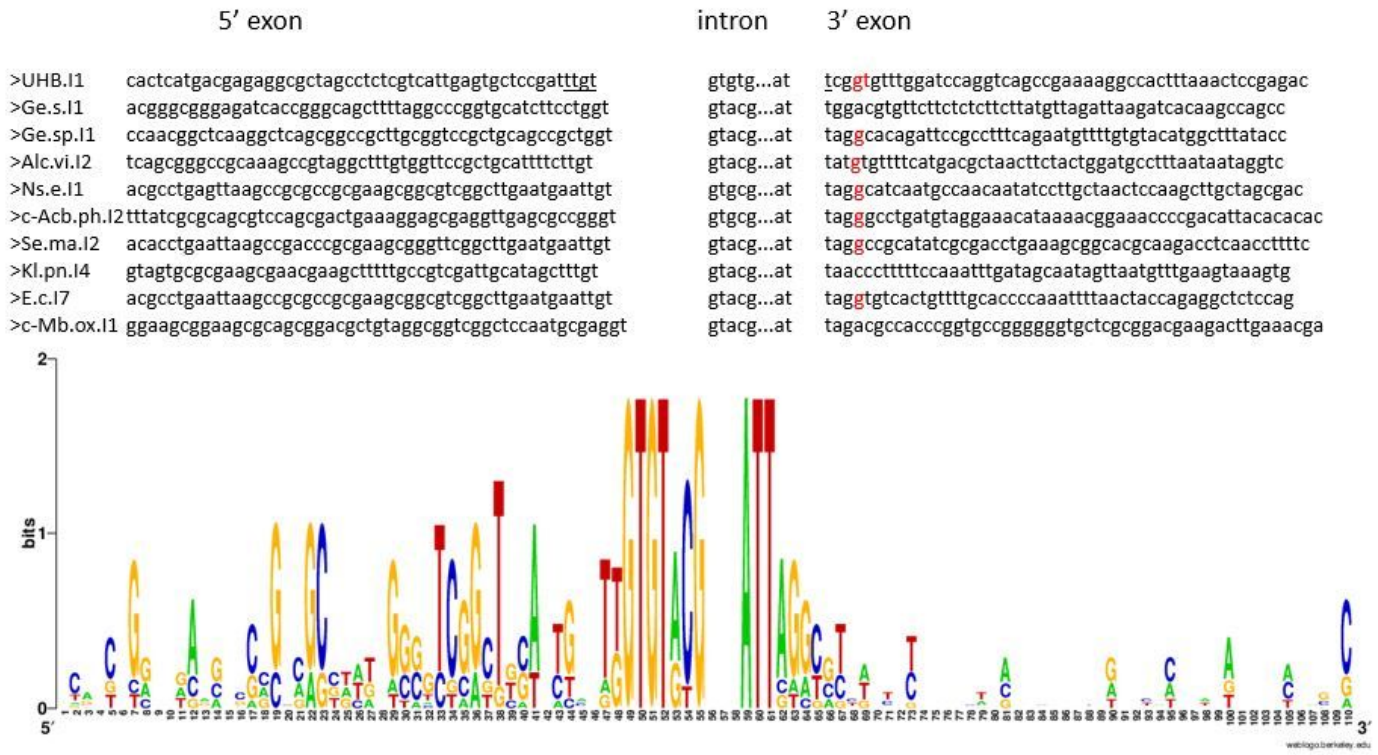

Figure 5

UHB.I1 flanking exons logos with closer hits. The logo shows a conservation in 5’ exon, intron boundaries and a +4G residue at exon 3’ (shown in red). IBS1 and IBS3 in UHB.I1 5’ exon and 3’ exon, respectively are underlined

