Summer 8-25-2021

Insights Into Halophilic Microbial Adaptation: Analysis of Integrons and Associated Genomic Structures and Characterization of a Nitrilase in Hypersaline Environments

Sarah Sonbol
sasonbol@aucegypt.edu

THE AMERICAN UNIVERSITY IN CAIRO
الجامعة الأمريكية بالقاهرة

Insights into halophilic microbial adaptation: Analysis of integrons and associated genomic structures and characterization of a nitrilase in hypersaline environments

School of Sciences and Engineering

A Thesis Submitted by

Sarah Ali Ahmed Sonbol. MSc

to the

Applied Sciences Graduate Program

Spring, 2021

In partial fulfillment of the requirements for the degree of

Doctorate in Applied Sciences (Biotechnology)

Under the supervision of: Prof. Rania Siam

American University in Cairo

May / 2021

Insights into halophilic microbial adaptation: Analysis of integrons and associated genomic structures and characterization of a nitrilase in hypersaline environments

A Thesis Submitted by

Sarah Ali Ahmed Sonbol

to the

Applied Sciences Graduate Program (Biotechnology)

Has been approved by

Dr. Rania Siam

(Thesis committee Chair/Advisor) _____

Affiliation _____

Dr.

(Thesis committee Reader/Internal Examiner) _____

Affiliation _____

Dr.

(Thesis committee Reader/Internal Examiner) _____

Affiliation _____

Dr.

(Thesis committee Reader/External Examiner) _____

Affiliation _____

Dr.

(Thesis committee Reader/External Examiner) _____

Affiliation _____

Dr. _____

(Moderator) _____

Affiliation

Graduate Program Director **Date** **School Dean** **Date**

Dedication

To my beloved family

Words cannot express how grateful I am for your presence in my life, your endless support and your continuous encouragements

Acknowledgements

I am really thankful to Dr. Rania Siam for her support, guidance and supervision of this thesis. I am also grateful to Dr. Karim Seddik the Associate Dean of graduate studies for his support and help throughout many logistics. My appreciation towards Dr. Ari Ferreira for his critical revision and participation in writing the nitrilase manuscript. My deep thanks to Mr. Amged Oaf for his great efforts and help throughout my PhD especially in my lab work. Thanks to Mr. Ahmed El Hosseiny for his help in my bioinformatics work. I am also deeply indeted to Dr. Nahla Hussein and Dr. Laila Ziko for their great help and support, fruitful lab discussions, revising manuscripts, providing suggestions and feeedback and for their continuous encouragement. I would also like to express my thanks towards Dr. Rehab Abdullah for her support and help in the lab and in providing some of the used reagents. Mr. Saifullah Soliman was a great assistant in LCL library screening, thus I owe him my thanks. Moreover, *E. coli* B548 and C319 were kindly provided from Dr. D. Mazel lab at Pasteur's Institute, Paris, France. Special Thanks to Dr. Céline Loot in Dr. Mazel's lab for the useful insights and tips concerning integron recombination assays. I acknowledge Dr. H. El Saied and his lab members for providing the Aghormy Lake filtered water sample, guidance through PCR screening of metagenomic DNA, and providing 16S rRNA sequencing results. Acknowledgments to Dr. T. Soliman for performing the Qiime analysis on the 16S rRNA sequences. Special thanks to Mr. Osama Ali for his help in ordering needed reagents. I am really thankful to Dr. Ghada Moustafa, Dr. Ali El Behery, Dr. Emanbellah Ramadan and Dr. Mohamed Maged for their help, advice and useful discussions at the beginning of my PhD. I am also indeted to Salma El Shafie, Shaimaa Farag and Marina Nabil for their help, support and encouragement. I would like to thank all biology department and biotechnology graduate program professors those whom I have TAed with or have taken courses with because I have learned a lot from them all. In addition, I am really thankful to the the biology department technical lab aides Mr. Zien and Mohamed El-Bagoury for their help during lab work.

Finally, this work could have never been done without th generous PhD fellowship that I have received from Youssef Jameel foundation. Thank you very much. I would also like to thank the AUC for funding my practical lab work by the AUC student research grant and for the AUC conference grant which allowed me to participate in ISME17 in Leipzig, Germany.

Abstract

Hypersaline environments are unique habitats in which residing microorganisms show distinctive adaptive measures that allow their survival under these stress conditions. Hence, halophiles are considered a unique resource of enzymes and gene cassettes with characteristics that could be exploited in different biotechnological applications. Here, due to the attention directed towards finding more stable nitrilases that has a great potential in green bioremediation processes and in different industries; we have biochemically characterized a nitrilase (NitraS-ATII) that we have previously isolated from the hypersaline and thermophilic Atlantis II Deep brine pool lower convective layer (LCL). Nitrilases can hydrolyze nitriles in a one-step reaction into their corresponding carboxylic acid and ammonia. NitraS-ATII showed higher thermal stability compared to a closely related nitrilase, and tolerance towards high concentrations of some heavy metals.

We have also focused on analyzing integrons and their associated genetic structures in hypersaline environments because of their presumed adaptive role. Integrons are genetic platforms in which an integron integrase (IntI) mediates the excision and integration of gene cassettes within the integron at specific recombination sites.

We constructed a fosmid library from the metagenome of hypersaline Aghormy Lake in Siwa Oasis. This library and the library of the Atlantis II Deep brine pool lower convective layer (LCL) were screened using pre-designed degenerate primers to amplify *int*I genes. However, we only detected two positive clones in the Aghormy lake library from which one (AGH-1G10) was further sequenced and analyzed and its integron components were all detected. The AGH-1G10 *int*I in addition to another identified IntI from Kebrit Deep brine pool Upper interface (KD UIN360) were synthesised and expressed in pBAD18 plasmid to test their *in vivo* excision activity. However, no activity was observed for both proteins. This could be attributed to using a deletion assay in which the used recombination sites cannot be identified by newly identified IntIs.

In addition, we used IntegronFinder software to analyze 80 halophilic bacterial genomes and 141 halophilic archaeal genomes. Our results revealed the presence of 19 new complete integrons and 44 clusters of *att*C sites lacking a neighboring integron-integrase (CALINs) in bacterial genomes and 1 complete integron in an archaeal genome. We also analyzed 28 hypersaline metagenomic assemblies in which we have identified eight complete integrons, 18 solitary integron integrases and 92 CALINs. Toxin-antitoxin (TA) gene cassettes were abundant in most detected integrons and CALINs, regardless of the length of the gene cassette array. Moreover, as expected, we have found different classes of insertion sequences (ISs) within and nearby integrons and CALINs. Surprisingly, this was only observed within analyzed genomes rather than assembled metagenomes which could be due to frequent concurrence of transposable elements' repetitive sequences with the peripheries of contigs. Some IS types were

more frequent than others such as IS*1182* elements and different ISs that are presumably able to mobilize adjacent genetic structures in presence of one copy of the IS element. Mining for group II introns revealed the presence of not only group IIC-*att*C, previously found embedded within different studied integrons, but also full and truncated group IIB introns (UHB.I2, H.ha.F1 and H.ha.F2) in CALINs within the extreme halophile *Halorhodospira halochloris* and a hypersaline metagenome. In addition, we have observed a relative abundance in arginine repressor (ArgR) binding sites within or overlapping with IntI promoters (P_{intI}) raising questions about possible regulation of IntI expression and recombination activity by these proteins.

Despite the reported absence of integrons in archaea, our search in halophilic archaeal genomes revealed the presence of an archaeal integron within a recently sequenced Natrialbaceae archaeon. Further investigation revealed the presence of other archaeal integrons within a thermophilic Euryarchaeota. The high similarity of the archaeal IntI to another bacterial IntI from a hypersaline environment would indicate its possible horizontal acquisition. Moreover, we detected atypical putative CALINs within archaeal metagenomes, showing arrays of successive *att*C-sites overlapping with archaeal ORFs.

Finally, the importance of assessing the prokaryotic diversity in studied sites led to our 16S rRNA-based analysis of the athalassohaline Aghormy Lake and comparing it to that of the thalassohaline Sebeaka saltern at the vicinity of Bardawil Lagoon (north coast of Sinai Peninsula). Aghormy Lake OTUs were assigned to 16 phyla, whereas, OTUs in Sebeaka saltern were assigned to 10 phyla. Both sites showed an abundance of Bacteroidetes, particularly family Rhodothermaceae. Aghormy Lake was characterized by phylotypes belonging to Deinococcus-Thermus, Spirochaetes, *Rhodovibrio* (Alphaproteobacteria), Chromatiaceae (Gammaproteobacteria) and GMD14H09 (Deltaproteobacteria). Phylotypes assigned to AT12OctB3 (Bacteroidetes), Rhodobacteriaceae (Alphaproteobacteria), Ectothiorhodospiraceae and Xanthomonadaceae (Gammaproteobacteria) formed Sebeaka saltern bacterial community. Cyanobacterial genus *Cyanothece* was abundant in both brines. In spite of the presence of shared phyla in both brines, differences were observed in lower taxonomic ranks which may reflect the differences in the biogeographical nature and physicochemical parameters between the two brines. Moreover, different identified halophiles in both sites have a potential to be exploited in different industries.

Our study may shed light towards a possible interplay of integrons along with different associated MGEs in the adaptation of microbial species in hypersaline environments. It also points out towards possible exploitations of identified genes within these harsh environments in different biotechnological applications.

Table of contents

Glossary and Abbreviations

ARG Antibiotic resistance gene

ArgR Arginine repressor

ATII Atlantis II brine pool

bs Bottom strand

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

cAMP cyclic AMP

cDNA Complementary DNA

CL Chloroplast-like

CRP cAMP receptor protein

DAP 2,6-diaminopimelic acid

DR Direct repeat

dsDNA Double-stranded DNA

DTR DNA transfer replication

DTT Dithiothreitol

EBS Exon binding site

EHB Extrahelical base

En Endonuclease

EPS Exopolysaccharide

GI Genomic Island

HGT Horizontal gene transfer

HTH Helix-Turn-Helix

IBS Intron binding site

ICE Integrative and conjugative element

IEP Intron encoded Protein

IntI Integron integrase

IPTG Isopropyl β-D-1-thiogalactopyranoside

IR Inverted repeat

IS Insertion Sequence

ISCR Insertion sequence common region

KD UINF Kebrit Deep brine pool upper interface

LCL Lower Convective Layer

MGE Mobile genetic element

MITE Miniature inverted repeat transposable element

ML Mitochondrial-like

MPF Mating pair formation

OD_{600} Optical density at 600 nm

ORF Open reading frame

PMSF Phenylmethylesulfonyl fluorid

qRT-PCR Quantitative real time reverse transcriptase ploymerase chain reaction

RBS Ribosomal binding site

RHH Ribbon-Helix-Helix

RNA-seq RNA sequencing

RNP Ribonucleoprotein

RT Reverse transcriptase

SDS-PAGE Sodium dodecyl sulfate-Polyacrylamide gel electrophoresis

SI Super Integron

ssDNA Single-stranded DNA

TA Toxin-Antitoxin

TSD Target site duplication

UCS Unpaired central spacer

VTS Variable terminal structure

List of Tables

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

List of Supplementary Materials

Appendix A: Chapter 4 Supplementary tables

Appendix B: Chapter 5 Supplementary tables

Appendix C: Chapter 6 Supplementary tables

Appendix D: Chapter 6 Supplementary figures

Chapter 1: Literature Review & Study Objectives

1.1. Microbial adaptation to hypersaline environments

Hypersaline aquatic environments are interesting unique extreme habitats. Microorganisms residing in there show different measures to adapt to the hypersalinity and other stress conditions encountered at these environments. For instance, modifications in the membrane lipid composition of halophiles were observed [1]. A relative increase in saturated and cyclopropanoic fatty acids and acidic phospholipids have been shown in *Pseudomonas halosaccharolytica* upon the increase in temperature and salinity [2]. Commonly, polar lipids are present in a higher frequency than non-polar lipids in the membranes of halophiles [2].

In general, halophilic microorganisms adopt one of two strategies for adaptation to hypersalinity [3]. The first and the most common one is a “salt-out” strategy in which the microorganism expels extra ions and accumulate organic osmolytes –named compatible solutes- intracellularly. Compatible solutes are polar, water-soluble, low molecular weight organic compounds that can accumulate to high concentrations inside the cells. They are uncharged at physiological pH and do not disturb the cellular metabolism or protein folding. Compatible solutes are either synthesized *de novo* or accumulated from the external environment [1,4]. Their intracellular accumulation level is determined based on the osmolarity of the surrounding environment [1]. Compatible solutes were found also to protect membrane integrity, and protein folding at different stress conditions such as freezing, heating and high ionic concentrations [1]. Glycine betaine, ectoine and proline are few examples of compatible solutes [1].

The other strategy used by extreme halophiles such as Halobacteriaceae archaea and *Salinibacter ruber* is a “salt-in” strategy [1]. This strategy is based on accumulating high concentrations of KCl intracellularly and expelling Na^+ ions to the outside to keep an osmotic balance. This requires the whole enzymatic machinery, and not only excreted enzymes, to adapt to the high ionic intracellular concentration [1]. Halophilic proteins are characterized by a significant increase of negatively charged acidic residues on the protein surface allowing their interaction with a network of hydrated cations in the medium [5]. In addition, halophilic proteins are characterized by a lower frequency of Lys, Cys and strong hydrophobic residues. They also tend to form more flexible coil-structured regions rather than helical structures [5].

Diferent halophiles are exploited in diverse biotechnological applications. One of the advantages of using halophilic enzymes in food industries is that they are not inhibited by high salt concentrations; at the same time, non-halophilic microorganisms cannot survive these conditions, thus limiting food contamination [4]. Moreover, different halophilic enzymes have been

used in food and pharmaceutical industries in addition to their potential use in treatment of saline wastewater and other bioremediation processes [4]. Halophiles were also utilized in the production of compatible solutes such as ectoine from *Halomonas elongate* and polymers such as exopolysaccharides. These products are extensively used in food and cosmetic industries [4].

1.2. Selected unique hypersaline environments with biotechnological promises

Different studies on the Red Sea discovered the presence of 25 unique brine pools in its depth [6]. Those are geothermal underwater lakes of high salinity that are found in depressions in the seafloor of the central and northern Red Sea [7]. One of the most extensively studied brine pools in the Red Sea is Atlantis II Deep. It is a hydrothermal ore-deposit at which the seafloor hydrothermal activity is associated with deposition of minerals on the seabed [8]. The brine has a maximum depth of about 2200 m [9]. It is stratified into four layers with the lowest (Lower Convective Layer -LCL) characterized by its harshest conditions: high salinity (250 ppt), high pressure, high temperature (68°C) , anoxic conditions, acidic pH and high heavy metal content [6] [7,10]. Discovery Deep is another Red Sea brine pool with a temperature of ~ 45°C and salinity of 100 ppt [11]. A third unique brine pool in the Red Sea is Kebrit Deep which is characterized by its high H_2S concentrations [12]. The high density of the brine waters separate each brine from its upper water column by a distinctive interface layer [9].

Other unique hypersaline systems that received little attention by researchers are salt lakes in Siwa Oasis. Siwa Oasis is located in the Northwest of the Egyptian Western desert. The deepest areas of the oasis are occupied by salt lakes surrounded by salt marches. These lakes are the natural discharge areas for water coming from the abundant artesian wells, springs and cultivated areas in the Oasis [13]. Aghormy Lake is one of these lakes, which has a 0.5 m depth and is located 18 m below sea level. It is characterized by total dissolved solids (TDS) of 220.03 g l^{-1} (ppt) and a pH of 7.83 [14].

Salterns of the hypersaline Bardawil lagoon are also of particular interest. Hypersaline lagoons are seawater bodies connected with the sea with salinities higher than 40% due to excessive water evaporation [15]. Situated at North Sinai [16] with an area of about 600-650 km^2 and a maximal depth of 3 m [15], Bardawil Lagoon is characterized by being oligotrophic and hypersaline lake [16] with salinity ranges from 39.5 -68.5 ppt according to the season and location within the lagoon [17]. This high salinity is due to the evaporation of Mediterranean seawater without having any other non-marine water source except for the scarce rain water [15] [16]. Themperatures there range from 21-30°C according to the season, and the pH ranges from 8.22-8.5 [18]. Salt flats covering the southern and eastern parts of the lagoon are described as “Sabkhas” [17]. Two types of Sabkhas can be encountered at the vicinity of the Lagoon; coastal sabkhas that are connected to the lagoon and inland sabkhas that are separated from it by sand

dunes [19]. An example of coastal sabkhas is Sebeaka saltern at the eastern part of Bardawil Lagoon [17,20]. The arid conditions of the area along with the availability of hypersaline water facilitate the formation of permanent halite thick crusts [21]. Thus, Sebeaka saltern is utilized in commercial salt production [17,20].

The hypersalinity of these environments combined with other unique characteristics of each site makes them promising candidates for identification of different enzymes with unique characteristics that allow their exploitation in different biotechnological applications. For instance, the Atlantis II Deep LCL was a source of unique halophilic, thermophilic and heavy metal-tolerant mercuric reductase [22] and esterase [23], in addition to thermostable antibiotic resistance enzymes [24].

1.3. Nitrilases: enzymes with potential biotechnological applications

An interesting group of enzymes with a biotechnological potential are nitrilases. Nitrilases can hydrolyze nitriles (cyanide containing compounds) in one step into their corresponding carboxylic acid and ammonia [25].

Most isolated nitrilases are inducible with different substrate specificities [26]. Moreover, they have a high stereo-selectivity which is exploited in the synthesis of specific isomers without production of toxic byproducts such as HCN gas [25,27]. In general, nitrilases have proven to be superior to conventional chemical methods in different pharmaceutical and chemical industries [28], and in environmentally-friendly bioremediation processes such as in the detoxification of cyanide containing wastes and herbicides [25]. However, the use of nitrilase-producing microorganisms as catalysts probably results in insufficient amounts of the enzyme with low reaction rates [26]. The instability of most nitrilases is another issue that limits their use [26]. Thus, genetic manipulation of unstable nitrilases in order to increase their stability in different extreme conditions seems promising [26]. Another alternative approach is to mine for nitrilases with higher stability profiles from extreme environments [26], as its more likely that enzymes isolated from hypersaline or thermophilic environments can withstand the harsh conditions at which they naturally reside in.

Some nitrilases were found to be encoded by a *nit*C gene and its homologues in a conserved gene cluster Nit1C that was identified in different species [29]. The gene cluster was also identified on a virulent plasmid suggesting its lateral transfer [29]. This cluster was found to be involved in cyanide and nitrile assimilation pathway [30,31] . In two reports, we have found that nitrilase gene cassettes were identified in class 1 integrons isolated from carriage water of ornamental fish [32] and from *Comamonas* sp. isolated from wastewater [33]. Detailed description of integrons along with their possible horizontal transfer will be discussed in next subsections.

1.4. Site-specific recombination reactions

1.4.1. Definition of site-specific recombination

A site-specific recombination reaction is a process in which DNA segments are exchanged at specific sites after breaking and re-joining resulting in an integration, deletion or an inversion event [34,35]. This process is considered “conservative”, as no loss or gain of sequence information occurs during the process, such as the formation of target-site duplication (TSD) known in transposition reactions [35]. A site-specific recombination reaction requires the presence of two DNA substrates, a site-specific recombinase that catalyses the reaction and a mechanism at which the phosphodiester bond energy is conserved [34] . An integration reaction occurs when the two recombination sites are located on two different DNA molecules in which one at least is circular [34]. On the other hand, if the two recombination sites are located on the same DNA molecule, an excision or an inversion occurs based on the orientation of the recombination sites. Directly repeated recombination sites result in a deletion, whereas inverted sites would result in an inversion [34,35]. Site-specific recombinases can either catalyse unidirectional (irreversible) recombination reactions between two different recombination sites, or bidirectional (reversible) reactions between identical recombination site. Exceptions to this classification were observed as well [36].

1.4.2. Mechanism of site-specific recombination

The recombination mechanism can be summarized as follows: two monomers of the site-specific recombinase (a dimer) bind to two binding sites of the recombination site at each DNA substrate, thus the process involves a tetramer of the catalytic recombinase. When the 2 DNA substrates along with the enzyme tetramer are brought into close proximity, a synaptic complex is formed. A nucleophilic attack by the OH of the recombinase catalytic residue on the DNA phosphate group of the DNA sugar-phosphate backbone cleaves the DNA. This allows an exchange of DNA segments followed by the dissociation of the synaptic complex, and the formation of new recombined segments [35].

1.4.3. Types of site-specific recombinases

Site-specific recombinases fall into two broad families: tyrosine recombinases and serine recombinases. This classification is based on whether the recombinase uses a tyrosine or a serine as a catalytic residue [34,35]. Although in both types a nucleophilic attack on the sugar-phosphate backbone takes place, the location of the formed protein-DNA bond differs [34] [35]. Serine recombinases form a 5’-phosphoserine bond with the DNA, while tyrosine recombinases form a 3’-phosphotyrosyl bond [34,35]. Another difference is that in case of serine recombinases all four monomers function simultaneously; breaking all DNA strands at once before an exchange of strands takes place [34,35]. On the other hand, two tyrosine recombinase monomers introduce one strand break in each DNA duplex, followed by a strand exchange forming a Holliday junction

structure. Isomerization of the recombinase tetramer converts the active monomers inactive and vice versa. The newly active monomers introduce DNA breaks in the other strands followed by a second strand exchange resolving the formed Holliday junction [34].

Tyrosine recombinases in general possess four conserved residues: R in box I motif and HRY in box II motif, in addition to three smaller motifs named patches I, II and III [37,38]. Integron integrases (IntIs) are a specific type of tyrosine recombinases with unique characteristics that sets them aside from other tyrosine recombinases [35]. They will be discussed in more details below.

1.5. Integrons

1.5.1. What are integrons?

Integrons are genetic elements where different open reading frames (ORFs) are captured and expressed according to the need of the microorganism [39]. They were first reported in 1989 as potential mobile genetic elements (MGEs) associated with antibiotic resistance genes (ARGs) [40]. They were initially connected to pathogenic Gram negative bacteria as a result of their apparent dissemination in clinical isolates; however, further discoveries showed their spread among different bacterial phyla in many environments harboring diverse gene cassettes [41].

An integron is composed of a functional platform containing all required elements for system operation and an array of gene cassettes (Fig.1.1). The functional platform is composed of : (1) *Int*I gene which encodes a site-specific tyrosine recombinase (IntI integron integrase) with its own promoter P_{intI} (2) a recombination site termed *att*I primary recombination site and (3) a promotor (P_C) for transcription of the associated gene cassettes as the vast majority of gene cassettes are promoterless [39]. A gene cassette is an independently mobilizable genetic element typically formed of an ORF followed by an *att*C recombination site (formerly named 59-base element [40]) recognized by the IntI. However, ORF-less gene cassettes and cassettes with more than 1 ORF were also observed [41]. The number of gene cassettes associated with an integron could vary from zero to more than 200 cassettes such as those observed in *Vibrio* spp. chromosomal super-integrons (SIs) [39].

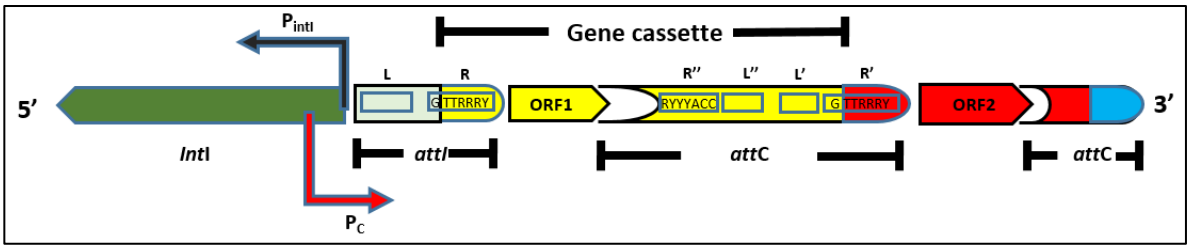


Fig.1.1. **Integron components.** An integron is composed of an integron integrase gene (*int*I) with its promoter (P_{intI}), *att*I recombination site, P_C promoter for cassettes transcription, and an array of gene cassettes. Each gene cassette is composed of an ORF followed by an *att*C site. An *att*I site is mainly composed of 2 simple binding sites R and L, and an *att*C site is composed of R", L", L' and R' binding sites at which L" and L' are separated with a region of variable length

Although some gene cassette ORFs carry their own promoters [39,41], the majority are promoterless [39,42]. Hence, expression of gene cassettes is driven by P_C promoter located commonly within *Int*I gene or within *att*I site. Different variants of P_C promoters were identified within those of class 1 integrons being the most extensively studied [39,41]. Studies have shown that as the strength of the P_C promoter decreases, the excision activity of the IntI increases [43]. Moreover, it was found that the expression levels decrease as gene cassettes become more distal from the P_C promoter [39,41]. This could be attributed to failure of the ribosome to progress through the gene cassettes transcript due to the formed stem-loops by the *att*C mRNA [44]. It has been shown that destabilization of the secondary structures formed by the *att*C site transcript by the presence of a translated ORF within the *att*C site increases the expression levels of the downstream genes through translation coupling [44]. In general, translation rate of ORFs in gene cassettes is affected by the presence of an upstream ribosomal binding site (RBS). Translation of genes lacking RBSs can be initiated from a RBS of an upstream gene proceeding towards the downstream gene as if they are parts of an operon [45].

1.5.1.1 Integron Integrases

Integron integrases (IntIs) are members of site-specific tyrosine recombinases. Their closest relatives in the tyrosine-recombinase superfamily are XerC and XerD [38,46]. They possess all conserved regions characteristic for tyrosine recombinases : boxes I and II and patches I, II and III [38,46]. However, IntIs are characterized by the presence of an extra motif (named IntI patch [46]) between patch II and patch III that is absent from all other tyrosine recombinases [38,46]. The α-helix in the protein (termed α-I2) within this motif is important for synapse formation in IntI-mediated recombination reactions [47].

Integrons are classified based on the similarities between their IntI proteins. IntIs with greater than 98% identity are considered from the same class [48]. Numbers were granted for

early discovered integrons and classes 1, 2 and 3 got a lot of attention due to their association with transposable elements and their role in antibiotic dissemination [39]. Later studies showed that more than 100 IntIs have been identified [49]. Despite the huge number of detected IntIs, most experiments use class 1 integrons as a model for integrons [39].

1.5.1.2 Recombination sites

The core site of *att*I recombination site is minimally composed of two IntI binding sites termed R and L that form imperfect inverted repeats, where R has the consensus sequence of 5'-GTTRRRY-3', while the L site is highly degenerate (Fig.1.1). Recombination occurs between G and TT in the conserved triplet GTT within the R site. In class 1 integrons, two direct repeat binding sites (DR1 and DR2) were also detected upstream the core *att*I site [39,41]. IntI can recognize its cognate *att*I site; however, identification of *att*I sites from other integron classes was observed but with much lower efficiency [39]. The detection of *att*I sites for different integron classes is difficult because of their divergence and the degenerate nature of the L site [45].

On the other hand, the structure of the *att*C site is more complex when compared to the *att*I site. It is composed of four binding domains R", L", L' and R', where L" and L' are separated by a central region that varies greatly in sequence and length. The only conserved domains are the R" and R' sites with the consensus of 5'-RYYYACC-3' and 5'-GTTRRRY-3', respectively [39] [41] (Fig.1.1). Although ORFs within gene cassettes typically end before or within R" [46,48], they may extend further into the *att*C site or continue until they terminate before the next *att*C in the array [48]. The lack of conservation among *att*C sites renders their identification challenging [50]. This raised questions about the mechanism of recognition of different *att*C sites by the same IntI. Crystallization of VchIntIA with its attached *att*C site revealed that *att*C site interacts with IntI by its bottom strand (bs) only after the formation of a hairpin loop secondary structure at which R'' binds to R' and L'' binds to L' forming R and L boxes, respectively [47]. Two flipped-out bases at positions 20" and 12" on the R"- L'' arm of the bs, referred to as extrahelical bases (EHBs) [47], orient the polarity of the recombination reaction by identifying the recombinogenic strand (bs) [39] [47] . Some *att*C sites have a third EHBs [48]. IntIs have different preferences for their EHBs [51], but in most cases the 20" base is a G and the 12" base is a T [48]. The unpaired central spacer (UCS) between R and L boxes in the *att*C bs has an essential role in stabilizing the formed synapse during recombination [39]. On the other hand, the variable terminal structure (VTS) formed by the remainder *att*C bs sequence and shows great variations in length and structure among different *att*C sites, is thought to have an important role in the modulation of *att*C folding when it extrudes from double-stranded DNA (dsDNA) to form a cruciform structure as this event is favored by *att*C sites with short VTS [52].

In general, IntIs from different classes with less than 50% identity can recognize the same *att*C sites [53]. However, some IntIs such as IntI1 have a broader substrate specificity identifying

more *att*C sites compared to other IntIs [45,53]. Thus, the *att*C site secondary structure appears to be of greater importance than its primary sequence.

1.5.2. IntI-mediated recombination reactions and their mechanism

Integration and excision of gene cassettes are catalyzed by IntI protein (Fig.1.2). Site-specific recombination between an integron *att*I site and an *att*C site within a free circular gene cassette (*att*I X *att*C recombination) results in the integration of a gene cassette and its positioning as the first cassette within an integron [39]. This results in the formation of a chimeric *att*I/*att*C site at one end and a chimeric *att*C site on the other side of the integrated cassette [50]. In contrast, intermolecular recombination between two *att*C sites within the same integron leads to gene cassette excision. Recombination between two *att*I sites (*att*I X *att*I) has been observed, but it was less efficient [39]. Finally, recombination into secondary sites having a GTT triplet might occur as well [39,41]. Recombination occurs between G and TT in the conserved triplet GTT within the R site in *att*I site and the R' within the *att*C site [54]. To be more precise the cleaved strand would be the opposite strand between the A and the C in the conserved AAC triplet [55]. In addition to single gene cassette excision reactions, rare events of excision of large gene cassette arrays could happen. In one study, an excision of a 38-cassette array has been observed [56].

As discussed earlier, in a tyrosine recombinase-mediated recombination, two DNA substrates and four protein monomers form a synaptic complex at which two sequential strand-exchange events take place [34]. As IntI-mediated recombination reactions involve a single stranded *att*C bs, a second strand exchange would result in abortive products [39]. Thus, in IntI-mediated reactions, a single strand-exchange takes place and the formed Holliday junction is resolved by a replication step [57]. Thus, the process is semiconservative [39]. Although four IntI monomers are bound to the synaptic complex, only two act as attacking subunits in which their α-I2 helices form contacts with the 20" G EHB on the *att*C bs [47,53]. In contrast, the two non-attacking subunits interact with the 12" T EHB resulting in conformational changes; pulling the

catalytic tyrosines away from the phosphate groups, thus preventing a second nucleophilic attack followed by a second deleterious strand exchange [39,47].

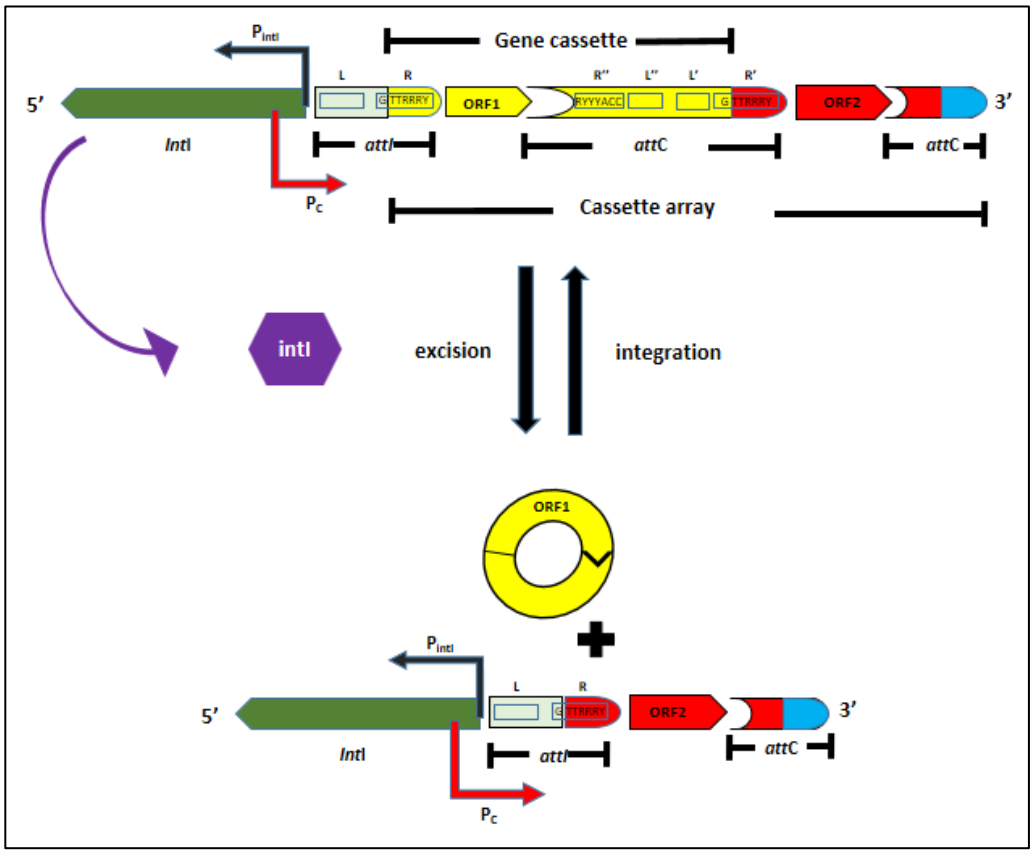


Fig.1.2 **IntI-mediated recombination reactions.** Gene cassettes can exist transiently in a circular form where it can be integrated within *att*I site of the integron. The process is reversible where a gene cassette can be excised out of the integron

1.5.3. Distribution of integrons

Integrons were first classified as either mobile integrons associated with plasmids and transposons, or as chromosomal integrons that are found to be widespread among many bacterial phyla. However, further discoveries showed that intermediate forms between these two extremes do exist [39,41]. Mobile integrons were classified into 5 classes based on sequence homology of their *Int*I genes. All 5 classes were associated with antibiotic resistance gene cassettes [39]. However, this could have been due to their first identification in clinical isolates [58]. Class 1 integrons harboring ARGs in particular are widespread in human and farm-animals commensal bacteria [41]. They have been even isolated from environments with low anthropogenic impact [59]. However, class 1 integrons were identified in different environments with gene cassettes unrelated to known ARGs [60]. Integrons can be found in almost any environment. They were isolated from desert soil, forest soil, hot springs, estuaries [41], polar sediment [61], glaciers [62] and marine environments [63].

Bioinformatics analysis on complete bacterial genomes, revealed that integrons are distributed through different bacterial phyla. However, they were completely absent in α-proteobacteria, Actinobacteria, Tenericutes and Chlamydiae [50]. Most identified integrons in that study were chromosomal integrons rather than integrons carried on plasmids [50]. As the number of identified integrons is increasing tremendously; the number of detected gene cassettes is growing as well. However, to the moment the majority of identified gene cassettes encode for hypothetical proteins of unknown functions [41]. Unfortunately, this limits our ability to exploit these gene cassettes to a great extent. Nonetheless, dissemination of ARG cassettes in clinical isolates and gene cassettes that are likely involved in the degradation of compounds of industrial wastes in environments heavily affected by industrial pollution [64] strengthens their suggested role in bacterial adaptation.

1.5.4. Toxin-Antitoxin systems widely distributed in integrons

Toxin-Antitoxin (TA) systems are addiction systems that encode a stable protein toxin and its cognate unstable antitoxin [65]. The rapid decay of the antitoxin leads to a biocidal or most probably a biostatic condition of the cells if the genes encoding the TA system are lost [66]. Based on the type of the antitoxin and its mechanism of interaction with its cognate toxin, prokaryotic TA systems can be divided into three main types [67]. In type I, the toxin mRNA translation can be inhibited by an antisense RNA antitoxin. In type II, both the toxin and the antitoxin interact at the protein level at which the antitoxin neutralizes the toxin. Type III antitoxins are small RNAs that can bind to and neutralize their cognate toxins [67].

One of the major systems extensively associated with integrons are type II TA systems [39]. Typically, these systems are arranged in operons in which the antitoxin gene is found upstream of the toxin gene with few exception such as in higBA TA module [66,67]. Coupled translation is observed as toxin and antitoxin genes are usually overlapped by few nucleotides [68].

Toxins of type II TA systems have different cellular targets. Toxins can affect translation by different mechanisms, but most toxins function as endoribonucleases. In addition, toxins could affect replication by inhibition of DNA gyrase, thus inducing DNA double strand breaks, which activates an SOS response, eventually leading to programmed cell death [66,67].

Many TA systems are now identified using bioinformatics [67]. For instance, VapC toxins are characterized by their PIN domain Mg^{2+} dependent RNAse activity. At the same time, many prokaryotic PIN-domain containing genes were found to lie downstream of genes believed to encode for transcription factors that are arranged as operons. It is thus predicted that these loci form TA systems, as each antitoxin is composed of two domains: one interacts with its cognate toxin while the other is a DNA binding domain for its autoregulatory function. The most commonly

found DNA binding domains in type II TA antitoxins are Ribbon-Helix-Helix (RHH) and Helix-Turn-Helix (HTH) domains [67].

TA systems exist in both plasmids and chromosomes. In plasmids they function as addiction modules stabilizing plasmids through post segregational killing and exclusion of co-existent plasmids from the same incompatibility group [66]. On the other hand, they seem to have diverse functions within chromosomes. Their abundance in chromosomal integrons, more specifically Super-Integrons (SI)s, led to the identification of their role in stabilizing these SIs [69,70]. The presence of TA cassettes minimizes the possibility of large cassette excision events, since deletion of TA cassettes would lead to cell death by the stable toxins [70,71]

In contrast to the majority of promoterless integron gene cassettes, TA operon gene cassettes are found within integrons with their own promoters [39]. In addition, they could be found in an opposite orientation to adjacent gene cassettes [72]. TA systems can also protect against invading phages [66,73] and plasmids [66], regulate biofilm formation [66], act as global regulators such as in their post-transcriptional regulation of sugar uptake and metabolism [67] and finally have a role in the formation of persisters upon stress conditions [66,67].

TA systems have a great biotechnological potential as well. They can be used as selection markers instead of antibiotics. They can also be used in selective cell killing in multicellular organisms, as some TA pairs were functional in eukaryotic cells, and finally as antiviral therapies by cleaving single-stranded RNA viruses [67].

1.5.5. Identification of integrons

Cultured microorganisms are the major source for studying integrons [46,59,74]. However, metagenomics could be a great mine for expanding our knowledge about integrons [59,61,75,76]. Screening for integrons in most studies is based on PRC amplification either to amplify integron integrase genes [42,61,77], their cassettes [78] or both [59,75]. Degenerate primers have been used to amplify *int*I genes or known gene cassettes [46,78]. Furthermore, different bioinformatics tools were developed for the identification of integrons; however, these early-developed programmes were restricted to few integron classes [79,80,81]. The high diversity in *att*C sites was a problem hindering the identification of new gene cassettes [50]. The development of IntegronFinder program [50] was a leap forward in the identification of integrons. The highly sensitive and specific pipeline can identify novel IntIs based on an HMM profile and *att*C sites based on a covariance model [50]. It also annotates *att*I sites, P_{intI} and P_C promoters for integron classes 1, 2 and 3 [50]. Developers of IntegronFinder showed that clusters of *att*C sites lacking a neighboring integron integrase (CALINs) are abundant in bacterial genomes [50].

1.5.6. IntI-mediated recombination assays

Binding of IntI protein to different recombination sites was assayed in different studies. Electrophoretic mobility shift assays were used to measure the binding of IntI proteins to different DNA fragments carrying *att*I sites [54] or both *att*I and *att*C recombination sites [55,82].

Other assays were developed to measure the recombination activity. The developed assays could either test the integration or the excision efficiency of the IntI. For instance, different conjugation assays have been developed for measuring IntI recombination activities. In these assays, a single-stranded DNA (ssDNA) can be transferred from a donor to a recipient cell mimicking the natural horizontal gene transfer (HGT) of gene cassettes via conjugation. This is particularly relevant with mobile integrons carried on conjugative plasmids [83]. In an integration conjugation assay, a pir^+ donor cell -encoding for a π protein with a π-dependant conjugative plasmid carrying an *att*C site (pAttC)- can be transferred by conjugation to a recipient cell devoid of a π protein (pir^-), thus cannot sustain the pAttC replication. However, the recipient cell carries a plasmid with an *att*I site and expresses an IntI protein. The only way for the transformed pAttC to be maintained is by integration into the *att*I site forming a co-integrate. Nevertheless, in some conjugation assays, co-integrates were retrieved in absence of IntI. This was found to be due to unexpected homologous recombination events between identical regions [54]. Recombination events can then be selected using the antibiotic marker carried on the pAttC plasmid [82]. In order to measure the excision efficiency using conjugation assays, a synthetic cassette surrounded by *att*C sites on a conjugative plasmid is transferred from a pir^+ donor cell to pir^- recipient cell expressing IntI. As the replication of the conjugative plasmid relies on the π protein encoded by the *pir* gene, it will not replicate inside the recipient cell unless the gene cassette is excised bringing together a promoter and a promoterless *pir* gene inside the conjugative plasmid. In this case the π protein will be expressed allowing the plasmid to be maintained within recipient cells [84]. Here again the excision events can be measured by selection using antibiotic marker on the conjugative plasmid [82].

Other developed assays were based on double transformation of cells with a plasmid carrying an *int*I gene and an *att*I site, and another plasmid carrying different gene cassette arrays. The expressed IntI would then excise one or more gene cassettes and integrate them into the *att*I site. Identification of integration and excision events would then be done by PCR amplification of extracted plasmids using proper primer sets and sequencing of amplicons [85]. The used gene cassette could harbor an antibiotic resistance gene (ARG). In this case, transformants can be screened for the loss of antibiotic resistance due to the excision of the corresponding cassette-encoded gene [54]. Plasmid extraction followed by PCR amplification using primers targeting cassettes-flanking regions was also used; as reduction in amplicons sizes would indicate a successful excision event [53].

As the majority of integrons are harbored on chromosomes rather than plasmids [50,83], chromosomal deletion assays were developed to measure the frequency of cassette excision in chromosomes. A plasmid harboring a *dap*A gene interrupted by a synthetic cassette is transduced and inserted into *att*B lambda site in *Escherichia coli* MG1655Δ*dap*A strain. This strain cannot synthesize 2, 6-diaminopimelic acid (DAP), thus the medium must be supplemented with DAP for their growth. Transformation of this strain with an IntI-expressing plasmid, leads to the excision of the gene cassette and restoration of a functional *dap*A gene, thus transformants can grow in absence of DAP [86].

1.5.7. Regulation of IntI expression and recombination reactions

Expression of IntI1 and IntIA (IntI4) was found to be under the control of SOS response [87]. Binding sites for LexA, a transcriptional repressor for the SOS response, were found overlapping the P_{intI} promotors [88]. SOS response can be induced by transformation of foreign DNA, conjugation [41] and antibiotic administration [39,41], thus upregulating IntI activity [39,41]. A study on class 1 integrons in biofilms has shown that the stringent response triggered by nutrient starvation led to an increase in IntI expression and mild induction of SOS response [89]. Thus, they have suggested a possible interplay between the SOS and stringent responses in biofilms inducing IntI1 expression [89]. Moreover, IntIA was found to be controlled by the carbon catabolite repression mechanism via cyclic AMP (cAMP)-receptor protein (CRP), independent of the SOS response regulation [90]. Finally, experimental evidence showed possible repression of IntI1 by nucleoid-associated proteins FIS and H-NS [91].

1.5.8. Biotechnological potential of integrons

As a unique system with the ability to acquire, rearrange and express exogenous genes, the integron system has a great potential as a platform for a variety of biotechnological applications. For instance, integron integrases can be used for the recovery of functional gene cassettes and their introduction into plasmids for further manipulations. Furthermore, integrons can be exploited in cloning techniques that do not depend on vectors or antibiotic markers [92]. Synthetic and natural gene cassettes can be transformed into bacterial cells to be incorporated within an existing integron. This could also be exploited for detection of cassette arrays in environmental bacteria by transforming these cells with marker cassettes such as a gene cassette for green fluorescent protein [41]. The inherent gene-shuffling activity of integrons can also be exploited in construction and optimization of different metabolic pathways for bioremediation or biosynthesis. This was successfully done with the tryptophan biosynthetic pathway in *E. coli*, yielding an 11-fold increase in tryptophan production by constructing the genes involved in tryptophan pathway in the form of gene cassettes and shuffling the order of these cassettes through IntI-mediated recombinations [93]. Moreover, as gene cassettes in environmental integrons are expected to encode for environmentally-adaptive proteins, these gene cassette arrays are considered as a

huge resource for the discovery of novel proteins [41]. Rowe-Magnus (2009) has engineered a tool based on IntI1 ability to recognize diverse *att*C sites to recover gene cassettes from different genomic libraries [94].

1.6. Mobile genetic elements (MGEs) and their role in integron dissemination

1.6.1. Different types of MGEs

Mobile genetic elements (MGEs) are DNA elements that mediate the mobilization of DNA segments intracellularly (within the same genome) or intercellularly (between different cells) [95] [96]. Intracellular mobilization of MGEs can be mediated through transposases and site-specific recombinases [95]. Although transposition mediated by transposases can occur at many different non-homologous genomic locations, the process is not really random. Specific target sequences were identified for some elements, in addition to the influence of the target DNA structure and supercoiling on the transposition process [97]. Unlike specific recombination which is a conservative process, transposition usually involves the formation of target site duplications (TSDs) to repair the formed gaps upon mobilization [35]. On the other hand, the intercellular movement of genetic material known as horizontal gene transfer (HGT) can be achieved via transformation, transduction or conjugation [95]. Transformation is the natural ability to uptake exogenous DNA from the surrounding environment. In contrast, transduction is the uptake of exogenous DNA through a bacteriophage. Finally, conjugation is the uptake of DNA from a cell to another through a conjugation or mating apparatus synthesized by the donor cell [98]. Here, we give a brief account on different MGEs, particulary those found to be associated with integrons.

1.6.1.1 Insertion Sequences (IS), Transposons (Tn) and related transposable elements

An insertion sequence (IS) is a mobile short DNA segment (0.7-3.5 kb) that encodes for a transposase and sometimes for other regulatory proteins as well. The transposase catalyzes the transposition of the IS and its insertion into different sites without need for DNA homology between the IS and its target. Most IS types are flanked by imperfect terminal inverted repeats (IRs) and some can generate target site duplications (TSD) upon insertion [99]. IS families can be classified into two major groups based on the type of their transposases into: DDE (and DEDD) transposases and HUH transposases [99]. Their names refer to their conserved amino acid motifs, and the “U” in HUH transposases refers to a large hydrophobic residue [99]. DDE transposases catalyze different transposition mechanisms. In the conservative or “cut and paste” transposition, the IS cleaves from its original site to be inserted in a new location. In replicative “copy and paste” transposition, a copy of the IS is produced by a replication step that fuses the donor and target DNA followed by resolution of the formed co-integrate by recombination of the

two IS copies. Finally, in “copy out-paste in” transposition, the IS is replicated and excised as a circular DNA before being inserted into its target site [99,100]. In contrast, HUH endonucleases catalyse transposition using an active site tyrosine residue that forms a transient covalent bond with its DNA substrate [101]. They encompass a large superfamily that includes: “Rep proteins” involved in plasmid and bacteriophages rolling-circle replication, “relaxases” involved in conjugative plasmid transfer and “transposases” involved in single strand transposition and presumed rolling-circle transposition [99,101]. In prokaryotes, HUH transposases are found in two IS families: IS*91* and IS*200/605* [99]. They lack IR sequences and do not create TSD [96].

Non-composite or simple transposons (Tns) are similar to ISs, in which they are surrounded by IRs and carry transposase genes essential for their transposition. However, in a transposon, other genes “passenger genes” that are unrelated to the transposition mechanism such as ARGs can be found as well [102]. Most non-composite transposons belong to Tn*3* family which includes subfamilies: Tn*7*, Tn*21*, Tn*501*, Tn*5393*, Tn*5403*, and Tn*1721* [102]. Tn*3* transposons are characterized by their long transposases and their movement via replicative transposition [100]. On the other hand, composite transposons are DNA segments with passenger genes unrelated to transposition, flanked by two copies of the same IS that allow their transposition by a cut and paste mechanism [100,102].

Unlike ISs and Tns, miniature inverted repeat transposable elements (MITEs) are non-autonomous IS derivatives that lack their own transposases, but can be mobilized *in trans* by transposases from related IS elements. They are short sequences of about 300 bp, flanked by IRs and usually generate TSDs [99].

In fact, the distinction between ISs, Tns and other related transposable elements is becoming blurred by time. The number of elements combining properties of ISs, Tns and other transposable elements is unceasingly increasing [99]. For instance, the ambiguous definition of a genomic island (GI) may encompass a large number of transposable elements. A GI is a relatively large DNA segment that is acquired horizontally and is usually flanked by DRs. They could carry genes that allow their mobilization such as transposases or genes related to a conjugation system [103]. Based on this definition, integrative and conjugative elements (ICEs) (discussed below) could be considered as GIs as well. In general, GIs are classified based on the type of their passenger genes. Those with ARGs are referred to as resistance islands and those with genes involved in virulence are named pathogenicity islands [96].

1.6.1.2 Plasmids and integrative and conjugative elements (ICEs)

A plasmid is an extrachromosomal DNA element that replicates independently of the bacterial chromosome. It carries genes essential for its replicative function and other accessory genes that encode for functions different than those encoded by the bacterial chromosome [95]. Conjugation function may exist in a plasmid forming a self-transmissible or a conjugative plasmid

that can be transferred horizontally [96]. A conjugative plasmid contains an origin of transfer *Ori*T and genes that encodes proteins for mating pair formation (MPF) and DNA transfer replication (DTR). Some non-conjugative plasmids that carry an *ori*T and a subset of DTR functions can be transferred horizontally. This can be achieved by utilizing the MPF apparatus (a specialized type IV secretion system) synthesised by a co-existing conjugative plasmid in the same cell [95].

Integrative and conjugative elements (ICEs) or conjugative transposons are MGEs that are integrated into the host chromosome [98], replicate as part of it, but can be excised and transferred via conjugation [95,98,104]. They carry their own modules that encode for conjugation machinery, integration/excision function and other regulatory functions encoded by different passenger genes that confer different phenotypes to the host cells [95,98,104]. Upon certain conditions, an ICE can excise out of the chromosome, circularizes, replicates and then transfers via its encoded conjugation machinery into a new host. The transferred copy of the ICE integrates into the recipient cell chromosome, whereas the remaining copy in the original host reintegrates into the chromosome [104]. Most ICEs integrate at the 3' ends of a tRNA gene [95,98], creating DRs flanking the ICE named as *att*L and *att*R [95].

1.6.1.3 Group I and group II introns

Group I and II introns are mobile catalytic RNA elements (ribozymes) that can self-splice themselves out of their mRNA transcripts. They can also integrate into homologous genomic locations (homing) or into new ectopic locations (ectopic transposition) by the aid of their intron encoded proteins (IEPs) [105].

Both types of introns are different in their distribution and structure [105]. Group I introns are distributed in bacteriophages, bacteria, organellar and nuclear eukaryotic genomes [105]. In contrast, group II introns can be found in bacteria, archaea, mitochondria and chloroplasts of lower eukaryotes and plants [105,106]. Group I introns are usually found within essential genes, whereas group II introns are mainly found within noncoding sequences [105]. Nonetheless, group II introns are usually found within MGEs such as plasmids, ISs, Tns and GIs [105]. Both group I and group II introns transcripts fold into conserved secondary and tertiary structures [107]. In group I introns, the secondary structure is composed of paired elements named P1-P10 [105], with a catalytic core formed by P3-P8 [107]. An ORF within the intron encodes for an endonuclease (En) that catalyzes the mobility of the intron [105]. On the other hand, group II introns transcripts form 6 double helical domains (DI-DVI) that radiate from a central wheel structure. The catalytic core is formed by DI and DV and an IEP is expressed from an ORF within DIV domain [108]. The IEP in group II introns can function as a reverse transcriptase (RT), a maturase and in some proteins as an endonuclease as well [108,109].

Splicing and mobilization mechanisms of both introns differ as well. Splicing happens through two transesterification steps in both; however, an external guanosine cofactor initiates the

nucleophilic attack on the intron 5' splice site in case of group I introns, while usually a bulged adenosine in DVI in a group II intron attacks the 5' splice junction forming a lariat structure (a circle with a tail) [105,106]. Homing of group I introns into other intronless alleles is a DNA-based mobilization mechanism that depends on homologous recombination between donor and recipient DNA strands and is catalyzed by the intron En [110]. On the other hand, retrohoming of group II introns is an RNA-mediated process catalyzed by the IEP associated with the spliced intron RNA forming a ribonucleoparticle (RNP). The intron RNA reverse-splices into ds- or ssDNA target site followed by a complementary cDNA strand synthesis [106].

1.6.2. Role of MGE in integrons dissemination

Association of integrons with MGEs has been documented especially with class 1 integrons being the most extensively studied class [111]. Identical integrons were identified within different bacterial species and in epidemiologically unrelated species [112,113,114] indicating the possible horizontal acquisition of these elements.

IS elements are commonly found adjacent to class 1 integrons [48]. Several studies have reported class 1 integrons flanked by IS*26* elements (IS*6* family) or flanked by ISs that are embedded within other Tns [111,115,116]. Insertion sequence common regions (ISCR) are always associated with complex class 1 integrons [117,118]. ISCR is a unique IS group that belongs to IS*91* family. Thus, as an IS*91*, it lacks IR ends and it starts with an *Ori*IS region and terminates by a *ter*IS region and presumably moves by a rolling-circle mechanism. This may allow the mobilization of adjacent DNA segments by just one copy of the ISCR [117]. IS*Ecp1* and similar ISs from IS1*380* family were also found to be able to mobilize adjacent ARGs in class 1 integrons by an unknown mechanism [96].

In general, class 1 integrons are associated with different Tns from Tn*3* family [71]. They can be found within a functional or defective Tn*402*, that could be itself embedded within a Tn*21* [39]. Although most isolated class 2 integrons contain a non-functional IntI due to a nonsense mutation, they are widespread in clinical isolates [119]. Most isolated class 2 integrons were associated with Tn*7*, a transposon that contains 5 transposition genes [71]. A class 3 integron form *Serratia marcescens* was found to be associated with a Tn*402*-like transposon [120]. Furthermore, class 5 integron was found within a composite Tn on pRSV1 plasmid in *Alivibrio salmonicida* [39,121].

Moreover, class 1 integrons in *Acinetobacter baumannii* and *Acinetobacter johnsonii* were flanked by MITEs [122,123]. Another defective class1 integron in a plasmid within *Enterobacter cloacae*, was found to be flanked by MITE-like structure that lacks TSD. However, when a transposase was provided *in trans*, it mobilized the entire structure and TSD were generated at their target site [124]. Furthermore, the presence of class 1 integrons within GIs in different

pathogens have been documented. Class 1 integrons were described within resistance islands found in *Salmonella enterica, Acinetobacter baumannii, Shigella flexneri, Proteus mirabilis* and *Pseudomonas aeruginosa* [111].

Plasmids harboring mobile integrons in different bacterial species were extensively reported [111]; however, later surveys found that the number of chromosomal integrons is much greater [50]. Different studies showed possible transfer of class 1 integrons carried on plasmids by conjugation [125,126]. Class 2 integrons could be disseminated via conjugation due to their association with Tn*7* that have shown an ability to be inserted into conjugative plasmids [71]. In addition, class 4 integrons were identified within an ICE in *Vibrio cholerae* [39,127].

Linkage of integrons to MGEs was not limited to ISs, Tns, plasmids and related structures. A class of group II introns, more specifically group IIC-*att*C , was found embedded after or within the *att*C site bs in an opposite orientation to adjacent gene cassettes [128,129]. It has been suggested that group IIC may have a role in the formation of gene cassettes and their assembly [128]; however, this hypothesis has been criticized [45].

Different methods of HGT along with different intracellular transposition mechanisms and homologous recombination may all contribute to the transfer of integrons among different bacterial species [111]. As shown earlier, integrons carried on conjugative plasmids and ICEs can be disseminated via conjugation [125,126]. Natural transformation is another method for dissemination of integrons. Natural transformation of class 1 integrons into *Acinitobacter baylyi* has been demonstrated, followed by the insertion of the integron via transposition stimulated by IS*26* elements, Tn*21*-like elements or homologous recombination [130]. Furthermore, synthetic gene cassettes and linear gene cassette arrays have been naturally transferred into *Pseudomonas stutzeri* in which the acquired gene cassettes were integrated into an integron *att*I site by site-specific recombination [131]. Moreover, self-replicating plasmids harboring class 1 integrons have shown the ability to be maintained in *A. baumannii* after being naturally transferred [132]. Finally, one study showed the possible acquisition of class 1 integron gene cassettes via transduction by P22-like phage ES18 and by phage PDT17 in *Salmonella enetrica* serovar Typhimurium [133].

1.7. Project objectives

This project with its branches had several objectives. The first objective was to biochemically characterize a nitrilase (NitraS-ATII) that we have previously isolated from the hypersaline and thermophilic Atlantis II Deep brine pool in the Red Sea to assess its possible biotechnological potential.

The second objective of this study was to identify and analyze integrons, CALINs and associated genetic elements in halophilic genomes and hypersaline metagenomes, in addition to

addressing the putative links between integrons and different MGEs such as ISs and group II introns and their role in microbial adaptation in hypersaline environments.

This was achieved by two approaches, the first was a PCR approach searching for integrons in hypersaline metagenomic libraries using predesigned primers for *int*I genes, and the second was a bioinformatics approach based on detecting integrons and CALINs in halophilic genomes and hypersaline metagenomic assemblies using the IntegronFinder pipeline.

We tried to assess the recombination activity of two *in vivo* expressed IntIs, from hypersaline metagenomes, using a pre-developed chromosomal deletion assay.

Moreover, annotation and identification of associated gene cassettes in all identified integrons and CALINs was also important in order to investigate possible roles of these gene cassettes in integrons mobilization and interaction in their environment. Identification of putative promoters for all identified *int*I genes was aimed in order to get a broader picture on possible regulatory mechanisms that control IntI expression and recombination events.

Finally, we aimed to unravel the differential microbial phylogenetic diversity in two Egyptian hypersaline aquatic environments: the athalassohaline lake “Aghormy Lake” in Siwa Oasis and the thalassohaline “Sebeaka saltern” at the eastern part of Bardawil Lagoon.

Chapter 2: Biochemical characterization of Atlantis II Deep Red Sea brine pool-nitrilase with unique thermostability profile and heavy metal tolerance properties

Abstract

Nitrilases gained increasing attention because of the abundance of nitrile compounds in nature and their use in fine chemicals and pharmaceutical industries. Nitrilases hydrolyze nitriles in a one-step reaction into their corresponding carboxylic acid and ammonia. In this study, we have biochemically characterized a nitrilase (NitraS-ATII) that we have previously isolated from the Lower Convective Layer (LCL) of the Atlantis II Deep Brine Pool in the Red Sea. The LCL environment is characterized by elevated temperature (68°C), high salt concentrations (250 ppt), anoxic conditions and high heavy metal concentrations. NtraS-ATII was selectivite towards dinitriles, suggesting a possible industrial application in the synthesis of cyanocarboxylic acids. Furthermore, NitraS-ATII showed higher thermal stability compared to a closely related nitrilase, in addition to its tolerance towards high concentrations of some heavy metals. The properties of NitraS-ATII may shed light on bacterial adaptation in extreme environments with high salinity and temperature, in addition to its potential use in bioremediation processes.

2.1. Introduction

Nitrilases are hydrolytic enzymes that can hydrolyze nitriles (R-CN) in a one-step reaction into their corresponding carboxylic acids (R-COOH) and ammonia (NH_3) [25,134]. They are classified based on their substrate specificity into aromatic, aliphatic and arylacetonitrile nitrilases. However, some nitrilases could have a broad substrate specificity [25]. All nitrilases are characterized by the presence of a catalytic triad of glutamate-lysine-cysteine [135]

Nitriles, which are organic cyanides, are abundant in nature [25,136]. They are used in the synthesis of different fine chemicals and pharmaceutical industries and could also be produced as industrial waste products [25]. Nitriles can be processed either chemically or enzymatically. [135]. Enzymatic processing of nitriles by nitrilases has proven to be superior than the use of conventional chemical methods. No toxic byproducts, such as HCN gas, are produced upon using nitrilases [25,27,137,138]. In addition, specific isomers can be synthesized due to the stereo- and/or regio-selectivity of nitrilases. Furthurmore, some nitrilases have the property of hydrolyzing a single cyano-group in dinitriles or polynitriles producing cyanocarboxylic acids. Those are used in different industries [25,27]. Nitrilases have also facilitated different bioremediation processes such as the detoxification of cyanide containing wastes and the degradation of nitrile-containing herbicides [25].

Nit1C gene cluster is a conserved gene cluster composed of seven co-transcribed genes. A *nit*C gene within the cluster encodes for a nitrilase [29]. The cluster was found to be involved in free cyanide and nitrile assimilation in *Pseudomonas pseudoalcaligenes* which can grow on cyanide as a sole nitrogen source [30]. A homologue from the same cluster in *Pseudomonas fluorescens* was found to be essential for cyanide assimilation as well [31] . This cluster was also identified in *Klebsiella pneumoniae* pLVPK plasmid suggesting its possible horizontal transfer [29]. In general, different nitrilase-producing bacteria were found to utilize nitriles as a sole source of carbon and nitrogen [25,137].

Isolation and characterization of nitrilases from extreme environments may lead to the identification of nitrilases with unique characteristics. Atlantis II Deep brine pool (ATII D) is a unique extreme environment. It is the largest of the 25 brine pools in the Red Sea [6,9]. It reaches a maximum depth of 2,194 m [9], and is characterized by its elevated temperatures and salinity [6]. The brine is segregated into four layers based on differences in temperature, salinity and oxygen content. The deepest is the lower convective layer (LCL) at which the highest temperature (68°C), salinity (250 ppt) and heavy metal concentrations are reached [6,7] .

In this study, we have biochemically characterized NitraS-ATII, a previously isolated nitralse form the LCL of the Red Sea Atlantis II Deep brine pool (KT354778.1) [139]. The gene encoding for NitraS-ATII resides in a Nit1C gene cluster [139] suggesting its role in cyanide and nitrile assimilation. Its thermostability and heavy metal tolerance were compared to *Rhodobacter sphaeroides* LHS-305 nitrilase [140], which has an 84% similarity to NitraS-ATII. NitraS-ATII has shown higher thermal stability and high tolerance towards different heavy metals.

2.2. Materials and methods

2.2.1. Expression of NitraS-ATII

A recombinant plasmid p-NitraS-ATII.A was obtained from GenScript, with a codon-optimized synthesized NitraS-ATII gene in pET-28b+ with a C-terminal His-tag. p-NitraS-ATII was transformed into *E.coli* BL21 (DE3) for expression. Cultures with kanamycin (50µg/ml) at an OD_{600} of ~0.6 were induced using 0.1mM IPTG for two hours at 37°C. Cells were pelleted and cell lysates were analyzed using 12% SDS-PAGE stained with Coomassie blue R250 [141].

2.2.2. Purification of His-tagged NitraS-ATII

Cell pellets were freezed (in ice cold ethanol) and thawed (42°C), followed by their re-suspension in binding buffer, pH 8 (20mM sodium phosphate buffer, 40mM imidazole & 500mM NaCl). Lysozyme (1mg/ml) and 1mM of phenylmethylsulfonyl fluoride (PMSF) were added to the formed suspension, then it was incubated on ice for 30 minutes with occasional shaking. Sonication of the cells was done for ten minutes with bursts of ten seconds interrupted by ten-second intervals. The supernatant was then separated from the cell debris by centrifugation. NitraS-ATII protein was

purified using Ni-NTA affinity chromatography with Ni-NTA agarose resin (Invitrogen™) at 4°C according to native condition specifications. The protein was eluted with 20mM Sodium phosphate buffer (pH 8) containing 500mM imidazole, 500mM NaCl and 50% glycerol. The eluted protein was visualized on Coomassie-stained 12% SDS-PAGE gels. The protein concentration was determined using the Pierce™ BCA Protein Assay Kit (Thermo Scientific).