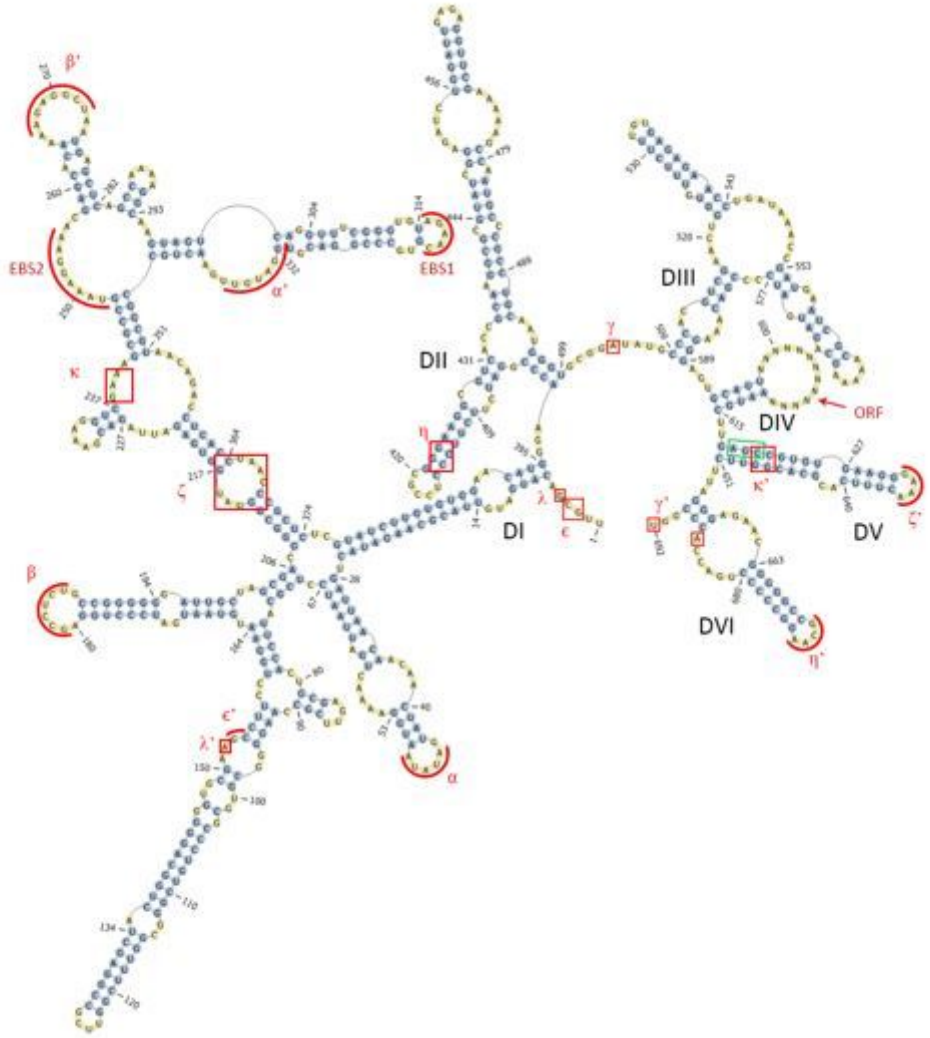

Figure 6

Secondary structure of group II intron UHB.I2. UHB.I2 identified in TSL2 contig shows necessary sequences required for intron splicing and reverse splicing. Important sequences are shown within red rectangles or curved lines. EBS1 and EBS3 are important for basepairing with target site in flanking exons, whereas other identified sequences are necessary for intron folding (Watson-Crick α-α', β-β', □-□'and γ-γ' and non-Watson-Crick κ-κ' and λ-λ' internal base-pairing and tetraloop-receptor interactions ζ-ζ' and η-η'). Conserved catalytic "AGC" triad in DV is shown in a green rectangle.

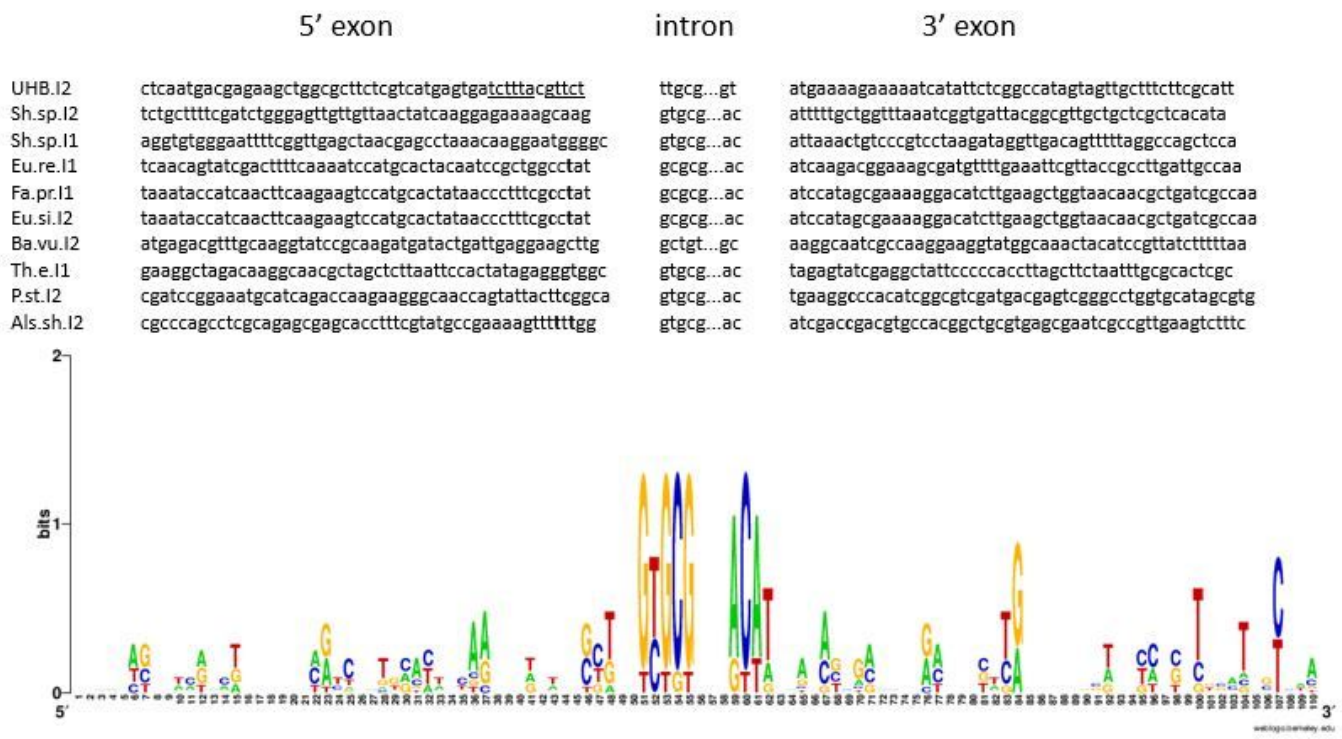

Figure 7

UHB.I2 flanking exons logos with closer hits. The logo shows a conservation in the target site. IBS1 and IBS2 sequences in UHB.I2 5′ exon are underlined.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfiles.docx
- Additionalfile2tables30Aug2020SS.docx
- Additionalfile1figures30Aug2020SS.docx