2.2.3. Nitrilase activity assay

Our developed quantitative nitrilase activity assay was based on a spectrophotometric-method and a fluorometric assay developed by Goyal *et al* [142] and Banerjee et al [143], repectively. In a 100 µl reaction, 10 µl of purified protein (100 µg/ml) were added to 50 mM potassium phosphate buffer (pH 7), 2 mM Dithiothreitol (DTT) and 400 mM succinonitrile as a substarte. DTT (2mM) was added to ensure that all cysteine residues are reduced. A 30-minute reaction was performed at 40°C/100 rpm and then stopped by the addition of an equal volume of 100 mM HCl. The reaction mixture was centrifuged at 5000 xg for 10 minutes. For the detection of liberated ammonia, 10 µl of the reaction mixture were added to 140 µl of buffered alcoholic *o*-phthaldialdehyde/β-mercaptoethanol reagent. The isoindole derivative was allowed to develop for 30 minutes at 30°C/100 rpm. The color intensity of the developed isoindole derivative was measured at 405 nm using the FLUOstar OPTIMA microplate reader (BMG LABTECH).

In order to prepare the working reagent used in the assay one day before the assay, alcoholic *o*-phthaldialdehyde and β-mercaptoethanol were prepared first. We dissolved 100 mg *o*-phthaldialdehyde in 10 ml of absolute ethanol, and 50 µl of β-mercaptoethanol in 10 ml of absolute ethanol. For preparing the working reagent, 2.25 ml of both alcoholic *o*-phthaldialdehyde and alcoholic β-mercaptoethanol were added to 45.5 ml of 200 mM potassium phosphate buffer (pH 7.4) [143]. NH_4Cl was used to draw a standard curve to determine the ammonia concentration. The specific activity of the enzyme was measured in U/mg of protein. One unit (U) of specific activity is defined as micromoles of ammonia produced in 1 minute by 1 mg of the enzyme (µmole min^{-1} mg^{-1})

2.2.4. Effect of pH on NitraS-ATII activity

The quantitative assay was done at different pHs ranging from 3.5 to 11, using acetate buffer (pH 3.5-5), phosphate buffer (pH 6-8) and carbonate buffer (pH 9-11).

2.2.5. Enzyme kinetics

The initial reaction rate was determined in a 10-minute reaction with 100 ug/ml of the purified NitraS-ATII and different concentrations of succinonitrile (0-975 mM). The pH of the reaction was set at 7 using phosphate buffer. We determined the Michaelis-Menten kinetic parameters (K_m, V_{max} and K_{cat}) usings GraphPad Prism® (version 5.00 for windows).

2.2.6. Effect of temperature on NitraS-ATII activity

In order to determine the enzyme thermal stability, purified NitraS-ATII was incubated for different time intervals (30 sec- 60 min) at different temperatures (0-80°C) before assaying the residual activity using the activity assay described above. The residual activity of NitraS-ATII after 30 seconds and 1 minute incubations at different temperatures were then compared to those of *Rhodobacter sphaeroides* LHS-305 nitrilase (GenBank accession number JN635494) [140]. In order to determine the thermosensitivity of the enzyme, the activity assay was done at different temperatures (15-70°C) and compared to the thermosensitivity profile of the *R. sphaeroides* LHS-305 nitrilase.

2.2.7. Effect of salt on NitraS-ATII activity

To examine the effect of salt on NitraS-ATII activity, the activity assay was performed at different NaCl concentrations (0-4M).

2.2.8. Effect of different metals on NitraS-ATII activity

We assessed the activity of NitraS-ATII in presence of different metals. Here, no DTT was added to the reaction mixture because of its known ability to form complexes with different metal ions such as nickel, copper, zinc, cadmium [144] and mercury [145]. Different concentrations ranging from 1 to 25 mM of $NiSO_4$, $CdCl_2$, $CoCl_2$, $ZnCl_2$, $MgSO_4$ or $MnSO_4$ were added to the reaction mixture. Concentrations of 6 and 12µM were used in case of $HgCl_2$, and 0.5, 1 and 2 mM in case of $CuSO_4$. The same reaction conditions were used with *R. sphaeroides* LHS-305 nitrilase to compare the activity of both nitrilases.

All experiments were done at least in triplicates and all graphs were created by GraphPad Prism® (version 5 for windows).

2.2.9. 3D homology modelling and identification of salt bridges

We used Phyre2 tool [146] for homology modeling of both NitraS-ATII (GenBank accession no. KT354778.1) and *R. sphaeroides* nitrilase.(JN635494). Visualization and superimposition of the produced 3D structures was done using PyMOL™ Evaluation Product - Copyright © 2008 DeLano Scientific LLC. In order to determine possible salt bridges in NitraS-ATII and *R. sphaeroides* nitrilase, we used ESBRI online tool [147,148,149,150].

2.3. Results

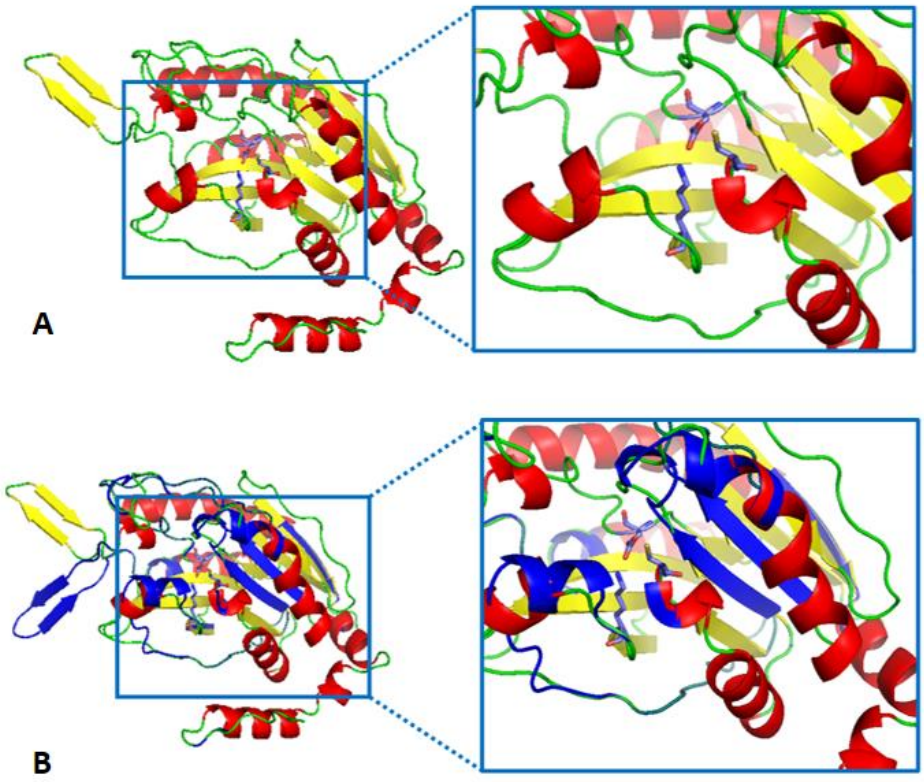


Fig.2.1 **NitraS-ATII 3D structure.** A. Three dimensional model of NitraS-ATII protein. The model was obtained using Phyre2 tool with 100% confidence and visualized using pyMOL and colored according to the secondary structure showing helices in red, sheets in yellow and loops in green. Residues of the catalytic triad are shown as a stick representation, showing carbon atoms in cyan, nitrogen atoms in red, oxygen atoms in blue and sulfur atoms in yellow. c. Superimposition of NitraS-ATII and control nitrilase from *Rhodobacter sphaeroides* LHS-305. NiraS-ATII, colored according to secondary structure, is superimposed on *R. sphaeroides* LHS-305 nitrilase, shown in blue, where the residues of the catalytic triad showed perfect superimposition.

2.3.1. Structural comparison between NitraS-ATII and *R. sphaeroides* LHS-305 nitrilase

R. sphaeroides LHS-305 nitrilase [140] was used as a control in our study. It showed 76% identity and 84% similarity with NitraS-ATII. A 3D structure model for both nitrilases was obtained with 100% confidence using Phyre2 tool [146], using the crystal structure of nit6803 nitrilase as a template (template c3wuyA). This template showed the highest identity (71%) to both nitrilases. Superimposition of the 3D structures for both nitrilases showed few variations; however, the catalytic triad (E-K-C) residues showed perfect superimposition (Fig.2.1). Using ESBRI tool, 119 and 102 salt bridges were predicted in NitraS-ATII and *R. sphaeriodes* LHS-305 nitrilase, respectively.

2.3.2. A thermostable NitraS-ATII without evident acidophilic or halophilic activity

The highest activity of NitraS-ATII on succinonitrile was achieved at pH 7 using phosphate buffer. Loss of activity was observed at acidic pHs and a minimal activity was retained at basic pHs (Fig.2.2A).

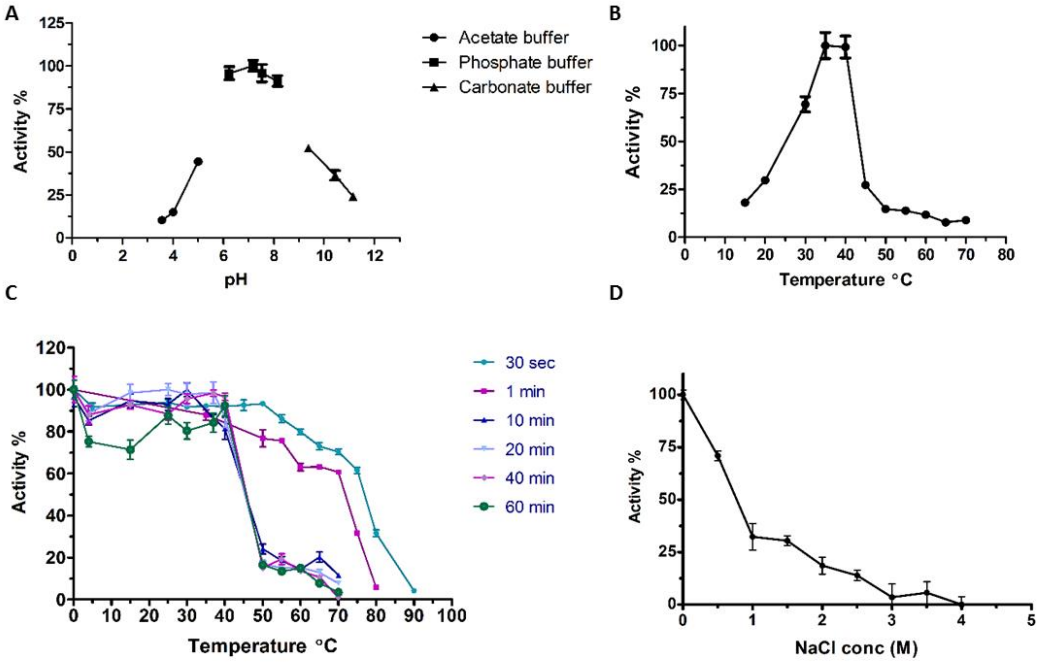


Fig.2.2 **Kinetics of NitraS-ATII at different pHs, temperatures and salt concentrations. A**. Effect of pH on NitraS-ATII activity was assessed in acetate buffer pH (3.5-5), phosphate buffer for pH (6-8) and carbonate buffer for pH (9-11). Optimum activity was achieved at pH of 7. **B.** Effect of the reaction temperature on NitraS-ATII activity. The optimum temperature of the reaction is shown to be 40°C and the activity was almost abolished at higher temperatures. **C**. The residual activity of NitraS-ATII after incubation at different temperatures for different periods. **D**. NitraS-ATII activity at different NaCl concentrations. A decrease in the nitrilase activity is observed with the increase in the salt concentration.

NitraS-ATII showed an optimum activity at temperature 35-40°C. The activity dropped sharply at temperatures higher than 40°C (Fig.2.2B). Upon incubating NitraS-ATII at different temperatures for different periods of time (10-60 min) before starting the reaction, the residual activity of the enzyme decreased sharply at temperatures higher than 40°C (Fig.2.2C). However, although *R. sphaeriodes* LHS-305 nitrilase showed an optimum reaction temperature of 50°C, a significant difference between the residual activity of both nitrilases was observed upon incubation at 70, 75 and 80°C for 30 sec or one minute, before initiating the reaction. For instance, NitraS-ATII retained 61.6% and 31.8% of its activity after incubation at 75°C for 30 seconds and one minute, respectively. On the other hand, *R. sphaeriodes* LHS-305 nitrilase maintained only 6.8% and 2.8% of its activity under the same conditions. NitraS-ATII retained 70.4% and 60.7% of its activity

following the 70°C pre-incubation for 30 seconds and one minute, respectively; while *R. sphaeriodes* LHS-305 nitrilase only retained 55.3% and 10.1% of its activity under the same conditions (Fig.2.3)

NitraS-ATII did not show any halophilic properties as its activity decreased upon increasing NaCl concentration (Fig.2.2D).

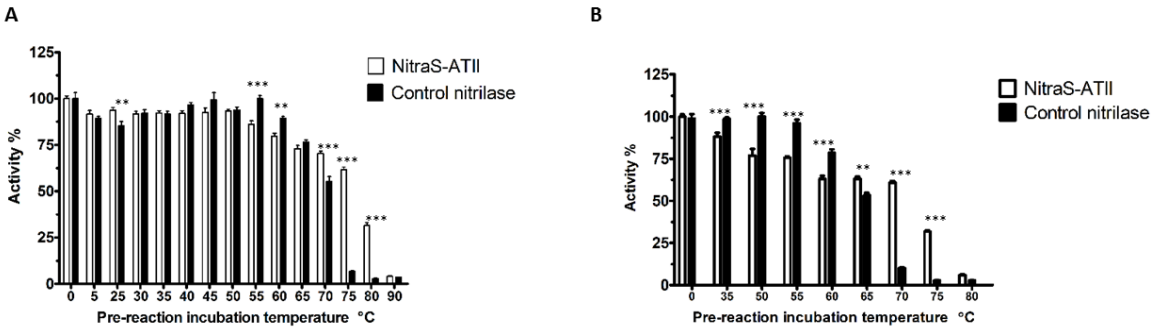


Fig.2.3 **Thermal stability of NitraS-ATII.** The enzymes were incubated at high temperatures for 30 seconds (A) or one minute (B) prior performing the reaction and measuring the residual activity. Two-way Anova test followed by Bonferroni post-hoc test was performed using GraphPad Prism® (version 5.00 for windows). *** indicates p-values lower than 0.001 and ** for p-values lower than 0.01.

Using a range of succinonitrile concentrations from 0-975mM, we measured the initial velocities of the reaction in a ten-minute-reaction. A typical Michaelis-Menten kinetics plot was obtained (Fig.2.4). We obtained a K_m of 59.4±6.831mM and V_{max} of 2.432±0.04993µM NH_3/min (6.081e-006 µmole NH_3/sec). The specific activity of NitraS-ATII was 0.73 U/mg of enzyme and k_{cat} was 0.4721 sec^{-1}.

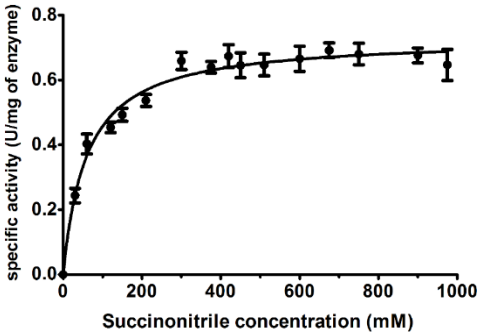


Fig.2.4 **Effect of substrate (succinonitrile) concentration on NitraS-ATII specific activity.** Specific activity was measured in U per mg of the enzyme. One unit (U) of specific activity is defined as micromoles of ammonia produced in one min by one mg of the enzyme (µmole min^{-1} mg^{-1}).

2.3.3. Maintained activity of NitraS-ATII at high concentrations of some metals

Both Cu^{2+} and Hg^{2+} inhibited the activity of NitraS-ATII at low concentrations (0.5, 1 or 2 mM $CuSO_4$ and 6 or 12µM $HgCl_2$). However, in presence of $HgCl_2$, the activity was nearly reversed when DTT (2mM) was added to the reaction mixture (97.6% with 6µM $HgCl_2$ and 90.3% with 12µM $HgCl_2$). However, a similar effect was not observed with $CuSO_4$.

NitraS-ATII retained most of its activity even in the presence of high concentrations of $ZnCl_2$, $MgSO_4$ and $MnSO_4$, whereas, a week inhibitory effect was observed when $NiSO_4$, $CdCl_2$ or $CoCl_2$ were added (Fig.2.5 and Table 2.1). Upon comparing the effect of some heavy metals on both NitraS-ATII and *R. sphaeriodes* LHS-305, the inhibitory effect of Ni^{2+} on NitraS-ATII was significantly higher than that on *R. sphaeriodes* LHS-305 nitrilase (p value = 0.0022). In contrast, NitraS-ATII showed significantly higher activity in presence of Zn^{2+} at concentrations lower than 16 mM $ZnCl_2$ (p value = 0.0096) and Mn^{2+} (p value = 0.0118) (Fig.2.6).

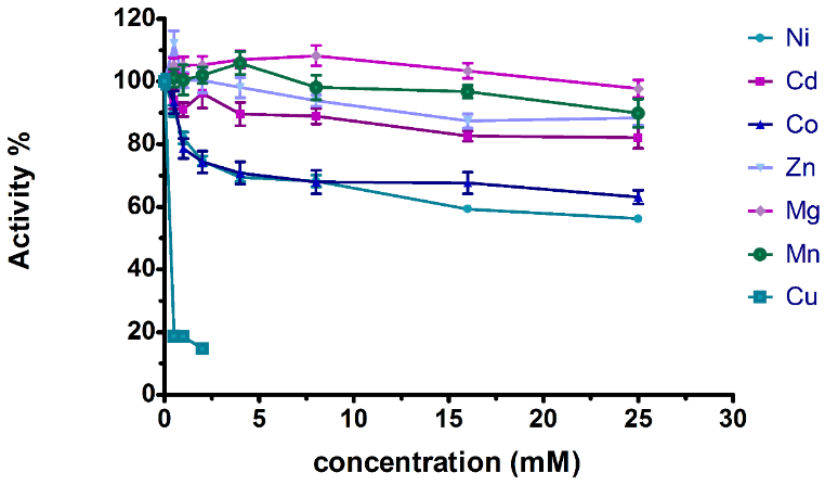


Fig.2.5 **Effect of different metal ions on NitraS-ATII activity.** The nitrilase activity is retained at high concentrations of Mg^{2+} and Mn^{2+} and a lower extent with Zn^{2+}. A high degree of tolerance is observed towards $Cd^{2+}$and Co^{2+} and to a lower extent towards Ni^{2+}; whereas, inactivation is observed even with low concentrations of Cu^{2+} and Hg^{2+}.

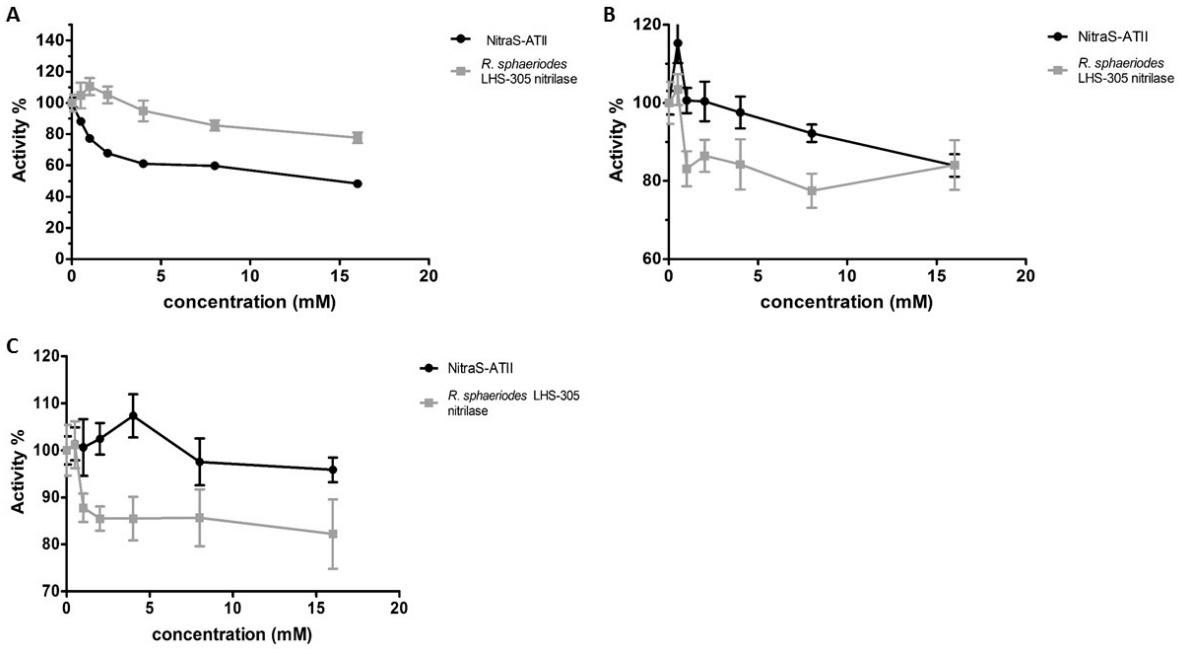


Fig.2.6 ***Comparison of the effect of selected metal ions on the activities of NitraS-ATII and*** **R. sphaeriodes** ***LHS-305 nitrilase.*** *Each panel shows the activity percentage in the presence of increasing concentrations of a metal ion. a. In presence of* Ni^{2+}, R. sphaeriodes *LHS-305 nitrilase retains higher activity (t test p-value =* $2.2x10^{-3}$*). b. In presence of* Zn^{2+}, *NitraS-ATII retains higher activity (t test p-value =* $9.6x10^{-3}$*). c. In presence of* Mn^{2+}, *NitraS-ATII retains higher activity (t test p-value =* $11.8x10^{-3}$*)*

Table 2.1 Effect of different metal ions concentrations on the activity of NitraS-ATII. The color code shows the activity percentage, with the highest activity shown in red and the lowest in yellow.

concentration (mM)	$NiSO_4$	$CdCl_2$	$CoCl_2$	$ZnCl_2$	$MgSO_4$	$MnSO_4$	$CuSO_4$	$HgCl_2$
0	100 ± 7.19	100 ±7. 19	100 ± 7.19	100 ± 7.19	100 ± 7.19	100 ± 7.19	100 ± 7.19	100 ± 7.8
6 µM	-	-	-	-	-	-	-	16.77 ± 1.98
12 µM	-	-	-	-	-	-	-	14.59 ± 1.56
0.5	90.63 ± 5.75	94.3 ± 9.39	93.28 ± 10.12	112.04 ± 12.12	105.13 ± 6.81	101.07 ± 8.3	18.64 ± 1.67	-
1	81.97 ± 5.75	91.05 ± 6.88	78.59 ± 9.52	100.48 ± 7.59	105.13 ± 8.11	100.48 ± 11.67	18.57 ± 1.77	-
2	74.58 ± 4.54	96.21 ± 14.01	74.29 ± 10.65	100.29 ± 11.95	105.24 ± 8.27	101.93 ± 7.91	14.71 ± 2.82	-
4	69.36 ± 4.23	89.58 ± 11.13	70.76 ± 10.61	98.05 ± 9.62	106.93 ± 8.4	105.79 ± 10.9	-	-
8	68.21 ± 5.74	88.94 ± 7.33	67.89 ± 11.26	93.86 ± 5.25	108.18 ± 9.6	98.07 ± 11.78	-	-
16	59.29 ± 2.39	82.52 ± 4.76	67.6 ± 10.34	87.39 ± 6.81	103.4 ± 7.19	96.73 ± 6.16	-	-
25	56.24 ± 3.23	82.06 ± 10.33	63.16 ± 6.6	88.29 ± 6.93	97.63 ± 8.54	89.84 ± 13.69	-	-

2.4. Discussion

Characterization of new nitrilases has gained some attention due to their potential uses in bioremediation and green industry. NitraS-ATII gene isolated from the LCL layer of the Atlantis II Deep brine pool in the Red Sea was found to be part of a conserved gene cluster, Nit1C, present in several bacterial phyla and in microorganisms that inhabit diverse environments [151]. This gene cluster was found to be involved in cyanide and nitrile assimilation [30], which increases the potential of exploiting NitraS-ATII and associated genes in biodegradation of cyanogenic wastes.

It was surprising to find that the optimum temperature of NitraS-ATII activity is 35-40°C, demonstrating its mesophilic nature. This was completely opposite to what was expected from an enzyme isolated from the Atlantis II Deep LCL known by its high temperature (68°C). However, when NitraS-ATII was pre-incubated at high temperatures for short periods of time, its activity was not greatly affected. NitraS-ATII retained more than 60% of its activity after incubation at 70°C for 30 seconds or one minute. This may indicate that NitraS-ATII requires additional interactions in order to maintain its properly folded structure and activity under the ATII LCL environmental condition for longer periods.

Since NitraS-ATII lost most of its activity at NaCl concentrations higher than 2.0 $mol.L^{-1}$, it would be expected that the microorganism from which its gene was isolated would produce compatible solutes to tolerate the hypersalinity of the ATII LCL environment. Compatible solutes are known by their possible role in protection against other stresses rather than osmolarity [152] Even though a consensus about their *in vivo* role towards increasing thermostability of proteins have not been reached, some halophilic (hyper)thermophiles were found to accumulate them at high temperatures, pointing towards a possible auxiliary role [153].

Upon comparing their short-period thermal stability profile, NitraS-ATII retained most of its activity at 70 °C, whereas *R. sphaeriodes* LHS-305 nitrilase almost lost its activity at the same conditions. The difference in predicted number of salt bridges between the two enzymes, 119 in case of NitraS-ATII and 102 in case of *R. sphaeriodes* LHS-305 nitrilase 102, could be high enough to account for the observed higher thermal stability of NitraS-ATII. The increased number of salt bridges is considered one of the characteristics that may enhance protein thermostability [154] [155,156,157]. For instance, in a study in which a mesophilic β-glucosidase from *Bacillus polymyxa* was engineered to contain four extra salt bridges at specific positions, a significant increase in the engineered enzyme thermal stability was observed [158].

The LCL of Atlantis II Deep is particularly characterized by its high heavy metal content [10]; thus, we expected that NitraS-ATII might be tolerant to certain metals. Several studies have shown the effect of metals on different nitrilases; however, they used low metal ions concentrations (1 or 5 mM) [138,140,159,160]. In our study, we examined the effect of different metals at concentrations up to 25 mM. Generally, NitraS-ATII showed high tolerance towards most of the tested metal ions. However, a strong inhibitory effect of Cu^{2+} and Hg^{2+} ions was observed. This

could be attributed to the possible complex formation between these ions and the thiol group in the catalytic cysteine residue of the enzyme.

Comparing the results of metal tolerance of NitraS-ATII with that of *R. sphaeriodes* LHS-305 nitrilase showed that NitraS-ATII has significantly higher tolerance to high concentrations of Mn^{2+} and Zn^{2+}, which are both present at high concentrations in the LCL [10]. On the other hand, *R. sphaeriodes* LHS-305 nitrilase showed higher tolerance towards Ni^{2+}. As no nickel was detected in the LCL, it seems that the inhibitory effect of nickel on NitraS-ATII has no adverse consequences in its natural environment. This points towards the intricate molecular adaptation of NitraS-ATII to its extreme environment.

2.5. Conclusions

Identification of new nitrilases holds a great potential for their exploitation in different industries and bioremediation processes. The plethora of available metagenomic databases of extreme environments is a gold mine for digging for extremophilic enzymes. In this study, we have biochemically characterized a nitrilase NitraS-ATII that we have previously isolated from the LCL of Atlantis II Deep brine pool. NitraS-ATII showed higher thermal stability when compared to a closely related nitrilase. In addition, NitraS-ATII showed high tolerance to different metals especially those abundant in the LCL. Further studies on NitraS-ATII to determine its stereo- and regio-selectivity and to assess its biotechnological potential are needed.

Chapter 3: Identification of integrons in two hypersaline aquatic metagenomes using PCR and bioinformatics approaches

Abstract

Integrons are genetic platforms that allow the expression of genes arranged in unique structures named gene cassettes. These gene cassettes can be integrated or excised from the integron by an integron integrase (IntI) encoded by the integron itself. Integrons are widely spread in bacteria in different environments. Here, using a PCR screening approach and an HMM scan for detection of *int*I genes in different hypersaline metagenomes, we have identified two integrons in two in hypersaline environments: Aghormy Lake in Siwa Oasis in Egypt and Kebrit Deep Brine Upper interface in Red Sea. Integron components were identified in both integrons and their intIs did not belong to well-studied IntI classes. The identified *int*I genes were synthesised and expressed in pBAD18 plasmid to test their *in vivo* excision activity. However, no activity was measured for both proteins. This could be attributed to the dependence on an excision assay primarily designed for IntI1 and the possible differences in recognition of different recombination sites. The identification and characterization of IntIs from hypersaline environments could have a great potential in biotechnological applications and in understanding microbial adaptation to high hypersaline environments.

3.1. Introduction

Integrons are widely spread genetic platforms for expressing different open reading frames (ORFs) arranged in unique structures named gene cassettes [40]. Free circular gene cassettes, typically composed of an ORF followed by an *att*C recombination site, can be integrated within an *att*I recombination site typically found at the 5' end of the integron integrase (intI) gene within the integron, in a reversible reaction catalyzed by the IntI as well [39,41]. An integron is composed of an *int*I gene, an *att*I recombination site and a P_C promoter for the transcription of downstream gene cassette ORFs which are usually promoterless [39].

IntIs belong to site-specific tyrosine recombinases, with XerC and XerD being their closest relatives [38,46]. Beside the conserved regions: box I and box II and patches I, II and III in all tyrosine recombinases at which the catalytic tyrosine is present in box II [38,46], intIs were found to have an extra domain near patch III that is absent from all other tyrosine recombinases [38]. This domain is referred to in some research articles as IntI patch [46]. This domain forms an alpha helix in the protein named I2, which has a role in folding the hydrophobic pockets within the intI 3D structure. This is important for stabilizing two extrahelical bases (EHBs) at *att*C site bottom strand (bs) [47].

An *att*I recombination site is mainly composed of two IntI binding sites termed R and L. The two sites form imperfect inverted repeats in which the R only has the consensus sequence of 5'-GTTRRRY-3', whereas the L site is degenerate [39]. On the other hand, *att*C site is more complex than *att*I site. It is composed of four domains: R", L", L' and R', in which only R" and R' sites are conserved with the consensus 5'-RYYYACC-3' and 5'-GTTRRRY-3', respectively [39,41]. The central region between L'' and L' varies greatly between different *att*C sites and shows different lengths. However, only the bottom strand (bs) of an *att*C is recombinogenic. It forms a hairpin structure at which R'' binds to R' and L'' binds to L' forming R and L boxes, respectively [47]. Two extrahelical bases (EHBs) at L'' orient the polarity of the recombination reaction by identifying the recombinogenic strand (bs) [39,47] .

Recombination reactions in integrons can occur between two *att*C sites, an *att*I and an *att*C site or two *att*I sites, with the latter being less efficient [39]. Intermolecular recombination reactions between *att*C X *att*C sites lead to gene cassette excision, whereas *att*I X *att*C recombination reactions lead to integration of gene cassettes into integrons [39]. Recombination occurs between G and TT in the conserved triplet GTT within the R site in *att*I site and the R box within the *att*C bs hairpin [54]. To be more precise the cleaved strand would be the opposite strand between the A and the C in the conserved AAC triplet [55]. Usually an IntI recognizes its *att*I site mainly; but it has been shown that some *att*I sites from other integron classes can be identified as well by IntI1 with lower frequency [54]. On the other hand, different IntIs can identify many *att*C sites as the secondary structure is the most important aspect for their identification [53].

P_C promoter located commonly within *int*I gene or within *att*I site drives the expression of the gene cassette ORFs. Variants of P_C promoters were identified in class 1 integrons [39,41]. However, the expression of gene cassettes differs based on their distance from the P_C promoter, as the first gene cassette shows the highest expression levels.

Screening for integrons was normally done using PCR primers designed for capturing *int*I genes or known gene cassettes [161], or later on using degenerate primers to capture a wider range of integrons belonging to different classes [46,78,162]. Degenerate primers designed to target diverse gene cassettes tried to target conserved regions of *att*C sites [46,78,162]. Later on, different computational pipelines were developed for identification of integrons, but they were restricted to certain well-studied integron classes, because of the difficulties in the identification of the diverse *att*C sites [79,80,81]. Identifying integron gene cassettes proved to be hard because of the high diversity of *att*C sites [50]. Based on *att*C sites restricted to certain integron classes, different computational programs were developed [79,80,81].

Measuring the recombination activity of identified integron integrases has been done using different assays that can measure integration, excision or both reactions. Conjugation

assays, based on the transfer of conjugative plasmids carrying either a single *att*C site or a gene cassette flanked by two *att*C sites into an IntI expressing recipient cell, were extensively utilized. The recipient cell either contains an *att*I site at which the transferred *att*C-harboring plasmid can integrate by an *att*I X *att*C recombination reaction to measure integration frequency or the expressed IntI can excise the gene cassette in the conjugative plasmid by an *att*C X *att*C recombination event to test excision frequency [82,84]. Some assays were based on double transformation of cells with an IntI-expressing plasmid that harbors an *att*I site, and another plasmid carrying a gene cassette. Excision of the gene cassette can then be detected by PCR amplification and sequencing [85]. If the used gene cassette encodes for an antibiotic resistance gene, then loss of antibiotic resistance would indicate a positive excision event [54]. The frequency of cassette excision in chromosomes can also be measured using a chromosomal deletion assay. In this assay, a *dap*A gene interrupted by a synthetic cassette is integrated into recombinant cells that cannot synthesize 2, 6-diaminopimelic acid (DAP). The cells need DAP to be supplemented in the medium to grow. Transformation with IntI-expressing plasmid would lead to the excision of the synthetic cassette, restoring *dap*A function and allowing cells to grow in absence of DAP [86]. All of these assays were mainly developed for intI1 [82,84,86] and some were used with intI2-4 as well [53,54,86].

Integrons were found in about 7.2% of bacterial complete genomes [50]. They were also isolated from different environments such as soil, hot springs [41], polar sediments [61], glaciers [62], clinical isolates [59] and marine environments [63]. Hypersaline aquatic environments are unique environments where microorganisms residing there should possess unique machineries to tolerate high salinity. The most common adaptation strategy is a salt-out strategy on which the microorganism expels extra salts, while accumulating osmolytes intracellularly to create an osmotic balance to the hypersaline surrounding [1]. The other strategy, is a salt-in strategy on which the microorganism accumulates high concentrations of KCl. This requires adaptation of the whole intracellular machinery to high intracellular salt concentrations [1].

Our analysis here focused on two hypersaline aquatic environments: Aghormy Lake in Siwa Oasis and Kebrit Depp Brine in Red Sea. Siwa Oasis at the Western Desert in Egypt, is a depression between latitudes 29° 05′ N and 29° 25′ N and longitudes 25° 05′ E and 265° 06′ E with an area of about 1200 km^2. It is characterized by an arid to semi-arid climate with scarce rainfall [163]. The deepest parts of the oasis are occupied by salty lakes surrounded by salt marches. Lakes in Siwa Oasis, are the natural discharge areas for water coming from the abundant artesian wells, springs and cultivated areas [13]. Aghormy Lake, located 18 m below sea level, is characterized by total dissolved solids (TDS) of 220.03 ppt and a pH of 7.83 [14]. On the other hand, Kebrit Deep is a brine pool at the Red Sea located at latitude 24° 43′ N and longitude 36° 17′ E [164] with an area of about 2.5 km^2 at the brine surface and a depth of 1580.2 m [165]. The brine

is characterized by its ambient temperatures and high H_2S content which is reflected by its name [164]. Its upper interface (KD UINF) is located at 1468 m depth [164] with 2.2 mol l^{-1} Chloride concentration (~ 79 ppt), whereas Cl^- concentrations reaches 5.06 mol l^{-1} at the brine itself [166].

Here we have identified two complete integrons, one from Aghormy Lake and one from Kebrit Deep Upper interface. IntIs identified from both integrons were synthesized and expressed *in vivo*. However, we need more specific assays to determine their recombination activity as no activity was measured upon using an already developed model for class 1 integron.

3.2. Materials and Methods

3.2.1. Sampling, DNA extraction and amplification of *int*I fragments from metagenomic DNA.

Filtered water samples through membrane 0.2 µm filters, Millipore, were obtained from Aghormy Lake in Siwa Oasis in the Western Desert, Egypt 29°11′ 44″N, 25°35′18″E (Fig.3.1) in Spring. DNA was extracted using PowerMax Soil DNA Isolation kit (MO BIO laboratories) according to the manufacturers 'recommendations. The extracted DNA was then concentrated using absolute molecular grade ethanol (Sigma Aldrich).

Fig.3.1 **Map showing Aghormy Lake located in Siwa Oasis in the Western Desert in Egypt.**

PCR on the extracted metagenomic DNA was performed using predesigned int1.F primer (5'-GGGTCAAGGATCTGGATTTCG-3) [161] with the degenerate primer int-948R (5′-NARTACRTGNGTRTADATCATNGT-3′) [162] to get 484 bp-bands. Bands were extracted using QIAquick Gel extraction kit (Qiagen). Then nested PCR on extracted bands using int1.F with intI-

864R (5′-YAGCAGATGNGTGGCRAAVSWRTGSCG-3′) to generate a band of ~404bp and intI528F (5′-CGNGAYGGYAARGGSRNVAAGGAYCGS-3′) with intI-864R (5′-YAGCAGATGNGTGGCRAAVSWRTGSCG-3′) to generate another band of ~363bp [162]. Obtained bands were ligated into pGEM-T easy vector (Promega) and the recombinant plasmid was transformed into *E. coli* top 10 and cultured on LB agar containing 0.5mM IPTG, 40ug/ml X-gal and 100ug/ml ampicillin. Plasmid extraction was done from few selected positive colonies using QIAprep Spin Miniprep kit (Qiagen). Inserted bands were amplified using M13 primers and sequenced to confirm the presence of *Int*I genes in the environmental DNA.

3.2.2. Construction of AGH fosmid library

The sheared DNA pieces with the proper size (40 kb) were subjected to end repair and ligation into pCC1FOS™ Vector, followed by packaging of the recombinant fosmids into MaxPlax™ Lambda Packaging Extracts (λ phage extracts). The Fosmids were transduced into Phage T1-Resistant EPI300™-T1R *E. coli* Plating Strain. Infected cells were cultured on LB agar plates with chloramphenicol and all obtained colonies were picked and cultured in LB broth with chloramphenicol and autoinduction solution to allow the extraction of fosmids from each colony and the creation of a glycerol stock for each colony (Fig.3.2). The creation of the library was done using CopyControl™ Fosmid Library Production Kit (Epicentre Biotechnologies) according to the manufacturers 'recommendations and the constructed library was referred to as AGH library.

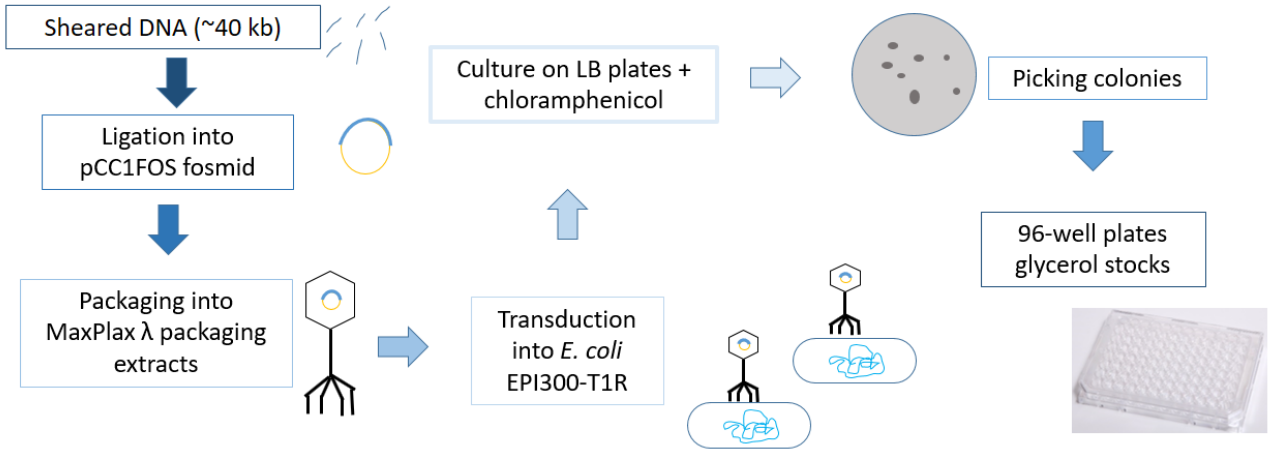


Fig.3.2 **Schematic representation showing steps of the construction of AGH fosmid library.**

3.2.3. PCR screening of the metagenomic libraries searching for integron integrases and fosmid sequencing

PCR screening to amplify existing *int*I genes was performed on AGH fosmid library and on ATII LCL fosmid library. The latter is a 10,656-clones-library, previously constructed from the lower convective layer (LCL) from Atlantis II brine pool in the Red Sea (21° 20.72' N and 38° 04.59' E) [23]. Using int1.F with int-948R, amplified bands near the expected size (~484 bp) using intI.F with intI-864R were extracted using QIAquick Gel extraction kit (Qiagen) and then reamplified using Int1.F with intI-864R to generate a band of ~404bp and Int528F with intI-864R to generate another band of ~363bp. The amplified bands were extracted and sequenced. Positive fosmids were shotgun sequenced using MiSeq platform.

3.2.4. Identification of *int*I sequences within Red Sea brine pool water and sediment metagenomes

In order to investigate the presence of integrons in the Red Sea brine pools and sediments assembled metagenomes (Atlantis II Deep, Discovery Deep and Kebrit Deep) [164,167,168], an HMM profile from six curated and biologically tested integron integrase sequences was done followed by an HMM search and blastx of the sequenced Red Sea brine pools' metagenomes (Table 3.1) against IntI sequences from INTEGRALL database [169] and NCBI nr protein database.

Table 3.1 Studied Red Sea brine pools assembled metagenomes.

Site	Description	Reference	Total assembled sequence length	Number of contigs
ATII SDM	Atlantis II Deep Brine Sediment, Red Sea	[6,167,168]	40413330	41726
DD SDM	Discovery Deep Brine Sediment, Red Sea	[6,167,168]	52421642	51829
ATII INF	Atlantis II Deep Brine interface, Red Sea	[164,168]	16014945	24317
DD INF	Discovery Deep Brine interface , Red Sea	[164,168]	11647401	18413
KD UINF	Kebrit Deep Upper interface, Red Sea	[164,168]	42652688	45750
KD LINF	Kebrit Deep Lower interface, Red Sea	[164,168]	50280352	74666
ATII LCL	Atlantis II Deep Brine, Lower convective layer, Red Sea	[164,168]	46518597	43555
ATII UCL	Atlantis II Deep Brine,Upper convective layer, Red Sea	[164,168]	21343827	29592
DD BR	Discovery Deep Brine , Red Sea	[164,168]	12244355	18850
KD BR	Kebrit Deep Brine, Red Sea	[164,168]	35162057	74666
ATII 50	Atlantis II 50 m water column, Red Sea	[168,170]	53647835	78510
ATII 200	Atlantis II 200 m water column, Red Sea	[168,170]	49971663	72359
ATII 700	Atlantis II 700 m water column, Red Sea	[168,170]	51443487	64636
ATII 1500	Atlantis II 1500 m water column, Red Sea	[168,170]	32542975	39190

3.2.5. Computational analysis on positive contigs and fosmids

Positive results were further investigated and ORFs were predicted using Metagene Annotator [171,172]. Identification of putative *att*I and *att*C recombination sites was done by manual investigation Identified gene cassettes were annotated and analyzed.

3.2.6. Integron components detection in 1G10 and KD UINF306

Recombination sites were detected by manual inspection. In case of *att*I sites, sequences at the 5' end of the *int*I gene sequence with R sites close to the consensus sequence 5'-GTTRRRY-3' were considered putative *att*I sites. Whereas in case of *att*C sites, putative *att*C where those showing R'' and R' sites close to the consensus sequences 5'-GTTRRRY-3' and 5'-RYYYAAC-3', respectively and with a bottom strand that forms a hairpin with EHBs at the L box. Putative P_{intI} and P_C promoters were identified using Bprom tool [173]. ORFs were identified using Metagene annotator [171,172] then blasted against NCBI nr protein database.

3.2.7. Gene synthesis of AGH-1G10 and KD UINF306 IntIs

*Int*I gene codon-optimized sequence from AGH-1G10 and contig00306 in KD UINF306 were synthesized and cloned into pUC57 cloning vector by GeneScript with *Sac*I and *Hind*III restriction site. The synthesized genes were then amplified from the pUC57 cloning vector with the introduction of a Shine-Delgarno sequence and *Nco*I, *Eco*RI restriction sites in the used forward primers. G10-F (CATCCATGGGAATTCTAACAAAGGAGCAAGCCATGGCCAGTTCGTCTTCCC) and G10-R (TACAAGCTTTTAGGTAACATCCGC) primers were used in case of AGH-1G10 *int*I and KU-F (CATCCATGGGAATTCTAACAAAGGAGCAAGCCATGGACCGTGTTAATAACGAGA) with KU-R (TACAAGCTTTTACAGGGTATCGCC) primers in case of KU UINF306 *int*I. An initial denaturation time of three minutes was followed by 35 cycles of 45 secs at 95°C, 45 secs at 52°C and one min at 72°C) with a final ten min extension at 72°C. The plasmid pBAD18 (ampicillin resistance) and the amplified genes were digested with *Eco*RI and *Hind*III, followed by ligation of each gene into digested pBAD18. The recombinant plasmids with AGH-1G10 and KD UINF306 were named G10-pBAD and KU-pBAD, respectively. The sequences of the cloned genes were confirmed by sequencing and PCR.

3.2.8. Quantitative *in vivo* excision assay

Transformants CG10 and CKU were constructed by the transformation of G10-pBAD and KU-pBAD into the Spectinomycin resistant *E.coli* B548 (derivative of MG1655 Δ*dap*A *rec*A269::tn10 in which *dapA* gene is interrupted by a synthetic 400bp lacZ cassette flanked by two *att*C sites: $attC_{aadA7}$ and $attC_{ereA2}$ [174]). Excision of the cassette restores a functional *dap*A gene allowing the strain to grow in a DAP-free medium. *dap*A is under control of P*lac* promoter; thus a functional *dap*A gene needs IPTG for induction. Transformants were grown on LB with DAP, spectinomycin, ampicillin and glucose.

We prepared an overnight culture of CG10 and CKU in LB broth in presence of spectinomycin (100 ug/ml), ampicillin (200ug/ml), diaminopimelic acid (DAP) (0.3mM) and glucose (10mg/ml). The later was added for repression of pBAD promoter. From the overnight culture, 250 ul were added to 10ml of LB broth with DAP(0.3mM), ampicillin (200ug/ml), arabinose (2mg/ml) for pBAD promoter induction and IPTG (0.8mM) for *dap*A gene induction which is under the control of P_{lac} promoter. The induction was done for six hours and the induced cells were inoculated on LB plates with spectinomycin and IPTG in presence and absence of DAP. The percentage of colonies growing in absence of DAP was compared to that growing in its presence (Fig. 3.3). Strain C319 (containing pBAD with IntI1 (p3938) was used as a positive control in the quantitative excision assay, whereas B548 strain was used as a negative control [174].

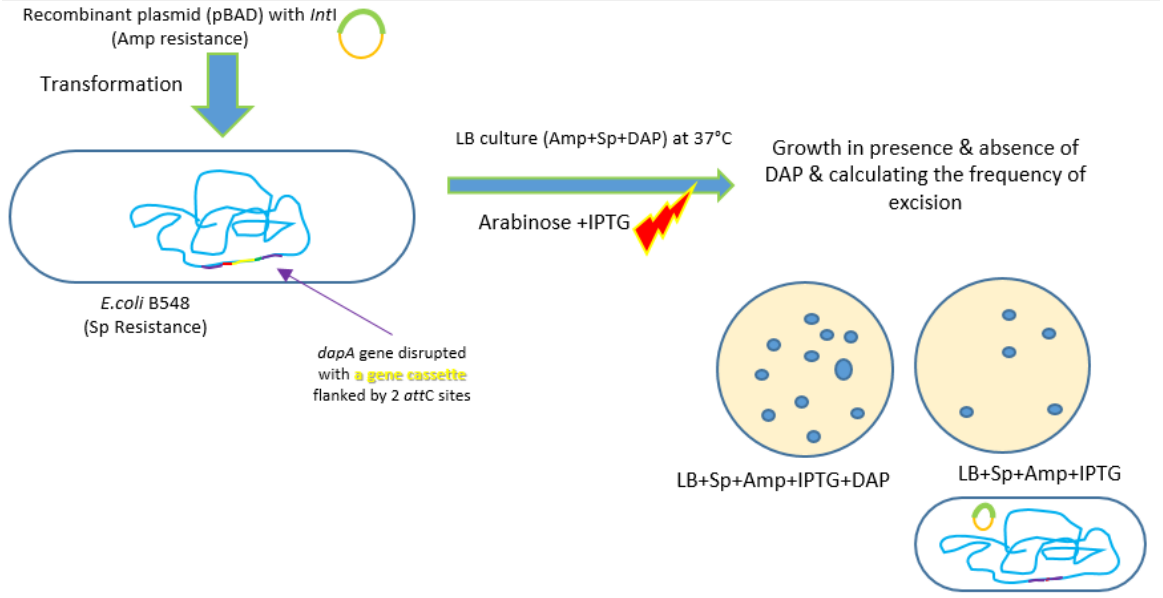


Fig.3.3 ***Quantitative excision assay.*** Recombinant plasmids with intI genes were transformed into Sp-resistant *E. coli* B548 (with a gene cassette-interrupted *dap*A gene Excision of the cassette restores a functional *dap*A gene allowing the strain to grow in a DAP-free medium. Abbreviations: Sp: Spectinomycin, Amp: Ampicillin, DAP: Diaminopimelic acid, IPTG: Isopropyl β-D-1-thiogalactopyranoside.

3.3. Results

3.3.1. Two positive results obtained in AGH library

We got bands at expected sizes (484, 404 and 363 bp) upon using int1.F with int-948R, intI.F with intI-864R and Int528F with intI-864R, respectively on AGH metagenomic DNA

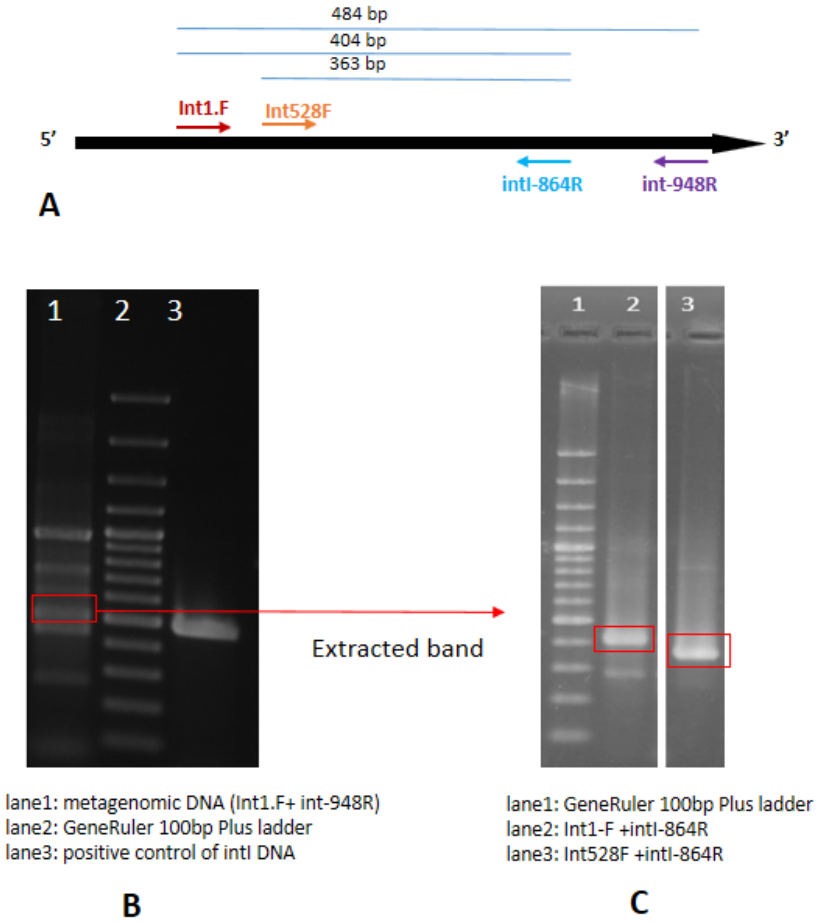


Fig.3.4 **PCR screening for *int*I detection**. A: Schematic diagram of *int*I gene showing relative positions of used primers and sizes of expected amplified bands. B: Gel showing PCR amplification results using IntI.F and int-948R primers on AGH metagenomic DNA. B: gel showing amplification results using Int1.F with int864R and int528F with int-864R on extracted band in Gel A.

(Fig.3.4). Sequenced bands confirmed the presence of *int*I genes within the metagenome. Thus, we have constructed a library of 4,556 clones from Aghormy Lake in Siwa Oasis (AGH). PCR screening for detection of *int*I genes in this library and in ATI LCL library (10,656 clones) resulted in two positive results only in AGH library and no positives in ATII LCL library. The two positives in AGH library were named 1G10 and 32A4 (Fig.3.5).

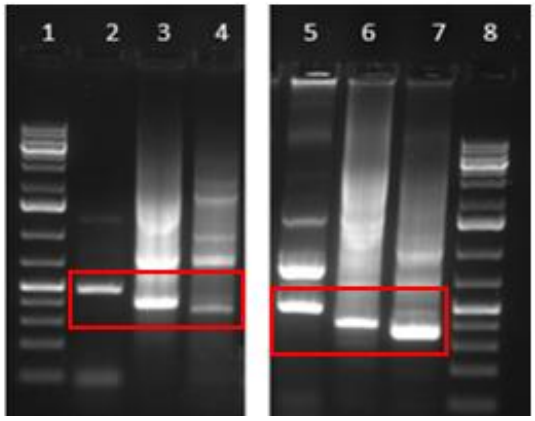


Fig.3.5 **Gel showing positive PCR screening results on two clones in AGH library.** Lanes 1 & 8: GenRuler 100bp Plus ladder, Lane2: 32A4 (Int1.F+int-948R), Lane3: 32A4 (Int1.F+ intI-864R). Lane4: 32A4 (Int528F+intI-864R), Lane5: 1G10 (Int1.F+int-948R), Lane6: 1G10 (Int1.F+ intI-864R). Lane7: 1G10 (Int528F+intI-864R).

3.3.2. IntIs detected in Kebrit Deep Upper interface (KD UINF)

Among all examined Red Sea brine pools, interfaces and sediments, six IntIs were detected in Kebrit Deep Upper interface (KD UINF) only. The six *int*I genes were identified in contig00306, contig01002, contig06491, contig12234, contig17426 and contig20623 (Appendix B: TableS5.2). All detected genes were partial due to the small sizes of contigs, except for the IntIs in contig00306 and contig06491. The first (KD UINF306) was further analyzed as the second was at the periphery of the contig preventing deeper analysis.

3.3.3. Identification of AGH-1G10 and KD UINF-306 integron components with no measured excision activity for both IntIs

Sequencing of AGH-1G10 fosmid revealed the presence of a complete integron (Fig.3.6) within a 24,734 bp contig. Blastp of the intI sequence (Fig.3.7) showed 86% similarity to an IntI

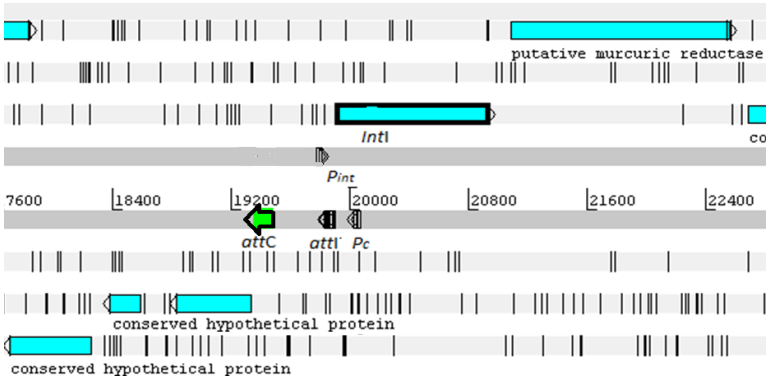


Fig.3.6 **AGH-1G10 integron.** Schematic representation using artemis software for the complete integron of AGH-1G10 showing all required components for a functional integron.

from a *Salinibacter* sp. (Fig.3.8). The acidic residues in the protein sequence were 12.5% of the total residues.

>AGH-1G10
MASSSSPDSCSSSSSSFLDRVRAACRRKGYTYRTEKTYLRWIVRYVK
YHGTEHPREFGKEEVRDYLSHLATDRNVAASTQNQALNALLFLHR
DVLGAEWDGVSDFDRAQEPERLPVVLTQEEVKELLGEMEGPNGL
VAHLLYGAGLRLSEALRLRVKDLDFDYEQITVRQGKGKKDRRTLLPG
MLIGSLRRQLRKSKAIWKEDLEAGYGTVSMPKALARKYPNAATEW
GWQYVFPSVRRSKDPRSGDIKRHHRSPSAVQKAVKRAVDATDISKS
ASCHTLRHSFATHLLEQGTDIRTVQELLGHRDLRTTQVYTHVLQDG
QAGTRSPLEGLGADVT

Fig.3.7 **Amino acid sequence of identified AGH-1G10 IntI**

integron integrase [Salinibacter sp. 10B]

Sequence ID: WP_105014151.1 Length: 352 Number of Matches: 1

Range 1: 28 to 352 GenPept Graphics ▼ Next Match ▲ Pr

Score	Expect	Method	Identities	Positives	Gaps
528 bits(1359)	0.0	Compositional matrix adjust.	258/325(79%)	281/325(86%)	0/325(0%)

```
Query  8    DSCSSSSSSFLDRVRAACRRKGYTYRTEKTYLRWIVRYVKYHGTEHPREFGKEEVRDYLS  67
            D   + SS FL RVRAACRR GYTYRTE+TY RWIVRYVKYH T HP + GKEEVR YLS
Sbjct  28   DRPGNGSSDFLGRVRAACRRSGYTYRTEQTYTRWIVRYVKYHNTRHPSKMGKEEVRKYLS  87

Query  68   HLATDRNVAASTQNQALNALLFLHRDVLGAEWDGVSDFDRAQEPERLPVVLTQEEVKELL  127
            +LAT R VAASTQNQALNALLFL+RDVLG EWD ++DF+RA EPERLP+VL++EE + LL
Sbjct  88   YLATKRRVAASTQNQALNALLFLYRDVLGREWDEITDFERANEPERLPIVLSEEETRALL  147

Query  128  GEMEGPNGLVAHLLYGAGLRLSEALRLRVKDLDFDYEQITVRQGKGKKDRRTLLPGMLIG  187
            GEMEG NGLVAHLLYGAGLRLSEALRLRVKDLDF YEQITVRQGKGKKDRRT+LP  L
Sbjct  148  GEMEGTNGLVAHLLYGAGLRLSEALRLRVKDLDFGYEQITVRQGKGKKDRRTILPDPLEA  207

Query  188  SLRRQLRKSKAIWKEDLEAGYGTVSMPKALARKYPNAATEWGWQYVFPSVRRSKDPRSGD  247
             LRRQL+KS+AIW+EDLEAGYG  SMP ALARKY NAATEW WQYVFPS RRS+DPRSGD
Sbjct  208  PLRRQLQKSEAIWREDLEAGYGQASMPLALARKYLNAATEWKWQYVFPSSRRSEDPRSGD  267

Query  248  IKRHHRSPSAVQKAVKRAVDATDISKSASCHTLRHSFATHLLEQGTDIRTVQELLGHRDL  307
            IKRHHRSPSAVQKAVK+AV    I+K AS HTLRHSFATHLL+ GTDIRTVQELLGH DL
Sbjct  268  IKRHHRSPSAVQKAVKQAVRDAGITKPASPHTLRHSFATHLLKHGTDIRTVQELLGHEDL  327

Query  308  RTTQVYTHVLQDGQAGTRSPLEGLG  332
            RTTQ+YTHVLQ G+AGTRSPL  +G
Sbjct  328  RTTQIYTHVLQKGKAGTRSPLSIIG  352
```

Fig.3.8 **Alignment of AGH-1G10 IntI sequence with Blastp first hit** (*Salinibacter* sp. IntI) showing high similarity (86%).

Manually, a possible *att*I site was detected: 5'-GCATAACGTTGTTATGC-3'. We detected one *att*C site where its bottom strand can form a hairpin structure typical of known *att*C sites (Fig.3.9). However, no ORFs were detected betwen the two recombination sites indicating the presence of an empty gene cassette. Putative P_{intI} (5'- CTGAAATACAGGCATTTGCGAAAA-3') and P_C (5'-TTGACGTAGCGGACAATCCAACGAAGGTACGTT-3') promoters were detected as well.

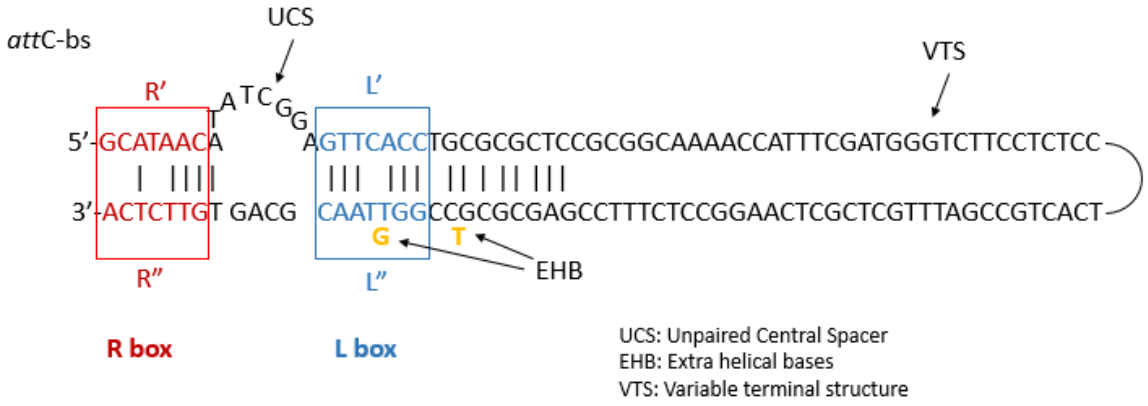


Fig.3.9 **Secondary structure of identified *att*C bs in AGH-1G10 integron** showing R and L boxes with EHBs at L".

KD UINF306 IntI (Fig.3.10) showed 9.6% acidic residues. A putative *att*I was identified: 5'-GTTTAAATGTTGTTCAAC-3'; however, no *att*C sites were detected. In fact, a long operon lies directly downstream the detected *att*I site until it reaches the contig periphery (Fig.3.11). It is thus unclear whether such a long operon could be a gene cassette or not. Proteins encoded by ORFS within this operon could be involved in lipopolysaccharide biosynthesis (Fig.3.11 and Appendix B: TableS5.2). Putative P_{intI} (5'-GTGCAAGACTTAAGCTTAGTTAAGTTTTTATAGT-3') and P_C (5'-TTAACTAAGCTTAAGTCTTGCACTGTTATTAT-3') promoters were detected as well. KD UINF306 intI and upstream ORFs have shown greater than 90% similarities with proteins from *Candidatus Marinimicrobia* bacterium.

For both tested intIs: AGH-1G10 and KU UINF-306, excision assay results showed no activity compared to negative and positive controls.

>KD-UINF306
MGAGRIGKVRLLLNVTKYNLNQHLREPKLLDRVNNEIQTRHYSRKT
GKTYRSWIKQFILYHHKQHPSKLGEVEINQFLSYLATEKHVSASTQN
QALSALLFLYKYVLHKELGDFGDVIRAKRSKKIPVVFTQDEVRSILKH
LKDEKQLMASLLYGSGLRLTECLRLRVKDVDFDNKQIIVRDGKGEKD
RVTLLSKKIIPHIKKHLSGVRKIYKADSKEGIGTTNIPYALERKYPTIAK
EWHWAYVFPSTKHAADKQTGELKRHHLNESVLQRAVKNAVKLAN
VEKHGGCHTFRHSFATHLLEAGYDIRTIQELLGHKKLETTMVYTHV
MNKGPMGVKSPGDTL

Fig.3.10 **Amino acid sequence of identified KD UINF306 IntI**

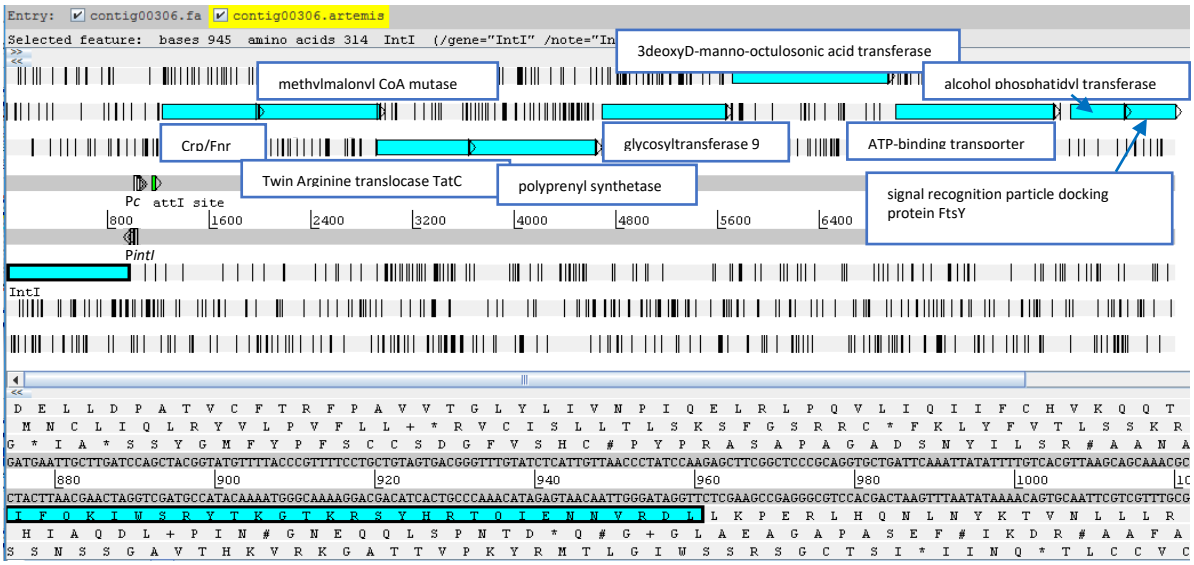


Fig.3.11 **KD UINF306 integron.** Schematic representation of KD UINF306 integron using artemis software with all necessary components for a functional integron and annotation of ORFs within the integron.

3.4. Discussion

Integrons have been identified in different environments [39] and have proven to be extremely diverse and not limited to few well studied classes [39,41,49]. We expected to get a high diversity of intI genes among our samples when using degenerate primers described in a study on hydrothermal vents [162]. Upon using degenerate primers: int480F with int-948R, we obtained no positive results at all; however, when we used int1.F primer [161] instead with int-948R , we got positive results with AGH metagenomic DNA and fosmid library. Still, getting only two positive clones within the whole library is considered extremely low, compared to the documented prevalence of integrons among different bacterial genomes [39,50]. It is not clear whether this is due to using a primer that is more likely to capture class 1 IntIs (int1.F) or because of possible sample enrichment with microorganisms that are devoid of integrons. For instance, integron integrases, integrons and related structures such as CALINs are totally absent in α-proteobacteria [50]. Moreover, no integrons have been reported in archaea which could be highly abundant in hypersaline environments. Taxonomic analysis of samples would reveal major taxa in these metagenomes and whether they contain any reported integrons.

AGH-1G10 has shown a relatively high percentage of acidic residues (12.5%) compared to many IntIs in publically available databases (data not shown) and the high similarity to *Salinibacter* sp. IntIs, increases the chances of this IntI to be from a halophile that uses a salt-in strategy, accumulating KCl inside the cell, thus requiring adaptation of the whole enzymatic machinery to the high intracellular salt concentration.

Integrons have a great biotechnological potential. For instance, they can be used for the incorporation of functional gene cassettes into plasmids and chromosomes. The increasing number of reported integrons with their new integron integrases can be exploited in different biotechnological application. However, the activity of most of these IntIs have never been assessed which limits their potential future exploitation. Developing excision and integration assays for emerging IntIs would thus be very useful. It should be noted that many factors could have an effect on the recombination activity of an IntI, such as the host, the used *att*I and *att*C sites and the nature of the recombination reaction.

Although we did not measure any activity for both AGH-1G10 and KD UINF-306 IntIs, using an excision activity assay, we cannot exclude the possibility of the proteins being active. The used assay is mainly developed for IntI1 [174]. Although different IntIs can identify different *att*C sites, some intIs were unable to identify certain *att*C sites [53]. Excision efficiency is also known to be affected not only by the cassettes own *att*C site, but also by its upstream neighbouring attachment site and whether it is an *att*I or an *att*C site [85]. AGH-1G10 probably belongs to a *Salinibacter* related species, whereas KD UINF-306 IntI most probably belongs to Candidatus Marinimicrobia bacterium on which no studies on integrons were done. It is also reported that some intIs -for instance, VchIntIA- are more active within their original hosts although no known accessory proteins or host factors affecting recombination frequencies are known to the moment [39]. The recombination efficiency may also be different for the same IntI depending on the nature of the reaction and whether it is an integration or an excision. The high diversity of *att*C sites makes it more likely that an IntI would identify an *att*C site rather than a non-cognate *att*I site, thus we tried an excision assay rather than an integration assay. However, we still did not observe a recombination activity which could be for reasons explained above. Using recombinant strains with putative detected *att*I and *att*C sites in the examined metagenomes, would facilitate further exploration of the recombination activity for theses IntIs. On the other hand, the presence of complete integrons with defective integrons does not impede the expression of gene cassette ORFs within the integron, such as in the majority of class 2 integrons that have a detective IntI with an internal stop codon [175]. In chromosomal integrons in general, bottom strands of *att*C sites are more likely found on leading strands rather than lagging strands limiting their availability as ssDNA segments and thus limiting their involvement in recombination reactions [45].

3.5. Conclusions

We have identified integrons in two different hypersaline environments: Aghormy Lake in Siwa Oasis in the Western Desert in Egypt and Kebrit Deep Brine Upper interface layer in the Red Sea. However, the low number of detected integrons and the absence of integrons in other examined Red Sea brines might be due to limitations of used primers and limitations of

metagenomic studies in general as the small sizes of obtained contigs. We did not detect any recombination activity for the two tested IntIs using an excision assay. This does not necessitate the absence of recombination activity as the used recombination sites, the host and the tested recombination activity could all have an effect on the assay results. More specific assays developed for each IntI might have a clearer picture on its activity.

Chapter 4: Abundance of integrons and CALINs in halophilic bacteria and the identification of integrons in archaea

Abstract

Integrons are genetic platforms used for expressing open reading frames (ORFs) arranged in gene cassettes. Excision and integration of gene cassettes is controlled by their associated integron integrase (IntI). Using IntegronFinder software, we analyzed all complete halophilic genomes available in the HaloDom database, along with selected partial halophilic genomes. We identified 19 new complete integrons and 44 clusters of *att*C sites lacking a neighboring integron-integrase (CALINs). Different classes of insertion sequences (ISs) were also identified within and nearby integrons and CALINs; with the abundance of IS*1182* elements and different ISs that can presumably mobilize adjacent genetic structures. Several promoters for intI genes (P_{intI}) showed nearby binding sites for arginine repressors (ArgR), raising possible regulation of IntIs expression and recombination activity by these proteins. Additionally, archaeal integrons were identified within a halophilic Natrialbaceae and a thermophilic Euryarchaeota. The high similarity of the Natrialbaceae IntI to another identified metagenomic IntI from a hypersaline environment would indicate its possible horizontal acquisition. Our findings reveal the existence of new integrons in halophilic bacteria and archaea with possible roles in adaptation to hypersalinity.

4.1. Introduction

Integrons are genetic elements where different open reading frames (ORFs) are captured and expressed according to environmental conditions [40]. These unique elements were first discovered as systems associated with antibiotic resistance in pathogenic bacteria [41]. Further explorations revealed that integrons are genetic elements harboring diverse gene cassettes, many of which encode for different adaptive proteins, and are widely spread among different bacterial phyla in many environments [41].

An integron is composed of a functional platform with all required elements for system operation and an array of gene cassettes each composed of an ORF followed by an *att*C recombination site. The functional platform is composed of: (1) *Int*I gene which encodes an integron integrase (IntI), (2) a recombination site termed *att*I site and (3) a promotor (P_C) for transcription of promoterless gene cassettes [39] (Chapter 1: Fig.1.1).

Integration and excision of gene cassettes are catalyzed by the IntI protein. Circular gene cassettes can be integrated within an integron by a site-specific recombination between *att*C site within the cassette and *att*I site within the integron which is located at the 5'end of the *Int*I gene; thus, the new integrated cassette will be positioned as the first associated gene cassette in the

integron where it can be expressed by the P_C promoter. This process is reversible as the gene cassette can be excised from the integron by a recombination event between its flanking *att*C sites [39,41] (Chapter 1: Fig.1.2).

IntIs are members of the site-specific tyrosine recombinase family. All members of this family are characterized by the presence of two conserved regions named box I and box II, with 4 highly conserved residues: R within box I and R-H-Y within box II. The conserved tyrosine residue in box II is essential for catalyzing the recombination reaction. Another short conserved motifs, named patches I, II and II, were also identified in tyrosine recombinases [38,46]. Upon further analysis of IntIs, they were found to possess an additional domain around patch III which was not detected in all other tyrosine recombinases [38] termed IntI patch [46]. This domain possesses an alpha helix named I2 and functions in the folding of the hydrophobic pockets essential for stabilizing two extrahelical bases (EHBs) in *att*C site bottom strand [47]. The core site of *att*I recombination site is minimally composed of two IntI binding sites termed R and L that form imperfect inverted repeats, where R has the consensus sequence of 5'-GTTRRRY-3' while the L site is highly degenerate. Recombination occurs between G and TT in the conserved triplet GTT within the R site (Chapter 1: Fig.1.2). IntI can recognize its cognate *att*I site; however, identification of *att*I sites from other integron classes was observed but with much lower rate [39]. The structure of the *att*C site is more complex when compared to the *att*I site. It is composed of 4 binding domains R", L", L' and R', where L" and L' are separated by a central region that varies greatly in sequence and length. The only conserved domains are the R" and R' sites with the consensus of 5'-RYYYACC-3' and 5'-GTTRRRY-3', respectively [39,41] (Chapter 1: Fig.1.2).

The lack of conservation among *att*C sites renders their identification challenging. This raised the questions on the mechanism of recognition of different *att*C sites by the same IntI. Crystallization of an integron integrase with its attached *att*C site revealed that *att*C site interacts with IntI by its bottom strand, only after the formation of a hairpin loop secondary structure [47]. The bottom strand is recognized by two EHBs at L" [39,47]. Thus, it has been proven that the secondary structure of *att*C is more critical than its primary sequence for proper recombination [47]. Unlike recombination with other tyrosine recombinases, in recombination catalyzed by IntI, single strand exchange occurs and the formed Holliday Junction intermediate needs to be resolved by a replication step [41].

Expression of gene cassettes is driven by P_C promoter located commonly within *int*I gene or within *att*I site. Different variants of P_C promoters were identified with those of class 1 integrons being the most extensively studied [39,41]. Moreover, it was found that the expression levels decrease as gene cassettes become more distal from the P_C promoter. However, some gene cassettes were found to carry their own promoters [39,41]. Expression of *int*I genes, at least in some integrons, were found to be regulated by SOS response based on the finding of LexA

binding sites overlapping the P_{intI} promotor. LexA is known as a transcriptional repressor for the SOS response [39,41]. This suggests that rearrangements, excision and integration of gene cassettes are driven by response to external stresses [63].

Integrons were first classified as either mobile integrons associated with transposons, thus can be transferred by conjugative plasmids, or as chromosomal integrons with long arrays of gene cassettes. However, further discoveries showed that intermediate forms between these two extremes do exist [39,41] as chromosomal integrons might be found within mobile elements [127] and could be associated with short arrays of gene cassettes [39,176].

Class I integrons were commonly found in clinical isolates, but later on they were detected in different environments with different degrees of urbanization [59]. Integrons were isolated from numerous environments such as desert soil, forest soil, hot springs, estuaries [41], polar sediments [61] and marine environments [63]. Analysis of gene cassettes associated with identified integrons revealed that the vast majority of gene cassettes encode for proteins of unknown functions [41].

Identifying integron gene cassettes proved to be hard because of the high diversity of *att*C sites [50]. Based on *att*C sites restricted to certain integron classes, different computational programs were developed [69,79,80]. Later on, IntegronFinder program was established, which can identify any IntI based on an HMM profile and *att*C sites based on a covariance model which is able to identify true *att*C sites based on their secondary structure with high sensitivity and specificity [50]. IntegronFinder pipeline can also annotate *att*I sites, P_{intI} and P_C promoters for integron classes 1, 2 and 3, in addition to pre-defined ARG cassettes [50]. Using IntegronFinder, clusters of *att*C sites lacking a neighboring integron integrase (CALINs) were found to be abundant in bacterial genomes [50]. It is hypothesized that these CALINS may have arisen as a result of chromosomal rearrangements separating *int*I genes from their adjacent gene cassettes in genomes that encode intI [50]. Chromosomal rearrangements could be induced by insertion sequences (ISs) [177,178]. In fact, many integrons were found to be associated with transposable elements such as ISs, transposons and conjugative plasmids, although these elements were not always functional [58]. Insertion sequence common region (ISCR) elements are found to be embedded within complex class 1 integrons [117]. Working on 2484 bacterial genomes, Cury et al had found that 12% of CALINs and 23% of complete integrons had internal IS elements; whereas, 38% of CALINS had adjacent IS elements [50].

Hypersaline aquatic environments are interesting habitats that require special adaptation measures by microorganisms living there to tolerate the high salt concentrations. Adaptation could either be done by a salt-in strategy or organic-solutes-in (salt-out) strategy. The first strategy is based on intracellular accumulation of high molar concentrations of KCl. Nevertheless, this necessitates the adaptation of all cellular machinery to high salt concentrations and an

increase in the acidity of the proteins produced inside the cell [1]. This could be observed in aerobic halophilic archaea mainly and in few bacterial species such as *Salinibacter ruber* [1]. The second strategy is based on expelling extra salts to the outside and accumulating organic solutes "osmolytes" such as glycine, betaine and sugars inside the cell [1]. Most halophilic bacteria and halophilic methanogenic archaea use the later strategy [1]. Understanding integron systems in hypersaline aquatic environments would be a great addition to our knowledge on microbial adaptation under extreme conditions.

In this study, we have analyzed integrons and CALINs in all publically available complete halophilic bacterial (45) and archaeal (41) genomes in HaloDom database [179], and in selected partial bacterial (35) and archaeal (100) genomes. We identified novel complete integrons within 17.5% of tested halophilic bacterial genomes. We have found that CALINs were more common than complete integrons (in 26.25% of examined halophilic bacterial genomes). Furthermore, we have observed a high frequency in arginine repressors ArgR and ArgR2 binding sites in putatively identified P_{intI} promoters. Moreover, for the first time, we report the presence of a complete integron in a halophilic archaeon and in another thermophilic one.

4.2. Materials and Methods

4.2.1. Analyzed samples

Our analysis included completely or partially sequenced genomes of halophilic bacteria (45 complete and 35 partial genomes with a total size of 287.48 Mb) and archaea (41 complete and 100 partial genomes with a total size of 486.787 Mb). Lists of complete halophilic bacteria, partially sequenced bacteria and archaea were obtained from the HaloDom database [179] in March 2021, November 2019 and September 2020, respectively: "halodom.bio.auth.gr" (Appendix A: TableS4.1 and TableS4.2) except for Natrialbaceae archaeon XQ-INN 246 which was directly obtained from NCBI database [180].

4.2.2. Identification of integrons and CALINs

IntegronFinder version 2.0 [50] was used to search for complete integrons, Integron integrase genes (*int*I) and CALINs in genomes of different halophiles. We used the option "local-max" on the command line with all genomes and contigs and an eight kb distance threshold between successive identified *att*C sites to ensure the detection of all potential *att*C sites. A positive result is reported when an *int*I gene and/or at least two *att*C sites are detected within the eight kb threshold. A search for integron cassette promoters (P_C), *att*I sites for known integron classes (1, 2 and 3) and known antibiotic resistance genes (ARGs) has been performed.

4.2.3. 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 P_{intI}, P_C promoters and promoters for toxin-antitoxin (TA) systems genes was done using bprom tool [173]. Visualization of sequences with identified ORFs and

*att*C sites was done using Unipro UGENE v1.19.0 [181]. Identification of transcription factors binding sites for detected P_{intI} in already known classes of IntIs (classes 1-5) was done using bprom (used accession numbers in Appendix A: TableS4.3)

4.2.4. Insertion sequences identification

ISEscan pipeline [182] was used to search for IS elements within halophilic genomes nearby integrons and CALINs. Further inspection of detected IS elements was done using blastn function on ISfinder [183] and comparing with curated IS elements. Identified complete and probably functional IS elements were submitted to ISfinder database (https://isfinder.biotoul.fr/).

4.3. Results

4.3.1. Organization of Integrons and CALINs in halophilic genomes

In 45 complete and 35 partial genomes of halophilic bacteria and 41 complete and 100 partial genomes of archaea, we detected a total of 19 complete integrons and 44 CALINs in 25 bacterial genomes (31.25% of examined halophilic bacterial genomes) (Table 4.1) and 1 complete integron in Natrialbacaea archaeon XQ-INN 246. All detected *int*I genes were parts of complete integrons. In bacterial genomes, integrons were confined to 14 genomes (17.5% of examined bacterial genomes), whereas CALINs were present in 21 genomes (26.25% of bacterial genomes) (Table 4.1). Ten genomes contained both integrons and CALINs (Table 4.1).

Table 4.1 Number of detected integrons and CALINs in Halophilic bacterial genomes

Bacterial analyzed genomes	Complete integrons	CALINs
Desulfohalobium retbaense DSM 5692	0	1
Chromohalobacter salexigens DSM 3043	0	3
Halorhodospira halochloris DSM 1059	0	1
Halomonas elongata DSM 2581	0	2
Halomonas titanicae ANRCS81	0	2
Marinobacter hydrocarbonoclasticus ATCC 49840	1	3
Marinobacter hydrocarbonoclasticus VT8	2	0
Nitrosococcus halophilus Nc 4	2	1
Salinibacter ruber DSM 13855	2	0
Halomonas huangheensis strain BJGMM-B45	0	1
Marinobacter salinus strain Hb8	1	1
Pseudomonas salegens strain CECT 8338	2	1

Chlorogloeopsis fritschii PCC 6912	0	5
Chromohalobacter japonicus CJ	1	1
Chromohalobacter japonicus SMB17	1	0
Ectothiorhodospira mobilis DSM 4180	0	3
Halomonas arcis CGMCC 1.6494	0	2
Halomonas halodenitrificans DSM 735	1	3
Halomonas meridiana ACAM 246	1	1
Halomonas saccharevitans CGMCC 1.6493	1	1
Halomonas subterranea CGMCC 1.6495	1	5
Salinovibrio costicola ATCC 33508 = LMG 11651	2	2
Salinovibrio costicola PRJEB21454	1	0
Salinivibrio costicola subsp. *alcaliphilus* strain DSM 16359	0	4
Salisaeta longa DSM 21114	0	1
Total	19	44

In *Marinobacter hydrocarbonoclasticus* ATCC 49840, the detected *int*I gene was interrupted by an IS3 element (IS*Maq2* isoform), in addition to a frameshift most probably rendering the protein inactive. In *S. ruber*, two identical IntIs were detected, but *IntI*-B showed an internal deletion missing patch II and the active site residue K174 [39] (Appendix A: TableS4.4).

All identified halophilic IntIs showed low identities to well-known IntIs ranging from 39-61% identities. We did not detect any increase in acidic residues (a range of 6.54-10.33% acidic residues) compared to IntIs from classes 1-4 (6.79 -10.95% acidic residues). The only exception was that of *S. ruber* intIs which have shown a high percentage of acidic residues (14.54% in IntI-A and 14.22% in IntI-B).

Our analysis on identified gene cassettes showed that the vast majority encode for hypothetical proteins or TA systems, regardless of the length of the gene cassette array (TableS2.4). No known ARG cassettes were identified by IntegronFinder in all analyzed genomes, except for a putative spectinomycin adenyltransferase in *Halomonas elongata* DSM 2581 CALIN (TableS2.4). All detected TA operons had their own promoters (Appendix A: TableS4.4) even if they have the same orientation of adjacent gene cassette arrays.

4.3.2. ArgR transcription factor binding sites abundant in putative halophilic P_{intI} promoters

We mined for putative P_{intI} and P_C promoters adjacent or within identified *int*I genes. LexA binding sites were detected in nine out of 21 putative P_{intI} promoters. We also detected an abundance of ArgR and ArgR2 binding sites in P_{intI} promoters (Appendix A: TableS4.4). In eight P_{intI} promoters: ArgR binding sites were found in two promoters, one ArgR2 in one promoter and both ArgR and ArgR2 binding sites in five promoters. LexA and ArgR or ArgR2 binding sites coexisted in five putative P_{intI} promoters (Appendix A: TableS4.4). Upon examination of sequences of P_{intI} promoters in other studied integrons (Appendix A: TableS4.3). ArgR binding sites were only detected in P_{intI} promoters of *Escherichia coli* and *Vibrio cholera*e class 2 integrons and in *Vibrio* sp. class 4 integrons (Appendix A: TableS4.3). No ArgR binding sites were detected in IntI1 and IntI3 P_{intI} promoters. The P_{intI} 1 promoter sequence is highly conserved [91], thus it is important to mention that the 18 inspected class 1 integrons showed the same promoter sequence even in presence of very few variations within the IntI1 sequences (Appendix A: TableS4.3).

4.3.3. Identification of IS elements within or nearby analyzed integrons and CALINs

Upon searching for IS elements within or nearby integrons or CALINs, we identified a great number of ISs from different families within complete or partial halophilic genomes. Coordinates of identified IS elements in each genome are shown in Appendix A: TableS4.4, and complete probably functional ISs were submitted to ISfinder database [183]. IS elements were found embedded or adjacent to 20 of the 64 integrons and CALINs identified within the examined halophilic genomes (31.7%) (Fig.4.1). Some IS elements were more common (Table 4.2) within and nearby integrons such as IS*1182* (27% of detected ISs) and those transposed by a rolling-circle replication mechanism and/or can presumably mobilize adjacent genomic elements (IS*91,* IS*1380 and* IS*200/605*) (15.5% of detected ISs) . Nine of the identified IS elements were with frameshifts within their transposases (Appendix A: TableS4.4); however, the four identified IS3 elements (Appendix A: TableS4.4) probably express their transposases using programmed -1 frameshifing as previously reported [99].

Certain genomes showed a high clustering of ISs within their integrons and CALINs, such as *Marinobacter hydrocarbonoclasticus* VT8 integron. The integron was packed with different IS elements; IS*91*, IS*1182*, IS*1380* and IS*21*, in addition to other ISs downstream integrons in the same genome (Fig.4.1 and Appendix A: TableS4.4). In *M. hydrocarbonoclasticus* ATCC 49840, the same IS*3* element (IS*Maq2*) existed in a complete integron, interrupting the *int*I gene, and within a CALIN in the genome (Fig.4.1 and Appendix A: TableS4.4). Five IS*1182* elements are present within and downstream *Marinobacter salinus* strain Hb8 integron, in which one contains

sindels and frameshifts and three are isoforms of the same IS (IS*Masa1*) with more than 99% identity (Fig.4.1 and Appendix A: TableS4.4). In Natrialbaceae archaeon XQ-INN 246 strain 2447, an IS*66* (IS*Narch2*) was found downstream of the integron with three other copies from the same IS identified in different locations throughout the archaeal genome (Fig.4.1 and Appendix A: TableS4.4).

Additionally, solitary transposases present as gene cassettes without being part of ISs were identified, in *Nitrosococcus halophilus* Nc4 and *Halomonas halodenitrificans* DSM 735 integrons, and within *Halomonas arcis* CGMCC 1.6494, *Halomonas meridiana* ACAM 246 and *Halomonas titanicae* ANRCS81 CALINs (Appendix A: TableS4.4).

Table 4.2 Distribution of different IS elements within or nearby integrons and CALINs in halophilic genomes.

IS type	**Number within or nearby integrons or CALINs**	**Number of genomes**
IS*1182*	12	4
IS*21*	3	3
IS*200/605*	1	1
IS*91*	4	4
IS*3*	4	2
IS*66*	2	2
IS*5*	1	1
IS*256*	2	2
IS*30*	4	1
IS*30*	4	1
IS*1380*	2	2
IS*As1*	1	1
IS*1634*	1	1
IS*4*	1	1
IS*110*	3	2

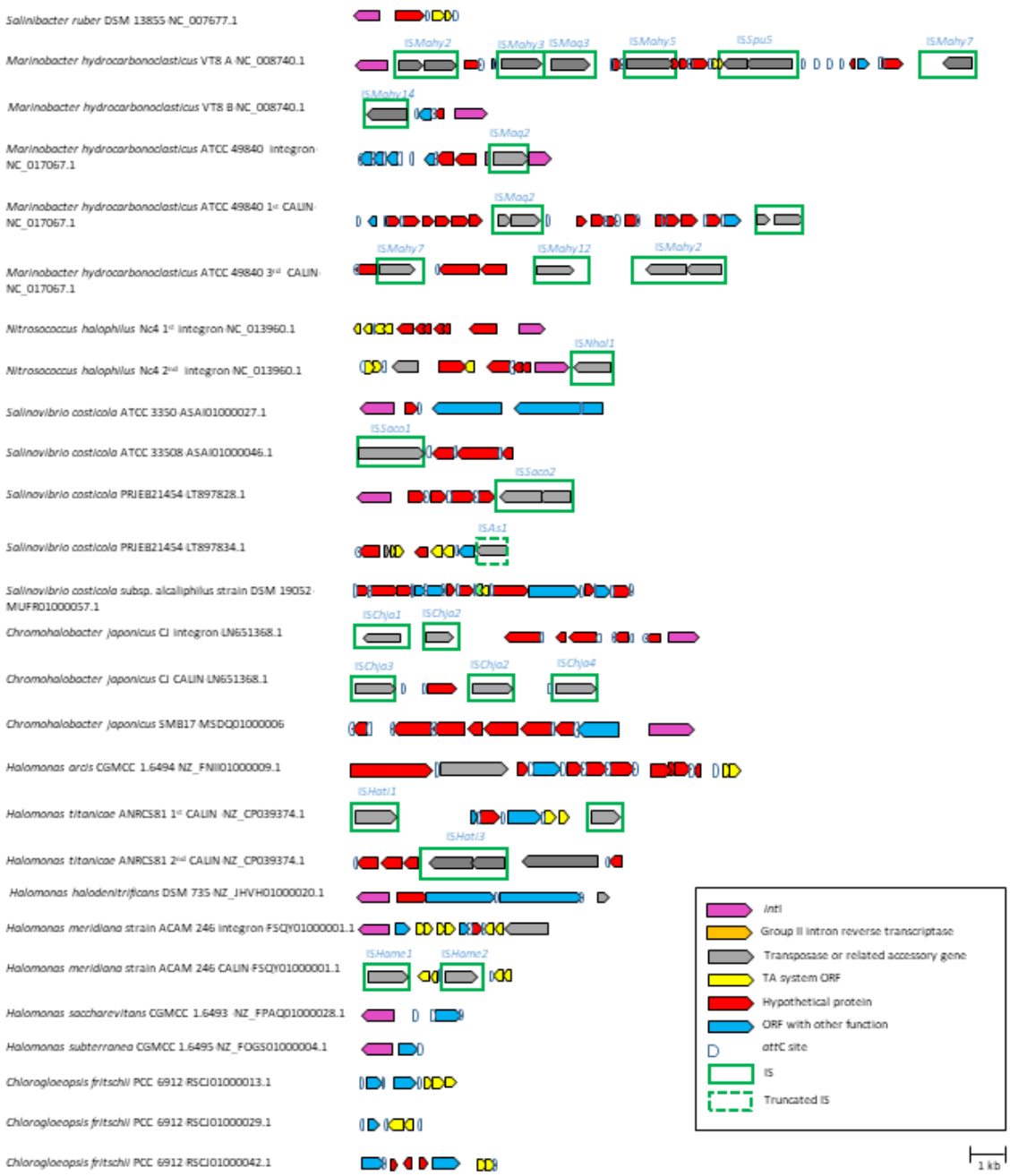


Fig.4.1 **Schematic diagrams of genetic components of integrons and CALINs in different halophilic genomes.** A colored key is provided within figure for explanation of different genetic components. All detected complete integrons with associated gene cassette ORFs are represented plus all CALINs with ISs and/or TA systems. No names were assigned to putatively defective IS elements with no known isoforms in ISfinder database. Genetic components are approximately to scale.

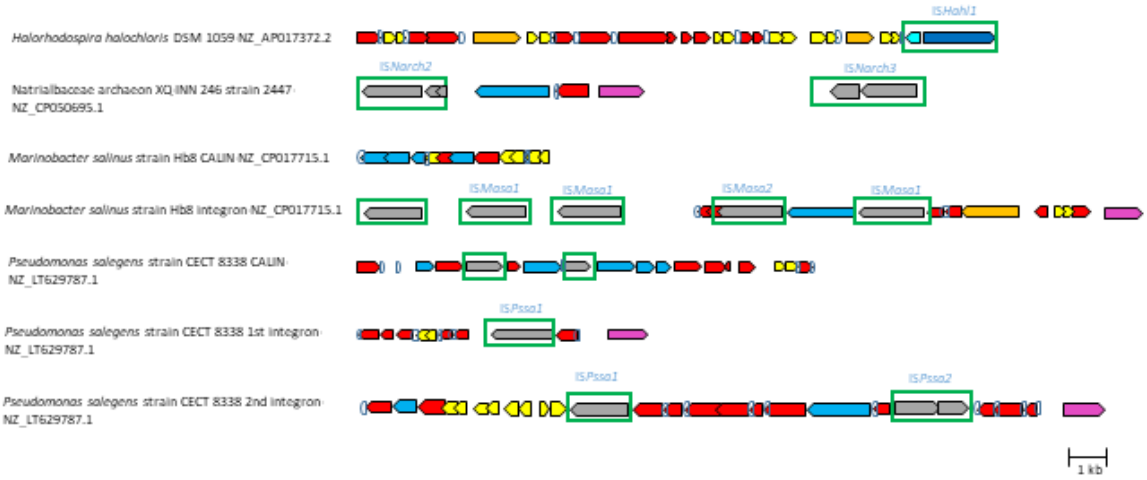


Fig.4.1. Continued

4.3.4. First reported archaeal complete integrons in halophilic Natrialbaceae archaeon and thermophilic Euryarchaeota archaeon

We identified a complete integron within the recently sequenced Natrialbaceae archaeon XQ-INN 246 (NZ_CP050695.1) from soda lakes and other hypersaline environments. This is the first reported complete integron in an archaeal genome. The identified IntI showed 92% similarity to an IntI that we have identified (Appendix B: Table S5.2) in a hypersaline soda lake metagenome (LFIK01016104.1). However, no halophilic characteristics are evident within this archaeal IntI and the percentage of acidic residues (8%) is almost similar to other known mesophilic IntIs (classes 1-4) as indicated above.

Upon inspection of intIs in archaea in general in the NCBI protein database, three other hits were found from metagenomic assemblies of archaeal isolates from marine hydrothermal vents; they were all marine and were not isolated from hypersaline environments. We have identified one intI in *Candidatus Aenigmarchaeota* archaeon B34_G1 metagenome scaffold from a Deep sea hydrothermal vent sediment in Guaymas Basin in Mexico (QMZW01000251.1) and two in Euryarchaeota archaeon isolate J059 from an iron-rich intertidal geothermal spring in Japan (RFHV01000400.1 and RFHV01000337.1). It is likely that the *int*I gene was partial in the latter because of the small size of the contig, which was found to be part of a complete integron with two ORFs in the first gene cassette (one hypothetical and an osmotically inducible oxidoreductase (OsmC family peroxiredoxin) (Appendix A: TableS4.4).

4.4. Discussion

4.4.1. New IntIs identified within halophilic genomes

Most research studies were directed towards integrons from bacterial pathogens and their role in the emergence of multiple resistant bacteria [48]. However, integrons are not only restricted to clinically relevant environments, but can also be found in almost all environments [45]. In our search for integrons in halophilic genomes, we have identified new types of *intI* genes, complete integrons and CALINs. None of the newly identified IntIs belonged to previous studied classes (class 1-4 integrons). This may indicate that they are either absent or poorly represented in halophiles in general. Unlike former surveys [39], in a wide study on integrons in complete bacterial genomes, integrons were found to be present in 7.2% of bacterial genomes and were reported to be absent in certain bacterial phyla such as Firmicutes and the class of α-proteobacteria [50]. However, our data shows a higher prevalence of integrons in halophiles (17.5%). This may indicate that integrons play a major role in adaptation to hypersaline environments. However, as the majority of detected gene cassette ORFs encode for hypothetical proteins, our ability to comprehend the role of integrons in adaptation to hypersalinity is impeded. The full potential of these gene cassettes will not be fully understood unless tested within their natural hosts. For instance, deletion assays of gene cassettes in *Vibrio rotiferianus* revealed an effect of a putative topoisomerase I-like protein in porin regulation, which may not have been detected if assayed in another host [184].

4.4.2. CALINs are more prevalent than complete integrons within halophilic genomes

Among our detected 64 complete integrons and CALINs in halophilic genomes, 44 CALINs were identified representing 69% of our positive results. This is higher than the recorded percentage of CALINs (56%) in a study on 2484 complete bacterial genomes [50]. The abundance of CALINs was then hypothesized to be the result of chromosomal rearrangements and/or integration of gene cassettes into secondary sites [50]. However, the rarity of these events could not account for the abundance of CALINs compared to complete integrons, especially in genomes devoid of *int*I genes. In our analysis, 44% of the genomes with positive results showed CALINs with no *int*I genes in the same genomes (Table 4.1). Perhaps, these arrays have unknown functions or they could act as reservoirs of gene cassettes on which *trans* acting IntIs can cause their mobilization. Nonetheless, chromosomal rearrangements could still partially explain the existence of CALINs especially in *int*I containing genomes. ISs could participate in chromosomal rearrangement events [177,178]; and they were found to be associated with integrons [12,58] and CALINs [12]. Thus, we have searched for IS elements within or adjacent to identified gene cassette arrays. IS elements were found within or adjacent to 28% of identified integrons and CALINs in halophilic genomes. It was intriguing to observe the clustering of

different ISs in the *M. hydrocarbonclasticus* VT8 integron. Although they belong to different IS families, they may still interact and recombine with closely related IS elements leading to an increase in genome plasticity and shuffling of genomic content. *M. hydrocarbonclasticus* ATCC 49840, which is another strain from the same species, shows different integrons and CALINs with different sets of gene cassettes and IS elements. This strengthens the notion of the increased plasticity in the *M. hydrocarbonclasticus* genome. In fact, a study has shown that gene cassette arrays can be naturally transferred between different bacterial species, followed by their chromosomal integration. Such process is facilitated by similarities in nearby mobile genetic elements that allow for homologous recombination [122]. Thus, it seems that abundance of integron and CALIN-adjacent IS elements would facilitate the transfer of gene cassettes not only intracellularly, but also horizontally between different bacterial species. Nevertheless, insufficient data concerning transposition mechanisms of different identified ISs, such as the highly abundant IS*1182* elements in our analysis, limits our ability to drive conclusions concerning their importance within or nearby integrons and related sequences. However, the presence of IS elements that can mobilize through a putative rolling- circle mechanism such as in IS*91* adjacent to integrons and CALINs may facilitate their mobilization as in case of IS*Mahy2,* IS*Saco2,* IS*Hati3* (Fig.4.1 and Appendix A: TableS4.4) and presumably by IS*Hahl1* [185] (Fig.4.1 and Appendix A: TableS4.4). Moreover, IS*Ecp1* and other similar ISs belonging to the IS*1380* family, were able to mobilize adjacent regions by an unknown mechanism [96]. In our data, we have identified two IS*1380* in *M. hydrocarbonclasticus* VT8 and *Chromohalobacter japonicus* CJ integrons (IS*Maq3* isoform and IS*Chja1*) and a partial IS*1380* transposase (found at contig's periphery) in *Halomonas halodenitrificans* integron. However, the possibility of these IS elements to mobilize adjacent cassettes needs to be verified experimentally.

4.4.3. Detection of archaeal integrons within halophilic and thermophilic archaea

It was reported that integrons have never been identified in archaea [63]; thus the presence of complete integrons within different archaeal species was surprising. The complete integron detected within Natrialbaceae archaeon XQ-INN 246 had only one gene cassette with a hypothetical protein ORF. The IntI showed high similarity with a hypersaline IntI from a soda lake metagenome (LFIK01016104.1). This similarity suggests its horizontal acquisition. Neither the archaeal integron, nor our identified halophilic IntIs had shown an increase in acidic residues, which is characteristic for proteins within halophilic microorganisms that adopt a salt-in strategy for high salt concentration adaptation [1]. This indicates that these IntIs are rather derived from halophilic microorganisms that use a salt-out strategy for high salt adaptation. This strategy, which is most common in halophilic bacteria, does not require modifying intracellular enzymes for adaptation to high salinity [1]. The only exception was the IntIs from *S. ruber* that have shown a higher percentage of acidic residues than other IntIs. This could be explained by the uniqueness

of *S. ruber* in adapting to high salt concentrations by a salt-in strategy that results in the accumulation of high molar intracellular concentrations of KCl [1].

The presence of an archaeal integron raised a question on whether we could find more archaeal integrons. However, we did not find *int*I genes, integrons or CALINs within the currently published halophilic archaeal genomes in the HaloDom database. Nonetheless, our search in NCBI protein database revealed the presence of additional three archaeal IntIs from metagenomic assemblies of archaeal isolates from marine hydrothermal vents. The short length of the contigs prevented further inspection of the genomic context of these *int*I genes; however, one of them was definitely part of a complete integron. These findings were interesting as they clearly show that integrons do exist, although rarely, within archaeal genomes. However, this implies a limited role in archaeal adaptation to their extreme environments.

4.4.4. Abundance of ArgR transcription factor binding sites in halophilic P_{intI} promoters

Regulation of IntI-activated recombination events has been shown to be affected by external stressors and controlled by LexA repressor that mediates an SOS response [39,88]. Thus, the abundance of LexA repressor binding sites within P_{intI} promoters was expected [39,88]. However, our results showed particular abundance of arginine repressors (ArgR1 and 2) binding sites in the vicinity of P_{intI} promoters. Coexistence of LexA and ArgR binding sites was observed in most of these predicted promoters. ArgR is a transcription factors that is arginine responsive and it regulates arginine biosynthesis, metabolism and transport, in addition to histidine transport [186], glutamate [187], lysine and ornithine biosynthesis [186] plus a suggested role in proline catabolism [188]. It was reported that a LexA dependent SOS response was induced in arginine-starved *E. coli* [189]. This only occurs in dividing cells and in conditions that allow high cyclic AMP (cAMP) production (presence of glycerol as a sole carbon source) [190]. It is worth mentioning that for resolving the formed Holliday junction in a recombination reaction mediated by an IntI, a replicative resolution step is required [191]. It is therefore likely that ArgR and LexA have a coordinated function in controlling integron recombination reactions in dividing cells in response to environmental stressors.

Furthermore, *att*C X *att*C intermolecular recombination reaction and *att*I X *att*I recombination reactions are known to be of much lower frequency than *att*I X *att*C and intramolecular *att*C X *att*C recombination reactions. Disfavoring these reactions was attributed to the formation of chromosome dimers that need special mechanisms for their resolution before cell division [39]. However, it is argued that the frequency of *att*I X *att*I reaction could be higher than what is observed under laboratory conditions especially with mobile integrons due to the presence of *att*I sites in a recombinogenic state [191]. As ArgR was found to function as an essential accessory protein for XerC/D in ColE plasmid dimer resolution [192], perhaps it has a

similar function with some IntIs to which XerC/D are the closest within the tyrosine recombinase family. It is possible that the presence of ArgR binding sites in P_{intI} promoters could only point out towards regulation of IntI expression by ArgR; however, it is intriguing to find that these regulators can also function as accessory proteins to a very close class of tyrosine recombinases. One could argue that IntIs do not require any accessory proteins and that they can function independently. Although this is correct with IntIs in mobile integrons such as IntI1 [39], other IntIs may require accessory proteins for optimum function. For instance, a 2600-fold higher rate of *att*I X *att*C recombination was observed in *V. cholerae* than in *E. coli* when a system derived from *V. cholerae* was used indicating that other host factors in *V. colerae* are required for optimum recombination reactions [39]. In fact, we found that intI4 promoters in *V. cholerae* have binding sites for both LexA and ArgR. Moreover, IntIA (IntI4) in *V. cholera* was found to be controlled by cAMP receptor protein (CRP) which is the main regulator of the carbon catabolite repression response important for adaptation of the cells to available carbon sources [45]. This indicates that IntI expression is affected by the host metabolism and the surrounding environment [45].

It is unclear whether ArgR proteins have a role in controlling IntI expression and regulating IntI-mediated recombination reactions in some environments. However, if that was proven to exist by experimental evidence, it would then add an extra layer of the complexity to the regulation of recombination reactions in integrons. It would be also interesting to inspect transcription factors binding sites in P_{intI} promoters from different environments to have a clearer image of their role in regulating recombination reactions.

4.5. Conclusions

Analyzing microbial halophilic genomes revealed the presence of novel integrons and CALINs, where CALINs were more abundant than integrons. Most ORFs, within gene cassettes, encode for proteins of unknown functions which impede further investigation of the role of these cassettes in adaptation to hypersaline aquatic environments. Furthermore, different IS elements within or nearby integrons and CALINs were identified. At least some of the identified types such as those moving by a rolling circle mechanism may have a role in mobilizing adjacent gene cassette arrays. We have also detected an increase in ArgR proteins binding sites within detected P_{intI} promoters, which may point out towards a role of these proteins in regulating IntI expression and/or function. Finally, the identification of archaeal integrons within a halophilic and a thermophilic archaeon for the first time indicates possible lateral transfer between microbial species. These findings suggest a role of integrons in bacterial adaptation to hypersaline environments and that more complex mechanisms could be involved in the regulation of integron integrase-mediated recombination reactions in aquatic environments. This role could be of limited importance in archaea, but it would be interesting to further study the role of integrons that are rarely found in archaea.

Chapter 5: Mining for integrons in hypersaline metagenomes

Abstract

Integrons are recombination platforms at which different gene cassettes can be excised, integrated and expressed. These recombination events are controlled by an integron integrase (Int) encoded by the *int*I gene within the integron. We analyzed different metagenomic assemblies from hypersaline aquatic environments using IntegronFinder. We identified 22 new *int*I sequences within hypersaline metagenomes. No gene cassettes with known antibiotic resistance genes (ARGs) were identified. The majority of gene cassette ORFs encode for hypothetical proteins, with abundance of Toxin-Antitoxin (TA) systems within and adjacent to integrons and clusters of *att*C sites lacking a neighboring integron-integrase (CALINs). All TA operons had their own promoters although the majority of them lied at the same orientation of adjacent cassettes. Insertion sequences (IS) were absent nearby detected integrons and CALINs. Finally, we detected atypical putative CALINs within archaeal metagenomes, showing arrays of successive *att*C-sites overlapping with archaeal ORFs. Our findings reveal the existence of new integrons in hypersaline environments that may have a role in adaptation to hypersalinity.

5.1. Introduction

The integron system has gained a lot of attention due to its ability to integrate, excise and shuffle different gene cassettes according to the need of the microorganism [40]. Although these systems were first associated with ARGs [41]. They were later found in different environments with different sets of gene cassettes, mostly of unknown functions [41]. A complete integron contains an *int*I gene that encodes for an integron integrase protein (IntI), an *att*I recombination site, most probably at the 5' end of the *int*I gene, and a P_C promoter followed by an array of gene cassettes. The P_C promoter drives the transcription of the promoterless gene cassettes, whereas, the intI gene has its own promoter P_{intI} [39]. A typical gene cassette is composed of an ORF followed by an attC recombination site [39]. Integration and excision of gene cassettes are all mediated by the IntI, which is a member of site-specific tyrosine recombinase family [38].

Class I integrons were commonly found in clinical isolates, but later on they were detected in many environments with different degrees of anthropogenic effect [59]. Most studies on integrons were based on cultured isolates [46,59,74]; however, metagenomics proved to be a great mine for isolation of different types of integrons [59,61,75,76]. Nevertheless, most of these

studies are based on PRC amplification either to amplify integron integrase genes [42,61,77], their cassettes [78] or both [59,75].

The high diversity in *att*C sites makes their identification challenging [50]. However, the highly sensitive and specific IntegronFinder pipeline was developed recently for identification of integrons and CALINs [50].

One of the most widely spread gene cassettes in chromosomal Super-Integrons (SIs), are those for type II Toxin-Antitoxin (TA) systems [39]. Those are addiction modules that can stabilize flanking regions in gene cassette arrays [69,70]. Typically, each module is composed of an upstream antitoxin gene followed by its cognate toxin gene arranged as an operon with its own promoter. Nonetheless, the arrangement could be inverted in some modules such as with higBA module [66,67]. In addition, the TA gene cassette could be oriented in an opposite orientation to adjacent gene cassettes [72].

Hypersaline aquatic environments are intriguing habitats that require special adaptation measures by microorganisms living there to tolerate the hypersalinity. The high plasticity in the integron systems may have a role in microbial adaptation to these extreme environments.

Here, we have analyzed integrons and CALINs in 28 previously assembled metagenomes (1,236,831,758 nucleotides and 658,054 contigs) from different hypersaline environments. We identified novel integron integrases (IntI)s and complete integrons within these environments. The identified CALINs were more common than complete integrons. Despite the many reports of presence of insertion sequences (IS) nearby integrons, we did not detect any adjacent to the identified integrons or CALINs, nor did we detect any known ARG cassettes. Nevertheless, TA systems were abundant in detected integrons and CALINs regardless to the size of the gene cassette array. We provide a more detailed account on detected TA systems. Finally, we detected atypical CALINs showing arrays of *att*C sites in archaeal sequences in the metagenome of Tanatar trona crystallizer in Russia and in *Caldivirga* spp. abundant in the metagenome of the hypersaline mat in Grendel Spring in Yellowstone National park, USA.

5.2. Materials and methods

5.2.1. Analyzed samples

Publicly available metagenomic assemblies from different hypersaline environments in addition to Red Sea brine pools metagenomes assembled in our lab (28 assemblies of a total of 1,236,831,758 bp and 658,054 contigs) (Table 5.1) were used in our analysis. Eight complete and partially sequenced *Caldivirga* archaeal species were analyzed in this study as well (Appendix B: TableS5.1)

Table 5.1 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, [193, 194]	8530607	11351
GNM2	Guerrero Negro mat, Mexico 1-2mm depth	GCA_000206565.1, [193, 194]	7390978	10551
GNM3	Guerrero Negro mat, Mexico 2-3mm depth	GCA_000206545.1, [193, 194]	8209846	11423
GNM4	Guerrero Negro mat, Mexico 3-4mm depth	GCA_000206525.1, [193, 194]	8130049	11724
GNM5	Guerrero Negro mat, Mexico 4-5mm depth	GCA_000206505.1, [193, 194]	9689398	14128
GNM6	Guerrero Negro mat, Mexico 5-6mm depth	GCA_000206485.1 , [193, 194]	8291075	11380
GNM7	Guerrero Negro mat, Mexico 6-10mm depth	GCA_000206465.1, [193, 194]	9759240	13649
GNM8	Guerrero Negro mat, Mexico 10-22mm depth	GCA_000206445.1, [193, 194]	7914434	11356
GNM9	Guerrero Negro mat, Mexico 22-34mm depth	GCA_000206425.1, [193, 194]	8308787	11596
GNM10	Guerrero Negro mat, Mexico 34-49mm depth	GCA_000206405.1, [193, 194]	7132956	10297
ATII SDM	Atlantis II Deep Brine Sediment, Red Sea	[6,167,168]	40413330	41726
DD SDM	Discovery Deep Brine Sediment, Red Sea	[6,167,168]	52421642	51829
Th	Thetis Mediterranean deep-sea hypersaline lakes	GCA_001684355.1	13102297	10347
ATII INF	Atlantis II Deep Brine interface, Red Sea	[164,168]	16014945	24317
DD INF	Discovery Deep Brine interface , Red Sea	[164,168]	11647401	18413
KD UINF	Kebrit Deep Upper interface, Red Sea	[164,168]	42652688	45750
KD LINF	Kebrit Deep Lower interface, Red Sea	[164,168]	50280352	74666
ATII LCL	Atlantis II Deep Brine, Lower convective layer, Red Sea	[164,168]	46518597	43555
ATII UCL	Atlantis II Deep Brine,Upper convective layer, Red Sea	[164,168]	21343827	29592
DD BR	Discovery Deep Brine , Red Sea	[164,168]	12244355	18850
KD BR	Kebrit Deep Brine, Red Sea	[164,168]	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, [195, 196]	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

5.2.2. Identification of integrons, CALINs, gene cassettes and all integron components

We used IntegronFinder version 2.0 [50] to search for complete integrons, Integron integrase genes and CALINs. Positive results were those with at least two successive *att*C sites within an 8 kb threshold. We annotated all predicted ORFs within the identified gene cassettes based on Blastx results against NCBI nr database. Search for P_{intI}, P_C promoters and promoters for TA systems genes was done using bprom tool [173]. Visualization of sequences with identified ORFs and *att*C sites was done using Unipro UGENE v1.19.0 [181]. We have searched for possible IS elements within and adjacent to detected integrons and CALINs using ISEscan pipeline [182].

5.3. Results

5.3.1. New integron integrases, complete integrons and CALINs identified within hypersaline metagenomes

In all examined metagenomes, most findings were CALINs (92 CALINs) rather than complete integrons (eight integrons) or solitary IntIs (18 solitary IntIs). Table 5.2 shows the number of contigs in all examined metagenomes and the number of those with positive results in each site.

Table 5.2 Number of contigs in each examined metagenomic assembly with the number of positive contigs and identified integrons, CALINs and IntIs.

metagenome	Number of contigs	positive contigs	Complete integrons-CALIN-intIs
GR	11,151	23	0-23-0
GNM1	11,351	3	0-3-0
GNM2	10,551	2	0-2-0
GNM3	11,423	4	0-2-2
GNM4	11,724	1	0-0-1
GNM5	14,128	5	0-4-1
GNM6	11,380	1	0-0-1
GNM7	13,649	5	0-3-2
GNM8	11,356	0	0-0-0
GNM9	11,596	3	0-2-1
GNM10	10,297	3	0-2-1
ATII SDM	41,726	6	0-6-0
DD SDM	51,829	3	0-3-0
Th	10,347	10	1-7-2
ATII INF	24,317	1	0-1-0
DD INF	18,413	1	0-1-0
KD UINF	45,750	15	1-9-5
KD LINF	74,666	5	0-5-0
ATII LCL	43,555	1	0-1-0
ATII UCL	29,592	0	0-0-0

DD BR	18,850	2	0-2-0
KD BR	74,666	3	0-3-0
TSL	19,350	11	5-5-1
TTCSL	9,426	5	0-5-0
PSL	25,098	2	1-0-1
Ty	15,008	2	0-2-0
Na	11,304	0	0-0-0
BSL	15,551	1	0-1-0

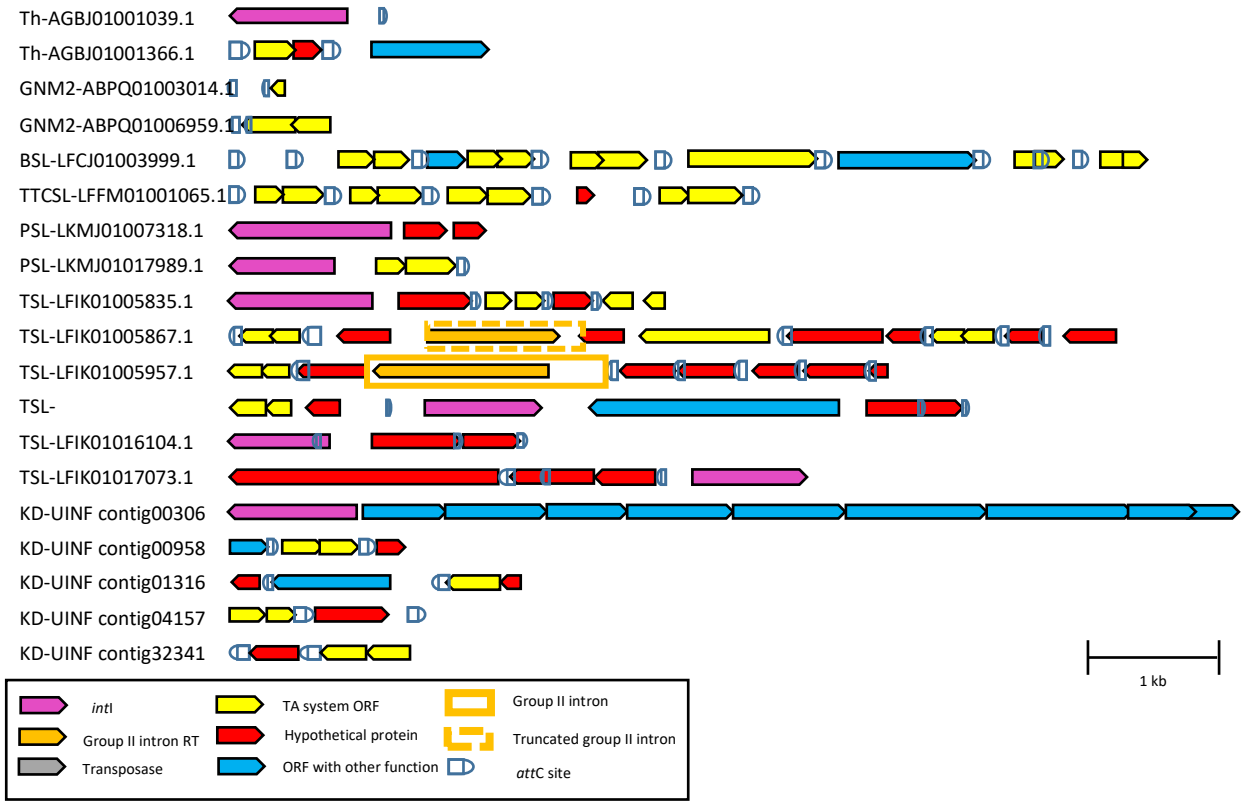


Fig.5.1 **Schematic diagrams of genetic components of integrons and CALINs in different metagenomic assemblies.** A colored key is provided within figure for explanation of different genetic components. All detected complete integrons with associated gene cassette ORFs are represented plus all CALINs with ISs and/or TA systems. Genetic components are approximately to scale.

The 18 identified solitary *int*I genes had no adjacent integron components, mainly because of the small sizes of contigs or existence near contig breaks. Multiple sequence alignment and blastx results showed that 13 identified IntIs were partial due to their presence at the edges of the assembled contigs or in very short contigs (Appendix B: TableS5.2). Thus, we were not able to detect all important domains for their putative recombination activity. For instance, box II region at the C-terminus was missing in IntIs from: Th-AGBJ01007148.1, GNM5-ABPT01000232.1 and KD UINF-contig12234 (which has also two frameshifts). Other identified *int*I genes showed one or more frameshifts within the ORF sequence casting doubt on their activity such as in case of: GNM7-ABPV01012279.1, GNM10-ABPY01004164.1, KD UINF-

contig12234, PSL-LKMJ01017989.1 and GNM3-ABPQ01010372.1. In the last two, in addition to the frameshifts detected, no box II regions were detected. Perhaps this is due to indels within their sequences.

We were able to identify eight complete integrons although some of the identified *int*I genes are probably pseudogenes as explained above (Appendix B: TableS5.2). All identified hypersaline IntIs are novel sequences with low identities to well-known IntIs (data not shown). IntI from the Tanatar-5 soda lake (TSL-LFIK01016104.1) showed high similarity (92%) to an IntI from a recently sequenced Natrialbaceae archaeon XQ-INN 246 strain 2447.

5.3.2. Neither known ARG cassettes nor IS elements were identified within and adjacent to identified integrons in hypersaline metagenomes

Our analysis of identified gene cassettes showed that the vast majority encode for hypothetical proteins (Appendix B: TableS5.2). No known ARG cassettes were identified by IntegronFinder in all analyzed metagenomes. However, putative betalactamase ORFs were found in TSL-LFIK01005867.1 CALIN and in Grendel Spring atypical CALIN GR-OFEH01000073.1 (Appendix B: TableS5.2). A GNAT family N-acetyltransferase was found in TSL-LFIK01005957.1 (Appendix B: TableS5.2). Aminoglycoside acetyltransferases belong to this family [197].

At the same time, we did not detect IS elements nearby integrons in hypersaline metagenomes.

5.3.3. TA systems are commonly found as gene cassettes or adjacent to CALINs and integrons regardless of length of the arrays

We have identified 22 putative Toxin-Antitoxin (TA) systems associated with integrons or CALINs. Table 5.3 summarizes the findings of the identified TA systems. More details are represented in Fig.5.1 and Appendix B: TableS5.2. TA systems were found to be common in most detected integrons and CALINs (Fig.5.1), regardless of the length of the gene cassette array. Sometimes, the presence of very short contigs hindered the search for complete TA systems. For instance, an orphan toxin in GNM2-ABPQ01003014.1 and an antitoxin in GNM3-ABPQ01010372.1 were detected; however, they could be parts of complete TA systems. Another putative orphan toxin (Fic protein) was found as a gene cassette in Bitter soda lake CALIN (BSL-LFCJ01003999.1), which is mainly composed of different TA system gene cassettes.

In general, all detected TA operons were at the same orientation of adjacent gene cassettes, except in a TA system in TSL-LFIK01005835.1 integron that lied directly downstream the last *att*C in the integron. In most detected TA systems, the antitoxin ORF was followed by a downstream toxin ORF. This is the common arrangement in the majority of TA operons [39,67].

On the other hand, in five TA systems, the toxin ORF was followed by a downstream antitoxin. The later arrangement was mainly found with *Brn*TA and *Hig*BA systems, which are normally present in this reverse arrangement [67]. We have also observed that in three cases, the TA system lies directly downstream the last *att*C in the integron or CALIN (Fig.5.1, Appendix B: TableS5.2, and [185]).

In Tanatar trona crystallizer metagenome (TTCSL- LFFM01001065.1), the first three cassettes are composed of a DUF344 domain-containing protein followed by a DUF5615-domaing containing protein (PIN-like domain). The same arrangement is seen in *Chlorogloeopsis fritschii* (contig RSCJ01000013.1) (Appendix A: TableS4.4). These proteins are related to uncharacterized VapBC45 proteins that are thought to form TA systems [198,199].

Table 5.3 A summary of identified types of TA systems within hypersaline metagenomes.

Contig	**type**	**order**
Th- AGBJ01001366.1	**BrnT T, hypothetical**	**T-> AT**
GNM2- ABPQ01006959.1	**HigB T, HTH (AT)**	**T-> AT**
GNM2- ABPQ01003014.1	HicA T (end of contig)	T
GNM3- ABPQ01010372.1	HicB AT	AT
PSL- LKMJ01017989.1	AT, VapC T (PIN)	AT ->T
TSL-LFIK01005867.1	YefM AT, Txe/YoeB T	AT ->T
	AT, RelE/ParE T	AT ->T
TSL-LFIK01005957.1	ParD-like AT, RelE/ParE T	AT ->T
TSL-LFIK01005835.1	AT, ParE T	AT ->T
	VapB AT, VapC T (normal but in inverted integron)	AT ->T
TSL-LFIK01007609.1	AT, T	AT ->T
BSL- LFCJ01003999.1	**RelE/ParE T, HigA AT**	**T-> AT**
	BrnT T, BrnA AT	**T-> AT**
	CopG AT, T (PIN)	AT ->T
	HicB AT, HicA T	AT ->T
	YefM AT, Txe/YoeB T	AT ->T
	Fic protein	T
TTCSL- LFFM01001065.1	DUF433, DUF5615	AT ->T
	DUF433, DUF5615	AT ->T
	DUF433, DUF5615	AT ->T
	Hypothetical, PIN domain	AT ->T
KD UINF-contig00958	BrnT T, BrnA AT	AT ->T
KD UINF-contig01316	Hypothetical, VapC T	AT ->T

KD UINF-contig03241	**RelE/ParE T, HigA AT(fs)**	**T-> AT**
KD UINF-contig04157	HicB AT, HicA T	AT ->T

5.3.4. Abundance of *att*C clusters in archaeal metagenomes from Grendel Spring belonging to *Caldivirga* sp.

Upon examination of metagenomes from different hypersaline environments, the majority of contigs showing positive results (i. e. the presence of complete integrons, CALINs or *int*Is) were most probably from bacterial sources. However, in one particular site: Grendel Spring hypersaline mat (GR) in Yellowstone National Park, Wyoming, USA (already known by its high archaeal content), positive archaeal contigs with neighbouring *att*C sites were detected. Three contigs in Tanatar trona crystallizer soda lake TTCSL (LFFM01001574.1, LFFM01002330.1 and LFFM01004875.1) showed neighboring *att*C sites with similarities to archaeal hits as well. However, in all of them, no typical gene cassettes were found. In case of Grendel lake metagenome, all detected *att*C sites overlapped with or laid within the identified ORFs which their blast hits showed high resemblance to or even 100% identity to archaeal proteins especially from different *Caldivirga* spp. Thus, we searched for a complete genome of a *Caldivirga* sp. We only found *Caldivirga maquilingensis* isolated from an acidic hot spring in the Philippines. It is a microaerophilic heterotroph and is able to use sulfur, thiosulfate, and sulfate as electron acceptors (this environment is of high salinity mainly due to iron and sulfate [200]). However, no CALINs were detected in this genome.

We analyzed several partially sequenced *Caldivirga* species (seven genomes) isolated from different Yellowstone National Park hot springs and other hot springs by IntegronFinder, we got similar results to those obtained from Grendel Spring metagenome (i. e. successive *att*C sites overlapping with ORFs, which makes it really hard to consider any of these ORFs as parts of gene cassettes). Some contigs from GR metagenome are clearly parts of larger contigs of some of the partially sequenced *Caldivirga* genomes as shown in Fig.5.2 and Appendix B: TableS5.2.

Fig.5.2 **Archaeal contigs within GR and TTCSL metagenomes** showing overlapping ORFs with putative attC sites. Genetic components are approximately to scale.

5.4. Discussion

5.4.1. New IntIs identified within hypersaline metagenomes with abundance of CALINs and absence of IS elements

It is now believed that integrons are widespread in almost all environments [45]. In our search for integrons in hypersaline metagenomes, we have identified new types of *intI* genes, complete integrons and CALINs. None of the newly identified intIs belonged to most studied classes (class 1-4 integrons). Although the nature of metagenomic studies may hinder the identification of all existing intIs, even the complete genomes didn't show any integrons from well-known classes. This may indicate the absence of these classes or their rarity in hypersaline environments.

Some of the identified *intI* genes are probably pseudogenes with frameshifts in their genomic sequences. However, it is unclear whether these frameshifts do actually exist or that they have resulted from errors in sequencing and assembly.

In our positive data (integrons, IntIs and CALINs), CALINs represented 78%. This is higher than the percentage of CALINs (56%) found on a study on 2484 complete bacterial genomes [50] and even higher than the percentage that we have found within halophilic genomes (69%) . This increase may be attributed to the nature of metagenomic contigs; these CALINs

could be actually parts of complete integrons that were not fully assembled. However, not all CALINs can be attributed to contig breaks, as they have been already reported in many complete genomes [50]. The function of CALINs is still unknown and perhaps they serve as reservoirs of gen cassettes [50]. However, it was suggested that these CALINs resulted from chromosomal rearrangements or integration of gene cassettes into secondary sites [50]. As ISs were found in association with integrons [50,58] and CALINs [50], and as they could have a role in chromosomal rearrangements [177,178]; we searched for ISs within or adjacent to identified gene cassette arrays. Surprisingly, no IS elements were found within or nearby examined metagenomes. However, this could be explained by the frequent existence of contig breaks at sequences of transposable elements [50].

5.4.2. TA systems abundance in integrons and CALINs regardless of the length of the gene cassette array

Previous studies have found that TA systems are mainly detected in long gene cassette arrays rather than short ones [39,71]. However, our results show abundance of TA systems in integrons and CALINs regardless of the length of these arrays. It has been also reported that most TA cassettes are oriented in opposite orientation to adjacent gene cassettes; thus they had their own promoters and their transcription cannot be controlled by P_C promoters [39]. Here, the majority of detected TA operons were at the same orientation of adjacent gene cassettes, but all of them had their own predicted promoters. This indicates that TA gene cassettes in general don't rely on P_C promoters for their transcription. Another observation is the positioning of some TA systems next to the last *att*C site in the gene cassette array (Fig.5.1 and Appendix B: TableS5.2), which means that they are most probably fixed in their positions and cannot be mobilized like adjacent gene cassettes. TA systems have shown to provide stability to different genomes and SIs [69,70]. This layout would actually lead to more genomic stability as the loss of the TA system would be more unlikely.

Orphan toxins and antitoxins were detected in our dataset as well. Different studies reported the presence of orphan toxins belonging to type II TA systems [201,202]. It is unknown however, whether these toxins could be regulated with related antitoxins within the genome. Within the CALIN detected in Bitter soda lake in Russia, we found a gene cassette with an ORF encoding for a putative orphan Fic protein toxin. This CALIN in particular is packed with different TA gene cassettes. Nonetheless, it is not clear whether this ORF is related to TA systems or not. Most Fic proteins are uncharacterized and those characterized showed different activities [203]. Some of them form toxins in TA systems with adenylylation function observed towards DNA gyrase and topoisomerase IV [204] or with phosphorylation function towards translation elongation factor EF-Tu [205]. However, unlike other TA systems Fic proteins are not found in operons [67]. In case of orphan antitoxins, one *hic*B antitoxin genes was identified in GNM3-

ABPQ01010372.1. However, it could be part of a TA operon as it lies at the periphery of the contig. Yet, it is hypothesized that orphan antitoxins may result from deletions in TA loci [187]. It has been also suggested that they serve new functions as anti-addiction modules, preventing MGEs integration [206] or that they interact with other toxins in TA pairs affecting their function [187].

5.4.3. Abundance of successive *att*C sites within some archaeal metagenomes

Another intriguing finding was the presence of atypical CALINs within archaeal metagenomes found in Grendel Spring in Yellowstone National Park in USA and in Tanatar trona crystallizer soda lake in Russia. These genomic structures showed successive *att*C sites overlapping with different ORFs and not showing the typical arrangement of a gene cassette. The identified ORFs in GR metagenome showed high similarities with ORFs from *Caldivirga* spp. Searching in the genomes of other *Caldivirga* spp. from different hot springs revealed the existence of the same arrangements of successive *att*C sites. Here, any recombination event would result in truncated ORFs. Thus, it is more likely that these arrays of successive *att*C-like structures have other unknown functions.

5.5. Conclusions

Analyzing different hypersaline metagenomes revealed the presence of novel *int*I genes, complete integrons and CALINs with more abundance of the latter. Most ORFs, within gene cassettes, encode for proteins of unknown functions which impede further investigation of the role of these cassettes in adaptation to hypersaline aquatic environments. However, different classes of type II TA systems as gene cassettes or adjacent to integrons and CALINs were extremely abundant supporting their role in stabilizing the integron systems. Finally, we have identified arrays of successive *att*C-sites within archaeal metagenomes and genomes, that do not resemble the typical structure of gene cassette arrays in CALINs. The role of such structures needs further investigation.

Chapter 6: Association of Group IIB Introns with integrons in hypersaline environments

Abstract

Group II introns are mobile genetic elements (MGEs) that can be used as gene targeting tools. They have the properties of both ribozymes and retroelements. So far, group IIC introns are the only class reported to be associated with integrons. Our aim was to study group II introns linked with integrons and CALINS (cluster of *att*C sites lacking a neighboring integron integrase) within halophilic microorganisms. Thus, we searched for integrons in 28 assembled hypersaline metagenomes and publically available 104 halophilic genomes by the aid of Integron Finder followed by blast search for group II intron reverse transcriptases (RT)s. Our results revealed the presence of group II introns from different classes associated with integrons and integron-related sequences. UHB.F1 and UHB.I2 group II introns were identified within putative integrons in the metagenome of Tanatar-5 hypersaline soda lake, belonging to IIC and IIB intron classes, respectively. Only UHB.I2 was a complete group II intron, whereas, UHB.F1 was a fragmented one. Two other group IIB truncated introns: H.ha.F1 and H.ha.F2 were detected in a CALIN within the extreme halophile *Halorhodospira halochloris*. Identified group IIB intron-encoded proteins (IEP)s belonged to CL1 class in UHB.I2 and to bacterial class E in H.ha.F1 and H.ha.F2. We have also identified a new insertion sequence (IS*Hahl1*) from IS*200/605* superfamily that was adjacent to *H. halochloris* CALIN. Finally, an abundance of toxin-antitoxin (TA) systems was observed within newly identified integrons and CALINs. Our analysis is the first study of group II introns within integrons in hypersaline metagenomes and halophilic genomes. Here, we report the existence of group IIB intron associated with integrons or CALINs in halophiles. This could provide a base for comprehending the potential role of group IIB introns in halophilic adaptation and their possible biotechnological applications.

6.1. Introduction

Group II introns are mobile genetic elements (MGE)s with properties of both catalytic RNAs (ribozymes) and retroelements [106,207]. They are found in bacterial and archaeal genomes, in addition to mitochondrial and chloroplast genomes of lower eukaryotes and plants [106,208]. The transcribed ribozyme catalyzes the excision of the intron and its integration into new locations with the aid of an intron-encoded protein (IEP) [208]. Despite of the poor conservation of the RNA sequence of the ribozyme [207], it can be classified into three major groups (IIA, IIB and IIC) [108]. Group II introns classification is based on their conserved secondary and tertiary structure where the transcribed intron forms six double helical domains (DI-DVI) radiating from a central wheel [108,208] (Fig.6.1A). Amongst the six double helical domains, DV and DVI are the only conserved domains [207]. DI and DV form the catalytic core of the ribozyme, while DIV contains the intron ORF [108]. Catalysis is promoted by the binding of Mg^{2+} ions to an AGC triad [208] (CGC in case of group IIC introns [209]) and to an AY bulge, located in DV [208] (Fig.6.1A).

Moreover, group II introns can be classified into subgroups based on their IEPs (Fig.6.1B):

mitochondrial-like (ML), chloroplast-like class I (CL1), chloroplast-like class II (CL2) and bacterial classes A-E [128]. Group II introns in bacteria contain all previously mentioned subgroups, whereas organelles contain only CL and ML subgroups [210]. The IEP acts 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) [108,109,128]. A "YADD" motif necessary for the reverse transcription activity is highly conserved in all bacterial IEPs within RT5 domain [109,208] (Fig.3.1B). Each IEP subgroup can be associated with one RNA subclasses as follows: ML (IIA1), CL1 (IIB1), CL2 (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) [108]. Most bacterial IEPs are found within MGEs such as plasmids or ISs [109].

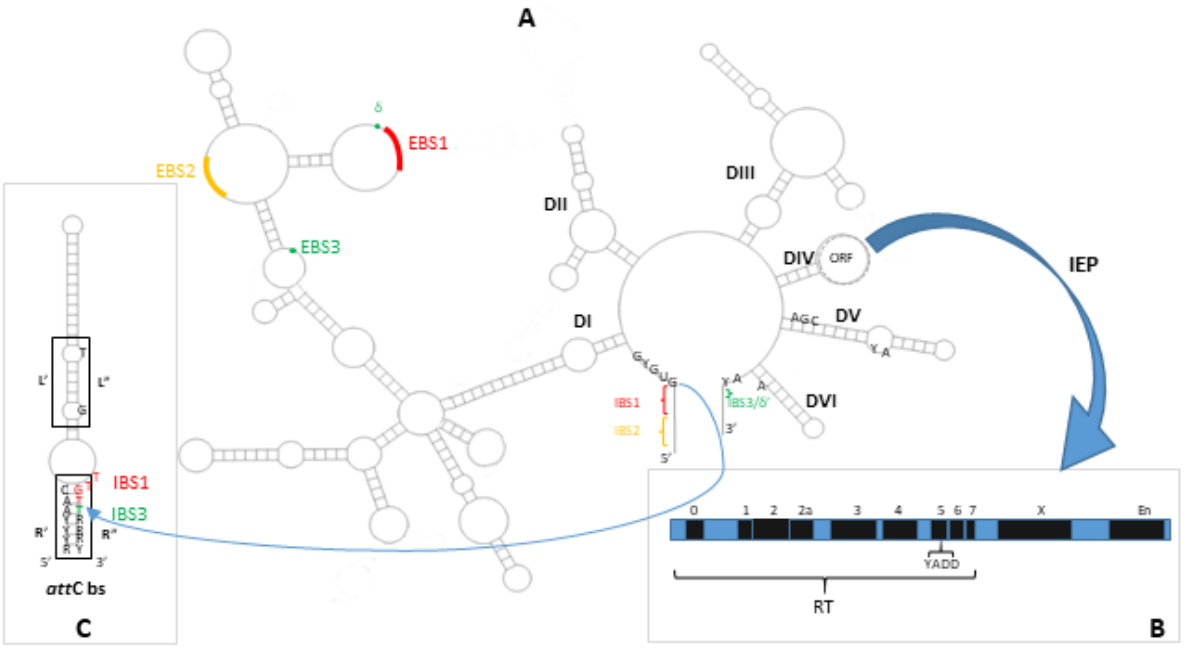


Fig.6.1 **General secondary structure of group II intron RNA, *att*C site and domains of IEP.** Group II intron is composed of six domains (DI-DVI) at which DI and DV form its catalytic core (A). Intron encoded protein (IEP) is encoded by ORF in DIV (A). The main domains of an IEP ORF (RT: reverse transcriptase, 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) in which the intron is inserted at the R'' sequence into the consensus sequence TTGT/T (IBS1/IBS3).

Mobilization of group II introns occurs through an RNA intermediate leading to their duplication [211]. The ribozyme in its conserved secondary structure can catalyze its own splicing (excision) from a precursor transcript [212]. Intron splicing usually occurs via two sequential transesterification steps [108] 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 (circle with a tail) and the ligation of the 5' and 3' exons [106,213]. 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 [213]. However, in group IIC introns, the hydrolysis pathway is more common [212]. The excised intron transcript (RNA) remains associated with the IEP forming a ribonucleoparticle (RNP), which can then be inserted (reverse splicing) into either an intronless allele (retrohoming) or into a non-cognate site (retrotransposition or ectopic transposition) which happens with a lower frequency [106]. 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 already formed eliminating the need for a second strand cleavage [106]. Yet, reverse splicing into ds or ssDNA independent of DNA replication can also occur with low frequency [106]. Different studies have shown that 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 [209]. For its insertion, 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) [106,214]. In group II A and B, 5' exon is recognized mainly by two base pairing interactions; IBS1-EBS1 (6 bp) and IBS2-EBS2 (6 bp) [212]. 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 [106]. On the other hand, group IIC introns exhibit some variations in their target site recognition; both IBS1-EBS1 and IBS3-EBS3 interactions are present. However, pairing in IBS1-EBS1 is 3-4 bp rather than six bp. To the moment, there's no evidence for an 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 [106,128].

*att*C sites are recombination sites found at the 3' end of an integron gene cassette, which can be recognized by an IntI protein, leading to integration or excision of integron gene cassettes [39]. An *att*C site is composed of four successive binding sites denoted by R", L", L' and R'. The only conserved domains are R" and R', with the consensus of 5'-RYYYACC-3' and 5'-GTTRRRY-3', respectively [39,41]. 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 [47] (Fig.6.1C). Group IIC-*att*C introns form a specific lineage of group IIC introns. They were found inserted directly after or into the stem-loop motif of the *att*C site bs, in an opposite orientation to the gene cassettes transcription [128,129]. 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) [50]. The majority of these introns were inserted into a consensus sequence of TTGT/T (IBS1/IBS3) within an *att*C site [128,129] (Fig.6.1C). 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 [128]. Thus, perhaps the presence of Group IIC-*att*C introns within gene cassette arrays is an intermediate step in the formation of some gene cassettes [128].

Members of group IIA and IIB introns have been successfully utilized as gene targeting vectors (targetrons) with high integration efficiency and target specificity [215]. On the other hand, group IIC introns have never been used in such applications, as their reverse splicing mechanism is not fully understood to the moment [215]. 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 [215].

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, November 2019). 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 ppt [216] with a highly active microbial sulfide cycle [217]. *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 sulfide, a pH of 8.1-9.1 [218] and salt concentration of 140-270 ppt [218]. 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 [183]. 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.

6.2. Materials and Methods

6.2.1. 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 6.1 shows all analyzed assemblies, whereas a list of halophilic bacteria and archaea was obtained from the Halodom database [179] in November 2019: “halodom.bio.auth.gr” (Appendix C: TableS6.1 and TableS6.2). 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 (Appendix C: TableS6.3). 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 6.1 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, [193, 194]	8530607	11351
GNM2	Guerrero Negro mat, Mexico 1-2mm depth	GCA_000206565.1, [193, 194]	7390978	10551
GNM3	Guerrero Negro mat, Mexico 2-3mm depth	GCA_000206545.1, [193, 194]	8209846	11423
GNM4	Guerrero Negro mat, Mexico 3-4mm depth	GCA_000206525.1, [193, 194]	8130049	11724
GNM5	Guerrero Negro mat, Mexico 4-5mm depth	GCA_000206505.1, [193, 194]	9689398	14128
GNM6	Guerrero Negro mat, Mexico 5-6mm depth	GCA_000206485.1 , [193, 194]	8291075	11380
GNM7	Guerrero Negro mat, Mexico 6-10mm depth	GCA_000206465.1, [193, 194]	9759240	13649
GNM8	Guerrero Negro mat, Mexico 10-22mm depth	GCA_000206445.1, [193, 194]	7914434	11356
GNM9	Guerrero Negro mat, Mexico 22-34mm depth	GCA_000206425.1, [193, 194]	8308787	11596
GNM10	Guerrero Negro mat, Mexico 34-49mm depth	GCA_000206405.1, [193, 194]	7132956	10297
ATII SDM	Atlantis II Deep Brine Sediment, Red Sea	[6,167,168]	40413330	41726
DD SDM	Discovery Deep Brine Sediment, Red Sea	[6,167,168]	52421642	51829
Th	Thetis Mediterranean deep-sea hypersaline lakes	GCA_001684355.1	13102297	10347
ATII INF	Atlantis II Deep Brine interface, Red Sea	[164,168]	16014945	24317
DD INF	Discovery Deep Brine interface , Red Sea	[164,168]	11647401	18413
KD UINF	Kebrit Deep Upper interface, Red Sea	[164,168]	42652688	45750
KD LINF	Kebrit Deep Lower interface, Red Sea	[164,168]	50280352	74666
ATII LCL	Atlantis II Deep Brine, Lower convective layer, Red Sea	[164,168]	46518597	43555
ATII UCL	Atlantis II Deep Brine,Upper convective layer, Red Sea	[164,168]	21343827	29592
DD BR	Discovery Deep Brine , Red Sea	[164,168]	12244355	18850
KD BR	Kebrit Deep Brine, Red Sea	[164,168]	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, [195, 196]	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

6.2.2. Identification of integrons and CALINs

We used IntegronFinder version 2.0 [50] to search for complete integrons, Integron integrase genes (*int*I) 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 eight kb distance threshold between successive identified *att*C sites to ensure the detection of all potential *att*C sites. At least two *att*C sites should be detected within the eight kb threshold to be reported as a positive result. A search for integron cassette promoters (P_C) and primary recombination sites (*att*I) for known integron classes (1, 2 and 3) was also performed.

6.2.3. 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/) [219,220] 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) [219] [220]. Identification of intron boundaries was done by the MFOLD webserver, which folds the introns RNA structure [221] based on known secondary structures of group II intron classes, that showed 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 [222,223]. 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’, Ɛ-Ɛ’ 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 [208] were determined manually whenever applicable. The final secondary structure was then depicted using Pseudoviewer3 [224].

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 [225]. Detection of introns upstream hairpin structures was done using MFOLD [221], respectively.

6.2.4. Insertion sequences identification

ISfinder [183] was used to search for insertion sequences within contigs or genomes in which integrons or CALINS were identified. ISEScan pipeline [182] was also used for further inspection of insertion sequences within *H. halochloris* DSM 1059 genome.

6.2.5. 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 TA systems genes, IEP ORFs and within group II introns was done using bprom tool [173].

6.2.6. Phylogenetic analysis

The 4 identified IEPs in this study were aligned with 34 bacterial IEPs from different classes using MUSCLE [223], along with Mitochondrial IEP from Liverwart *Marchantia polymorpha* as an outgroup. Molecular phylogenetic analysis was done with MEGA7 [226] using the Maximum Likelihood with WAG 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 analyzes with 1,000 samplings.

6.2.7. Determination of *H. halochloris* leading and lagging strands

GammaBOriS tool specifically designed for identification of origin of replication (*Ori*C) sequences in gammaproteobacterial chromosomes [227] was used for identification of probable *H. halochloris Ori*C. Based on the approach used by Mao et al [228]. 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 [228].

6.3. Results

6.3.1. 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 (Chapter 5). Annotation of the identified gene cassettes revealed the presence of two-group II intron RT/maturases in two 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, on which the first was truncated, were referred to as uncultured halophilic bacterium introns 1 and 2 (UHB.F1 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 seven 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 two 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 [218] were detected. Apart from the identified CALIN, only one other group II intron RT 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.F1, UHBI2, H.ha.F1 and H.ha.F2 to specific intron classes, we constructed maximum likelihood phylogenetic tree with different classes of bacterial IEPs (Fig.6.2). The phylogenetic tree revealed that UHB.F1 belongs to Group IIC-*att*C class, known to be associated with integrons [128] [129,229]. 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 [219,220] confirmed the results of our phylogenetic analysis. The closest hit to UHB.F1 was Ge.s.I1 of group IIC-*att*C class from *Geobacter sulfurreducens* with 50% identity and 62% similarity. Since the group II intron database is limited in number of sequences, we blasted the sequence against the vast NCBI nr database, closer hits were obtained, as the best hit was a group II intron RT from a *Verrucomicrobia* bacterium (sequence ID: NBB81160.1) with 80% identity and 87% similarity. However, the X domain of the IEP was detected in three different frames due to a small indel and an 11 bp insertion at the C-terminus. In addition, we were not able to locate the exact start of the translated protein as the predicted start by Prodigal [230] in Integron Finder tool detected the start at c(3899) missing few amino acids upstream that are actually part of the RT0 domain (Appendix D: Fig.S6.1 and Fig.S6.2). The whole RT0 domain was still incomplete missing few upstream amino acids indicating a possible deletion (Appendix D: Fig.S6.1). In case of UHB.I2, the closest hit was Sh.sp.I2 (CL1/IIB1) from 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.F1 and UHB.I2 IEPs, each with its closely related IEPs showed all required domains for IEPs lacking the endonuclease domain (En^-) (Appendix D: Fig.S6.1 and Fig.S6.3).

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 (Appendix D: Fig.S6.2 and Fig.S6.4).

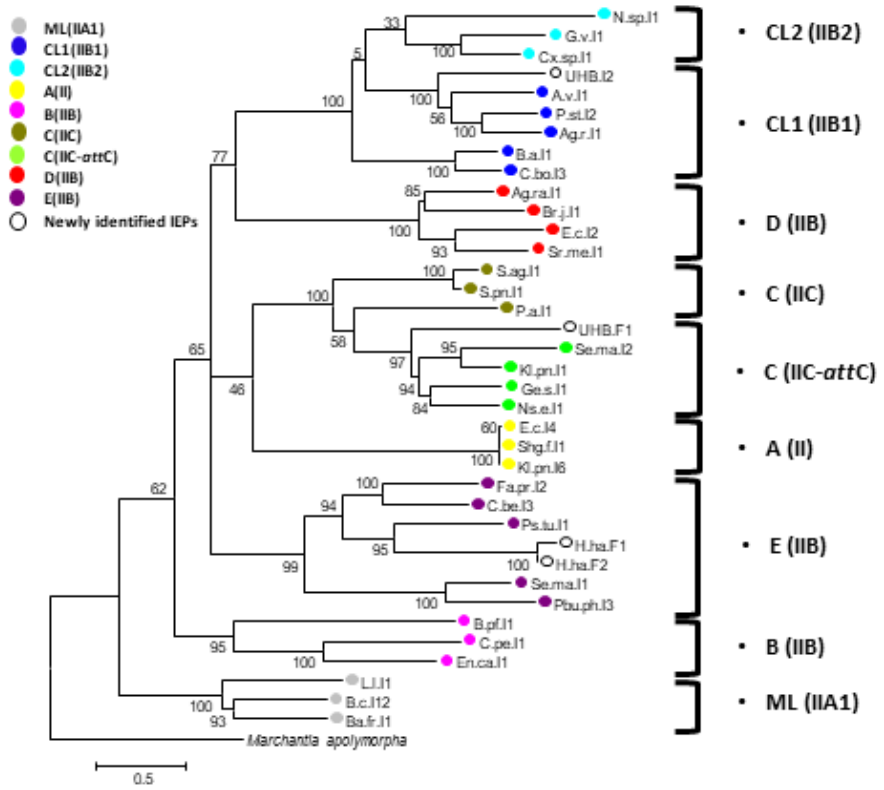


Fig.6.2 **Phylogenetic tree of identified putative IEPs with IEPs from different bacterial groups.** UHB.F1 clusters with group IIC-attC, UHB.I2 with group IIB (Chloroplast-like1 class) and both H.ha.F1 and H.ha.F2 with group IIB (bacterial class E). IEPs abbreviations are based on their introns nomenclature 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.

6.3.2. Metagenome of Tanatar-5 hypersaline Soda Lake (TSL1) harbors a truncated group IIC-*att*C intron within a gene cassette array

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 truncated group IIC-*att*C intron (UHB.F1) was detected (Appendix D: Fig.S6.5) within an array of gene cassettes. The intron was inserted in an opposite orientation to the adjacent gene cassettes. Being at the periphery of the TSL1 contig (9835bp), 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 (Fig.6.3A and Appendix C: TableS6.4).

We identified the 3' end of the intron which showed typical folding of DV and DVI loops (Appendix D: Fig.S6.6). However, although all RT domains were detected in the identified IEP, the RT0 domain missed few amino acids indicating a deletion at its N terminus. It is more likely that the identified intron is a 5' truncated intron as it was challenging to find a proper start or a properly full folded intron.

The gene cassette in which the intron is inserted has three other ORFs, two 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.6.3A and Appendix C: TableS6.4).

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) [208], we searched for the presence of putative promoters within UHB.F1 and its upstream region that could drive the transcription of gene cassettes at the 3' end of the gene cassette array and upstream of the intron. Four potential promoters were detected (Appendix C: TableS6.4 and Appendix D: Fig.S6.5). Perhaps one or more of these putative promoters is responsible for the expression of just one downstream ORF encoding for a hypothetical protein, since the TA operon in the next gene cassette had two predicted promoters in addition to two more predicted promoters within the antitoxin gene that could drive the transcription of the toxin gene (Appendix C: TableS6.4).

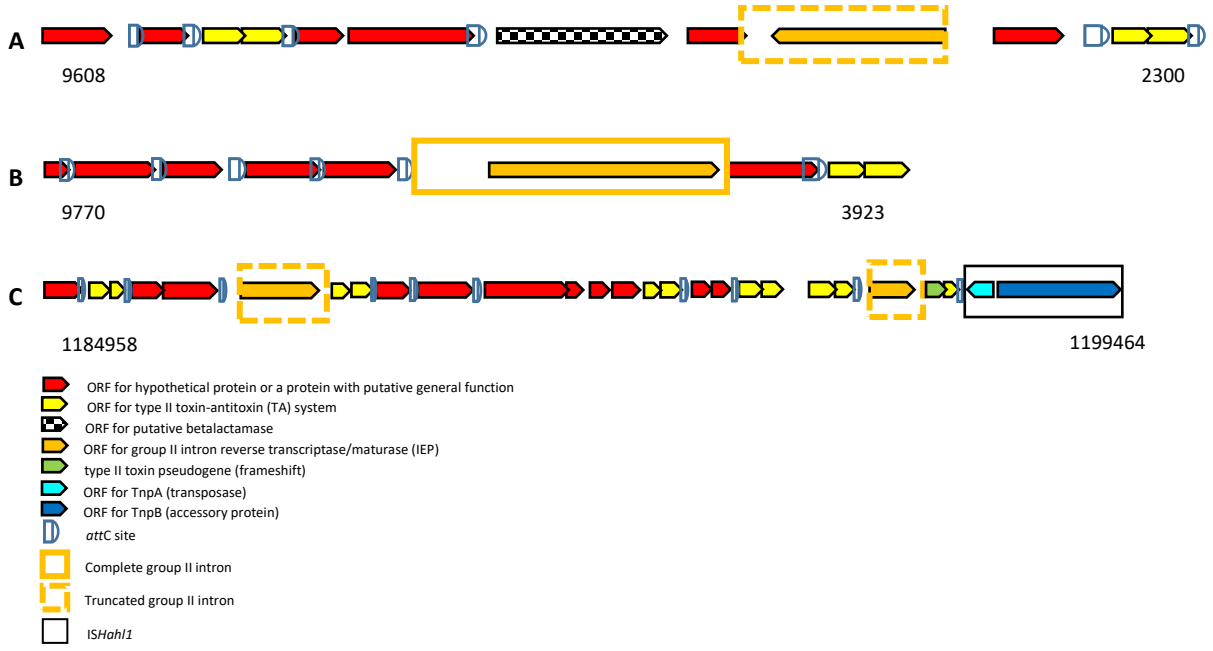


Fig.6.3 **Schematic representation of identified gene cassette arrays where group II introns are inserted.** A: TSL1 (LFIK01005867; c(2300..9608)), B: TSL2 (LFIK01005957;c(3923..9770)), C: *H. halochloris* CALIN (NZ_AP017372.2; 1184958..1199464). Arrow heads of different ORFs show the direction of their transcription. Colored legend show different genetic elements depicted. Map coordinates are indicated below each schematic representation.

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

Following the same steps described for the identification of UHB.F1 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.6.3B, Appendix C: TableS6.4, Appendix D: Fig.S6.2 and Fig.S6.5). 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.6.3B and Appendix C: TableS6.4).

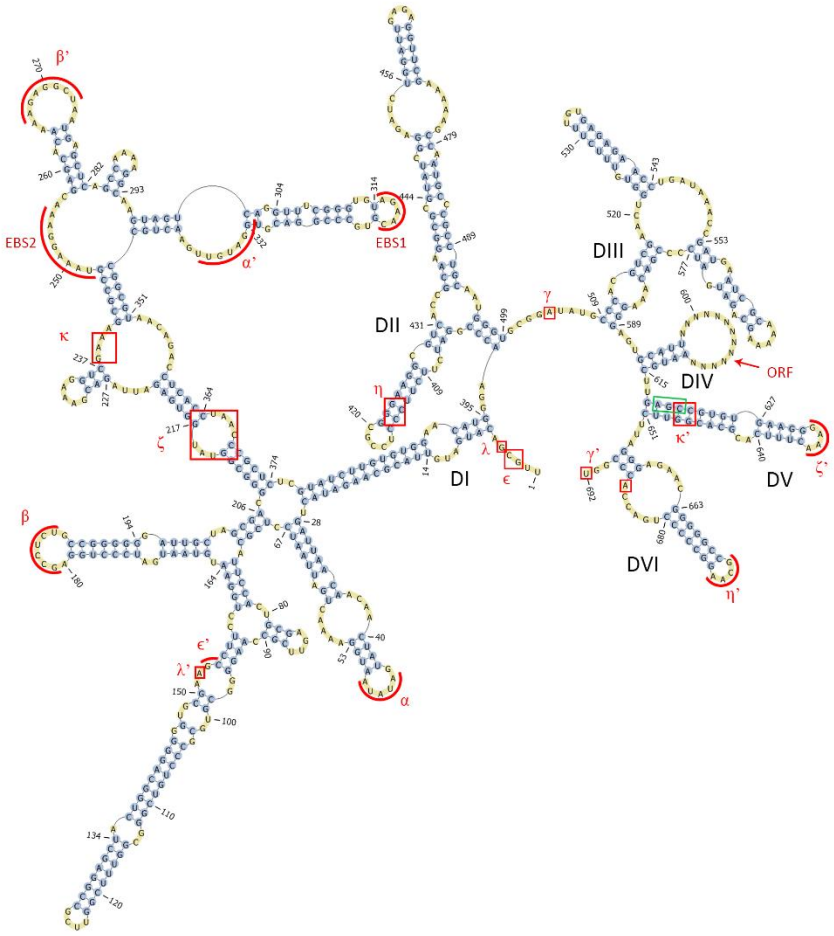


Fig.6.4 **Secondary structure of group II intron UHB.I2.** UHB.I2 identified in TSL2 contig shows necessary sequences required for intron splicing and reverse splicing. Important sequences are shown within red rectangles or curved lines. EBS1 and EBS3 are important for base-pairing with target site in flanking exons, whereas other identified sequences are necessary for intron folding (Watson-Crick α-α', β-β', Ɛ-Ɛ'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.

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.6.4). 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 the IEP ORF or downstream ORFs in the array (Appendix C: TableS6.4 and Appendix D: Fig.S6.5). 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 (Appendix D: Fig.S6.7). Examination of UHB.I2 flanking exons with homologous introns showed poor conservation for both exons except for the first two nucleotides in 3' exon (Fig.6.5).

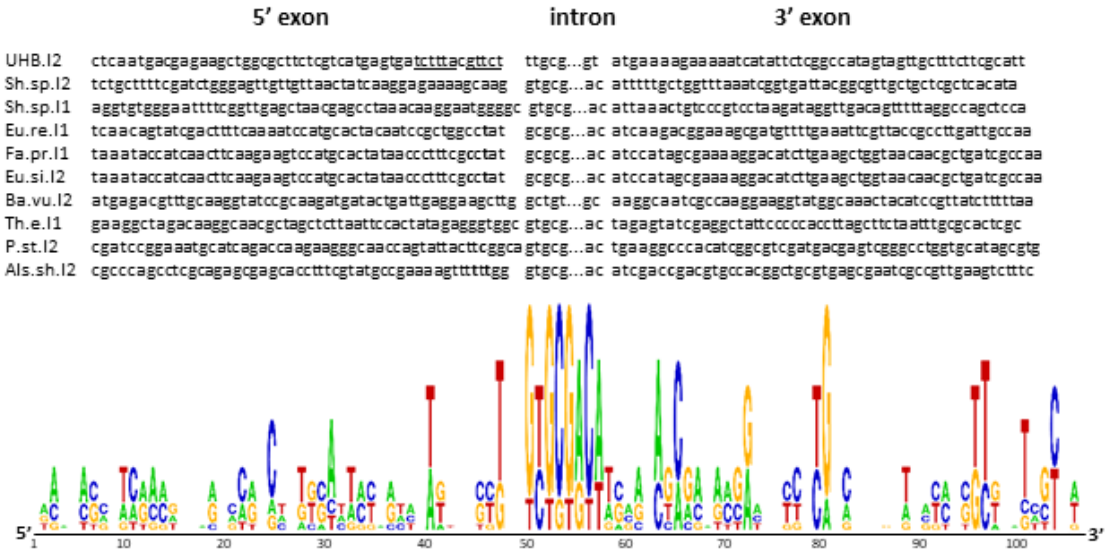


Fig.6.5 **UHB.I2 flanking exons logos with closer hits.** The logo shows a conservation in the target site. IBS1 and IBS2 sequences in UHB.I2 5' exon are underlined.

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 (Appendix D: Fig.S6.5). Folding of DV and DVI, depicting the 3' part of a group IIB intron were predicted in both intron RNAs (Appendix D: Fig.S6.6). Here again, putative promoters were predicted within H.ha.F1 and H.ha.F2 (Appendix C: TableS6.4 and Appendix D: Fig. S6.5). In all cases putative promoters directly upstream of all identified introns were detected (Appendix C: TableS6.4).

6.3.4. 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 detected ORFs within these integrons. All ORFs were BLASTed and annotated. We found that the three examined arrays contain type II TA systems of various types (Fig.6.3 and Appendix C: TableS6.4). Two TA system 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 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 where 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. Upstream to all identified TA operons, putative promoters also existed.

6.3.5. An insertion sequence (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 ten gene cassettes, with six ORFs in one gene cassette (detailed annotations in Appendix C: TableS6.4). Directly, downstream of the last *att*C in the identified CALIN, we found a new insertion sequence (IS) (Fig.6.3C), that we submitted to the ISfinder database [183] 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 [231]. The hairpin structures were compared to that of IS*CARN6*, the closest homologue in ISfinder database showing 66% identity to IS*Hahl1* (Appendix D: Fig.S6.9).

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 [231]. The configuration of the two ORFs is characteristic of IS*605* group within the IS*200/605* family [231]. With the aid of ISEScan pipeline [182], An isoform of IS*Hahl*1 was found with 98% identity (Appendix C: TableS6.4) about 70 kb downstream. Several other IS*605* group elements within the genome were identified; however, they were either partial, with frameshifts or missing parts in *tnp*A and *tnp*B genes (Appendix C: TableS6.4).

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 [227] results showed that the most probable *Ori*C position lies between 2787842-2789091 bp in the 2,834,560-bp-*H.haochloris*-genome. 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 its isoform are on the leading strand, while the *att*C sites' bottom strands in the CALIN are on the lagging strand.

6.4. Discussion

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

Group II introns has been identified in different bacterial, archaeal and organeller genomes [106]; however, their association with integrons has been limited to group IIC introns [128,129,229,232], and in most reported cases, this connection was confined to IIC-*att*C subclass [128,129,229]. 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 halophilic 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 these MGEs in environmental adaptation of halophiles. Our analysis revealed the presence of class IIC-*att*C and class IIB introns associated with integrons or CALINs 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 [215].

Our newly detected group IIC-*att*C intron, UHB.F1, 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 [128,129,229]. However, it was a 5' truncated intron with frameshifts near its 3' end. 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. In one reported case in an integron in *Enterobacter cloacae*, an unusual group IIC intron (not a IIC-*att*C) was at the same orientation of adjacent gene cassettes, as it was inserted within the top strand of an *att*C site rather than the usual bs target site [232]. Unlike other group II introns, group IIC intron target site possess a stem-loop structure upstream of the insertion point [106,128]. *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 [128]. 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 UHB.I2, there was no sequence conservation in its flanking exons. However, it showed an AT rich 3' exon (Fig.6.5). 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 [214]. Further experiments should be performed to determine the role of the UHB.I2 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.6.4). It is likely that IBS3 on the target site interacts with an alternative EBS3 site or position.

6.4.2. Identification of putatively essential upstream secondary structures for group II intron mobilization in *H. halochloris*

All our identified IEPs lacked an endonuclease domain (En^-), which is in more than half the bacterial group II introns IEPs [106,108]. Since En^- IEPs are incapable of a second strand cleavage, they depend on the host replication machinery for insertion into new target sites [106].

Based on GammaBOriS [227] identification of the origin of replication in *H. halochloris*, H.ha.F1 and H.ha.F2 are inserted within the leading strand rather than the lagging strand; a documented yet rare phenomenon [106]. 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 [106].

In *att*C recombination, replication is not only important for the formation of the folded bs, but also for the resolution of recombination products [57,174]. However, the presence of single stranded proteins (SSP) hampers the formation of a fully folded *att*C bs in absence of integron integrase (IntI) [233,234]. 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 [234]. 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 their 3' ends were identified, and important RT domains within their IEP ORFs were also absents. It is already documented that fragmented introns with frequent frameshifts are more commonly found than full-length introns in bacterial genomes [210]. Yet, a putatively functional IEP ORF (80% identical to H.ha.F1 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 [221], we have detected an *att*C-like structure upstream of H.ha.F1 (Appendix D: Fig.S6.10A) and a putative *att*C site upstream of H.ha.F2, showing a nearly typical *att*C site bs secondary structure (Appendix D: Fig.S6.10B). Again, the presence of these secondary structures before group IIB introns further suggests their possible role in recognition of target sites.

6.4.3. 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, where 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 [101,231] and 4- the presence of hairpin structures at both ends [29]. Transposition is strand specific and follows a "peel and paste" mechanism in which an excised circular single stranded intermediate integrates into a single stranded target site [231]. 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 [231]. 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 [231]. In fact, it was suggested that identical IS*605* elements in *H. pylori* had caused rearrangement within its genome [235]. The presence of an isoform to IS*Hahl*1 (98% identity) and other IS*605* elements with high homology to IS*Hahl*1 (Appendix C: TableS6.4) may allow such rearrangements to occur by homologous recombination. The rationale behind our mining for similar IS element was to inspect the possibility of mobilization of the adjacent CALIN sequence. Yet the large distance between the nearest homologous upstream IS605 element at 460538-461995 bp (~735 kb) 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 [236]. Moreover, it is worth noticing that IS*200/605* elements belong to HUH endonuclease superfamily to which IS*91* and ISCRs (Insertion sequence common regions [117]) belong as well. IS*91* and ISCRs are postulated to transpose their ssDNA sequence with a rolling circle replication mechanism that starts at a specific site named *Ori*-IS and ends at a termination sequence *ter*-IS. However, high frequency of termination failure at the *ter*-IS site can be observed, leading to a one ended-transposition, mobilizing adjacent sequences at the 3' end of the IS element [101]. Although this mechanism could explain the associated antibiotic genes commonly found downstream ISCRs, it cannot explain those lying at its upstream part [237]. Perhaps, a common minor transposition mechanism for ISCRs and IS*200/605*, other than the one already established for IS*200/605* transposition, exists allowing mobilization of adjacent genes to the IS elements in both directions. If this is true, this may allow the transfer of a CALIN without requiring the activity of an integron integrase for excising and integrating separate gene cassettes. Definitely, this needs a lot of investigation and experimental work to be verified.

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 [106,174,231,238]. 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.

6.4.4. 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. The abundance of TA systems as gene cassettes within integrons has already been observed in different studies [39]. It is hypothesized that TA systems could have a role in maintaining the integrity of these integrons by preventing deletions of existing arrays [39,67]. 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, three 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 [67]. 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 [67]. Similarly the identified TA cassettes from hypersaline environments can increase–mobilization of different MGEs such as integron gene cassettes, prophages and transposons [67].

6.5. Conclusions

Integrons and CALINs 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. 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.

Chapter 7: Differential Prokaryotic Consortia in Athalassohaline and Thalassohaline Brines

Abstract

Documentation of prokaryotic diversity is an essential primary step towards understanding microbial contribution to ecosystem dynamics in hypersaline environments. The bacterial composition of two Egyptian brines, athalassohaline Aghormy Lake in Siwa Oasis, and thalassohaline Sebeaka saltern on the eastern side of Bardawil Lagoon (north coast of Sinai Peninsula), was assessed based on metagenomic 16S rRNA high throughput sequencing. A total of 488828 reads from both sites grouped into 17741 operational taxonomic units (OTUs) were obtained. 3030 OTUs were shared in both sites, while 2255 and 9426 OTUs were unique to Aghormy Lake and Sebeaka saltern, respectively. Aghormy brine OTUs were assigned to 51 bacterial families, belonging to 16 phyla. OTUs in Sebeaka saltern were assigned to 37 families, belonging to 10 phyla. Unassigned reads represented 3.6% and 2.5% of total reads from Aghormy and Sebeaka brines, respectively. Both sites showed an abundance of Bacteroidetes, particularly family Rhodothermaceae. Aghormy Lake was characterized by phylotypes belonging to Deinococcus-Thermus, Spirochaetes, *Rhodovibrio* (Alphaproteobacteria), Chromatiaceae (Gammaproteobacteria) and GMD14H09 (Deltaproteobacteria). Phylotypes assigned to AT12OctB3 (Bacteroidetes), Rhodobacteriaceae (Alphaproteobacteria), Ectothiorhodospiraceae and Xanthomonadaceae (Gammaproteobacteria) formed Sebeaka saltern bacterial community. Cyanobacteria-like phylotypes were assigned to class Oscillatoriophycideae, in both brines. Archaeal family, Halobacteriaceae, represented 4.8% of Sebeaka brine reads. In spite of the presence of phylotypes belonging to the same phyla in both brines, differences were observed in lower taxonomic ranks which may reflect the differences in the biogeographical nature, physicochemical parameters and different stresses between the two brines. Here, we report the different prokaryotic phylotypes in these hypersaline environments.

7.1. Introduction

Brines are intriguing habitats; their study helps in comprehending how extreme environments shape the microbial community and enable adaptation under different physical and chemical conditions. Hypersaline habitats are either thalassohaline, of marine origin, or athalassohaline, inland saline aquatic environments. While the molarity in thalassohaline environments is mainly attributed to Cl^- and Na^+ ions, athalassohaline environments are characterized by different ionic composition from the general marine environment [239]. Athalassohaline environments are less abundant than thalassohaline ones and their ionic composition may differ from each other, depending on their origin [239]. However, the dominance of divalent cations is observed in many

athalassohaline environments, such as the Dead Sea [239,240]. Others are dominated by K^+, Mg^{2+}, Na^+ and $CO3^{2-}$ ions [241].

Siwa Oasis at the Western Desert in Egypt, is a depression between latitudes 29° 05′ N and 29° 25′ N and longitudes 25° 05′ E and 265° 06′ E with an area of about 1200 km^2. The climate is arid to semi-arid with scarce rainfall [163]. The deepest parts of the oasis are occupied by salty lakes surrounded by salt marches. Lakes in Siwa Oasis, are the natural discharge areas for water coming from the abundant artesian wells, springs and cultivated areas [13]. Aghormy Lake, athalassohaline environment, is located 18 m below sea level and is fed by springs with orifices within the lake. The lake is characterized by total dissolved solids (TDS) of 220.03 g l^{-1}(ppt) and a pH of 7.83 [14]. The lake with its surrounding sediments occupies an area of about 80 km^2. It is predominated by evaporite minerals, mainly halite (NaCl) and to a lower extent gypsum ($CaSO_4.2H_2O$) and polyhalite ($K_2Ca_2Mg(SO_4)_4.2H_2O$) [242]. Collective samples from different drainage water lakes in Siwa have shown that these waters are of (Na-Cl-SO_4) type [243]. The chemical composition of Aghormy Lake is characterized by the dominance of Mg^{2+} (52 g l^{-1}), Na^+ (55 g l^{-1}), Ca^{2+} (10 g l^{-1}) and K^+ (10.4 g l^{-1}) cations, in addition to Cl^- (246.8 g l^{-1}) and SO_4^{2-} (10.7 g l^{-1}) anions [13]. In addition, the sediment there is significantly enriched with different heavy metals such as Cu, Cd, Se, Co, Pb, Mn and Zn [14]. Several studies have documented seasonal blooming of microbial mats along the margins of the lake, in spring and early summer [14,242].

Bardawil Lagoon is a shallow hypersaline lagoon at the north coast of Sinai Peninsula in Egypt [17]. It covers an area of about 600 km^2, where it is mainly connected to the Mediterranean Sea via one natural inlet and two artificial ones [20]. The salinity of the Lagoon ranges from 39.5 -68.5 ppt [17]. However, the water concentrates in the southern and eastern parts resulting in the precipitation of gypsum and halite [244]. Hence, supratidal salt flats cover the southern and eastern parts of the lagoon, and are normally described as “Sabkhas” [17]. In general, two types of Sabkhas can be encountered in the vicinity of the Lagoon; coastal sabkhas that are connected to the lagoon and inland sabkhas that are separated from it by sand dunes [19]. Sebeaka saltern is an example of a coastal Sabkha in the eastern part of Bardawil Lagoon in the Zaranik Protectorate wetland [17] [20], representing thalassohaline environment. It is divided into 3 zones based on differences in elevation and sedimentary structures; a higher outer dry zone, an intermediate wet zone and a central basin. The red coloration of the saltern water is characteristic of hypersaline environments, with red halophilic archaea and carotene-rich algae [1]. Being a coastal Sabkha, Sebeaka saltern is connected to the lagoon and occasionally flooded by its water. The frequent evaporation of water results in halite precipitation and the formation of a permanent thick halite crust. The arid conditions of the area along with the availability of hypersaline water facilitate the formation of these salt crusts [21]. Thus, Sebeaka saltern is utilized in commercial salt production [17,20]. Sabkhas in this area are generally composed of sand, gypsum, halite and calcite ($CaCO_3$) [21]. In general, salterns are

characterized by sequential precipitation of different salts, starting by CaCO3, followed by gypsum accumulating on the bottom of the ponds and finally NaCl precipitating at salinities above 300 ppt, resulting in waters concentrated with Mg, K, Cl and sulfate ions [245]. A study on the soil from the eastern and south-eastern Sabkhas has shown that the pH ranges from 7.5 to 8.3 [246].

Halophilic prokaryotes represent diverse groups of bacteria and archaea and their distribution and metabolic activities depend on salt concentration [247]. The difference in salinity tolerance of halophiles shapes their diversities in brines and makes them suitable candidates for several biotechnological applications [248].

Lack of knowledge about biodiversity in Egyptian hypersaline environments leaves gaps in understanding the ecological role of halophilic microorganisms within these habitats. The objective of this study was to unravel the differential phylogenetic diversity of prokaryotes inhabiting Egyptian athalassohaline, Aghormy Lake, and thalassohaline, Sebeaka saltern. Sequencing of seven hypervariable regions of the 16S rRNA gene, using ion S5 high-throughput sequencing, recorded unique phylotypes, characterizing each site.

7.2. Materials and Methods

7.2.1. Sampling

Ten-litre water samples were collected from each of two Egyptian brine environments during spring season. The first site was the athalassohaline Aghormy brine, Siwa Oasis in the Western Desert, 29°11′ 44″N, 25°35′18″E (Fig.7.1A). The second site was the thalassohaline Sebeaka saltern field, Bardawil Lagoon, in Northern Sinai, 31°5′44″N, 33°28′46″E (Fig.7.1B). Water samples were filtered through membrane 0.2 µm filters, Millipore, and washed with sterile TE buffer, 50 mM EDTA, 50 mM Tris-HCl, pH 7.6, and processed for molecular analysis.

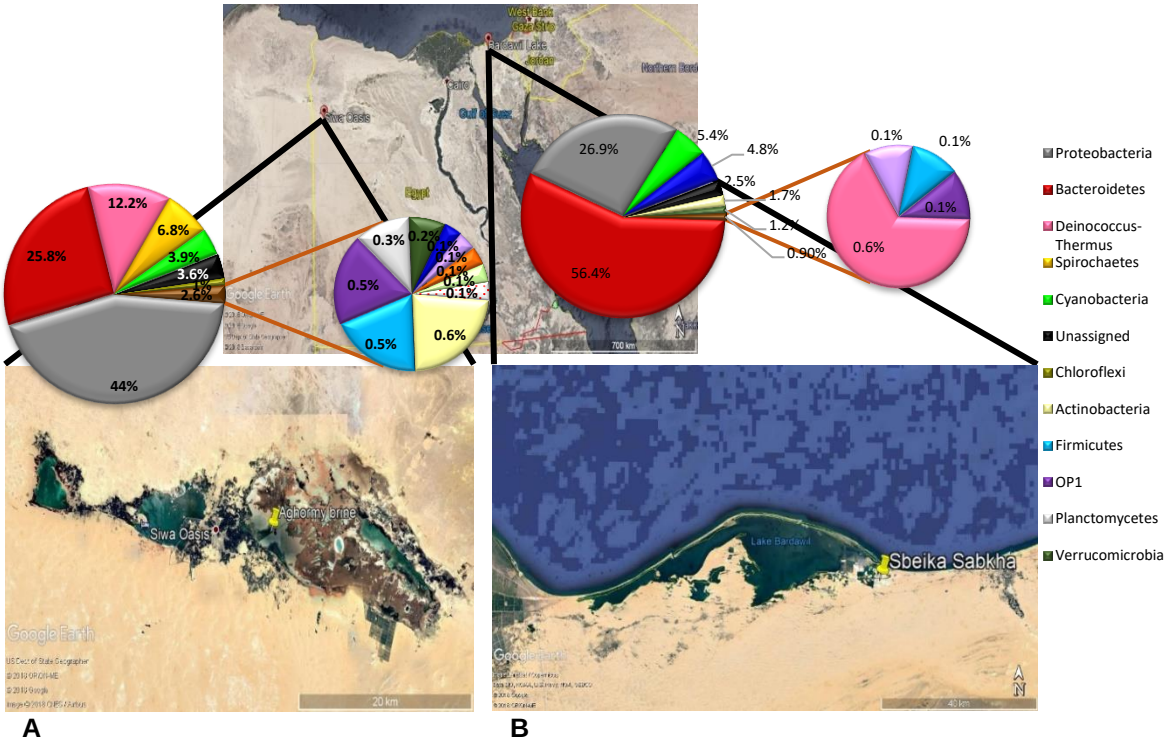


Fig.7.1 Map showing locations of the studied Egyptian brines along with the prokaryotic phyla distribution in each site. A: Aghormy Lake, Siwa Oasis, Western Desert, Egypt. B: Sebeaka saltern on the eastern side of Bardawil Lagoon, North Sinai, Egypt

7.2.2. Molecular analysis

Metagenomic DNA was extracted from filter-concentrated microbial communities, using a DNA isolation kit (MO BIO Laboratories, 12888-50, Carlsbad, CA), according to the manufacturer's specifications with some modifications. Microbes were lysed, using a mixture of 5M guanidine thiocyanate (Sigma) and 10% sodium dodecyl sulfate (SDS), followed by incubation at 75°C for 20 min with shaking. Cellular debris was then removed by centrifugation for ten min at 10000x g and the supernatant was collected in sterilized propylene tube. DNA was purified using the Sephadex columns, included in the kit.

Ion 16S Metagenomics kit (Thermo Fisher Scientific, Cat. no. A26216), which includes V2, V3, V4, V6-7, V8 and V9 primers, targeting seven hypervariable regions of the 16S rRNA gene, was used for PCR amplification according to the manufacturer's specifications. The workflow described in Ion 16S Metagenomics kit user guide was followed and recommended kits by the manufacturer were used.

7.2.3. Bioinformatics analysis

Primary data analysis was performed using the Ion Torrent Suite™ Software within the Ion S5™ platform. Acquisition of raw data, well validation, base calling and quality check for each well were done. Trimming of adaptor sequences and low quality 3' ends of reads were performed, in

addition to the removal of both short reads (less than eight bp) and reads resulting from polyclonal Ion sphere particles. Reads that passed quality check were exported to an unmapped BAM file. The obtained sequenced read BAM files were registered in DNA Data Bank of Japan, DDBJ, under the accession number DRA006839.

BAM files were converted to FASTQ files with Galaxy tool (version 2.26.0) [249]. Further analysis and open-reference OTU picking was done using QIIME bioinformatics pipeline version 1.9.1 [250]. Chimeric sequences were filtered out, and the filtered reads were clustered against Greengenes reference database using the script "pick_open_reference_OTUS.py". Reads that did not align to the reference database were subsequently clustered as *de novo*. A threshold of 97% identity was used for defining any distinct OTU. Alpha diversity analysis was performed and rarefaction curves were plotted based on Chao1 estimator. Heatmap was drawn on RStudio version 1.0.136 using pheatmap package (https://CRAN.R-project.org/package=pheatmap).

7.3. Results and Discussion

7.3.1. Phylotypes profiles of studied brines

The 16S rRNA gene sequencing of Aghormy Lake and Sebeaka saltern generated a total of 85291 and 403537 valid reads, respectively (Table 7.1). The amplified reads covered seven of the 16S rRNA gene variable regions. There were 3030 common OTUs detected in both studied sites, while 2255 and 9426 OTUs were uniquely identified in Aghormy Lake and Sebeaka saltern, respectively (Fig.7.2). Alpha-diversity analysis, based on Chao1 richness index, showed a plateau in the rarefaction curves in both samples, indicating deep coverage, in which saturation was observed following 20000 and 80000 sampled sequences in Aghormy Lake and Sebeaka saltern, respectively (Fig.7.3).

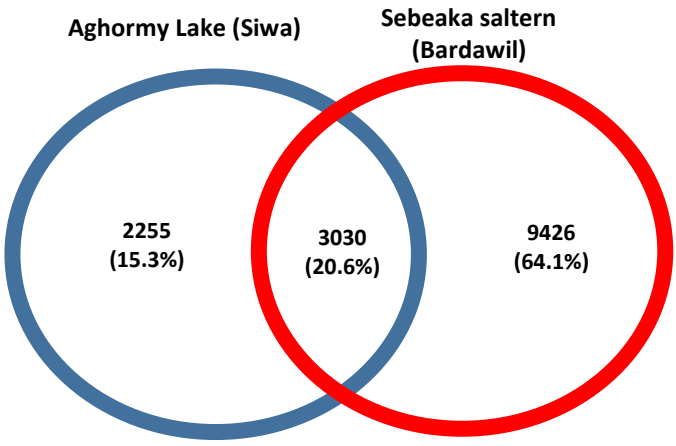


Fig.7.2 Venn diagram showing distribution of detected OTUs in Aghormy Lake and Sebeaka saltern

Table 7.1 Number of reads and OTUs in Aghormy Lake and Sebeaka saltern.

Sequenced reads	**Aghormy Lake**	**Sebeaka saltern**
Total reads	107962	517313
Valid reads	85291	403537
Number of OTUs	5285	12456
Taxonomy assigned reads	82229	393411

The Aghormy Lake OTUs were assigned to 51 known bacterial families with 37 classes belonging to 16 phyla (Fig.7.1 and Fig7.4). On the other hand, OTUs in Sebeaka saltern were assigned to 37 known families with 19 classes belonging to 10 phyla (Fig.7.1 and Fig.7.4). This suggests a relatively higher species richness, but low evenness, in the Aghormy Lake when compared to Sebeaka saltern microbial community (Fig.7.3 and Fig7.4). Unassigned reads represented 3.6% and 2.5% of total sequences from Aghormy and Sebeaka brines, respectively (Fig.7.4).

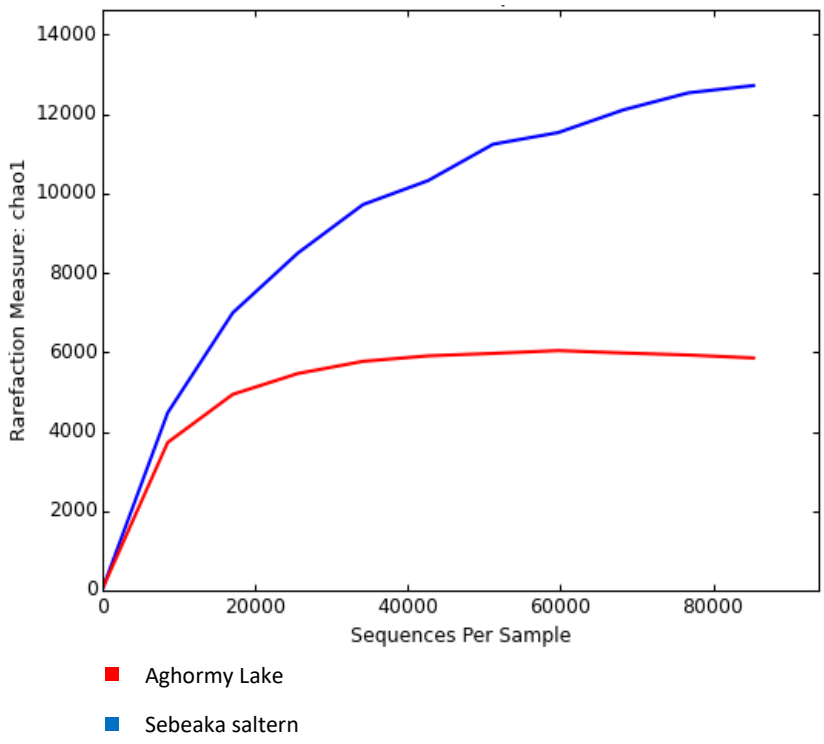


Fig.7.3 Rarefaction curves with Chao1 estimator corrected numbers of observed OTUs in both Aghormy Lake and Sebeaka saltern

The abundant phyla in both sites were Bacteroidetes and Proteobacteria. Bacteroidetes-like OTUs represented 25.8% and 56.4% of the total valid reads in Aghormy Lake and Sebeaka saltern, respectively (Fig.7.1). This observation was in accordance with other findings from different hypersaline environments, such as the Dead Sea [251], Soda lakes in Russia [216], Tirez Lagoon in Spain [252] and Lake Tebenquichi, Chile [253]. Proteobacteria represented 44% and 26.9% of the total sequences in Aghormy Lake and Sebeaka saltern, respectively (Fig.7.1).

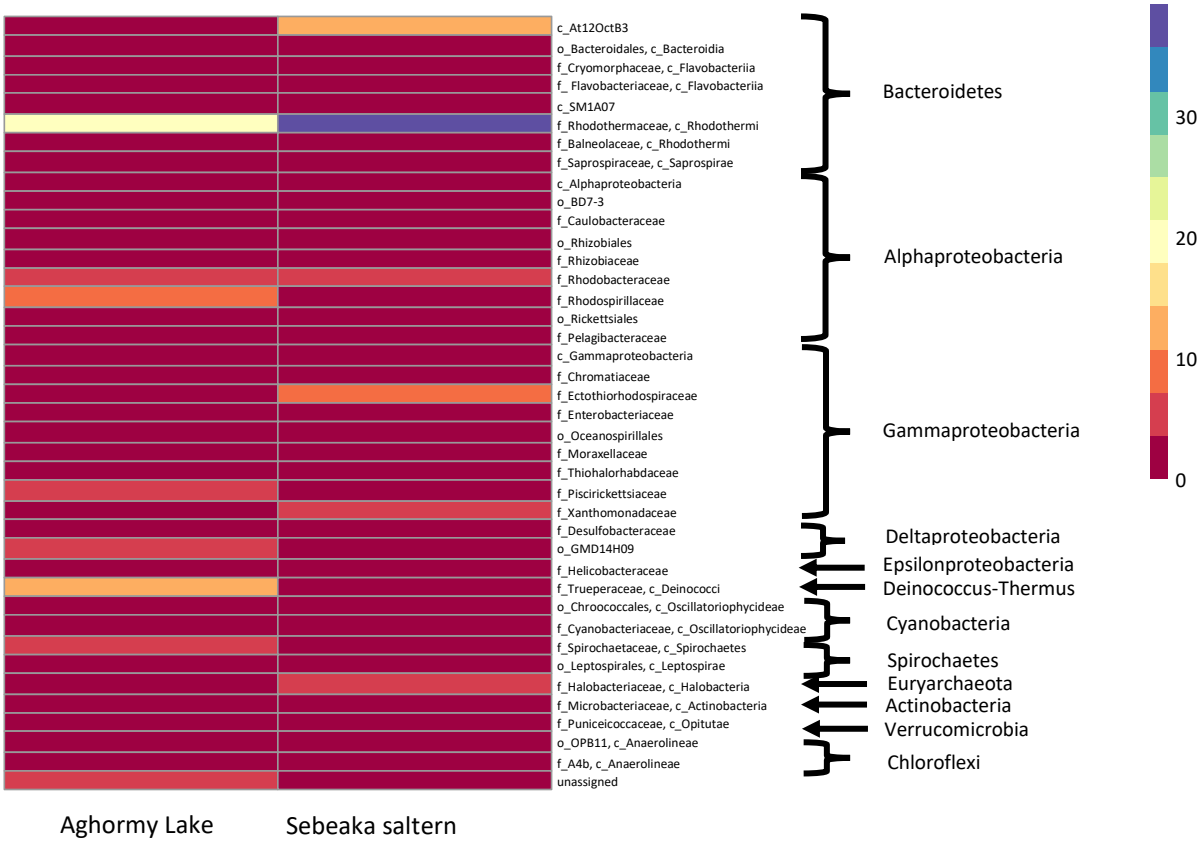


Fig.7.4 Distribution of prokaryotic families, each of which had abundance of ≥ 0.5 % of total sequences in Aghormy Lake and/or Sebeaka saltern, across different phyla. Higher taxonomic ranks were represented when OTUs could not be assigned to family level. Abbreviations: (f: family, o: order, c: class).

7.3.2. Differential halophilic Bacteroidetes in Aghormy Lake and Sebeaka saltern

Rhodothermaceae was the most dominant family in both sites (Fig.7.4). Within Rhodothermaceae, OTUs related to known halophilic members, such as *Salisaeta,* constituted 13.4% of bacterial population in Aghormy Lake; and those related to the extreme halophile *Salinibacter* represented 4.4% of the bacterial population there. *Salisaeta* was similarly reported in several hypersaline environments [254,255,256,257,258].

In case of the Sebeaka saltern, family Rhodothermaceae, represented about 39.32% of total bacterial phylotype composition in the brine. It was also observed that most of the detected OTUs were assigned to the extreme halophile *Salinibacter* representing 38.86% of the bacterial population in Sebeaka saltern. Different studies revealed that *Salinibacter* is widespread in hypersaline environments. It also shares many characteristics with extremely halophilic archaea [259]. The red color in the saltern could be partially attributed to this red halophile as seen in other crystallizers [1].

The abundance of *Salisaeta*-like phylotypes in Aghormy Lake in contrast to the *Salinibacter*-like phylotypes in Sebeaka saltern may reflect differences in salt concentration and chemical composition in both sites. The only known species belonging to genus *Salisaeta* is *Salisaeta longa,* a halophile requiring 10% NaCl and Mg^{2+} cation for optimum growth [260]. The dominance of Mg^{2+} cations in the lake [13] could support the growth of *S. longa* which can tolerate $MgCl_2.6H_2O$ up to 20% [260]. On the other hand, as *Salinibacter* requires 20%-30% NaCl concentrations for optimum growth [259], the abundance of *Salinibacter*-like OTUs in Sebeaka saltern, which is characterized by halite precipitates [20], can be explained.

Flavobacteriaceae (3.23%) and Cryomorphaceae (1.17%) were also detected in both brines. Most genera belonging to both families require NaCl for their growth. Furthermore, in Sebeaka saltern, a high percentage of detected OTUs were assigned to class AT12OctB3 (11%), which was only identified in hypersaline environments such as Lake Tebenquiche, Chile [253], Tuz Lake, Turkey [261], and hypersaline lakes of the Tibetan Plateau [262]. In fact, OTUs belonging to this class in addition to Rhodothermaceae were also found to be the major Bacteriodetes-related OTUs in the hypersaline Lake Gasikule, Qaidam Basin, Tibetan Plateau [262]. Finally, family Balneolaceae represented 0.6% and 1% of sequenced reads in Aghormy Lake and Sebeaka saltern, respectively. This family was first proposed in 2016 by Xia *et al*, in a study that identified a new moderately halophilic member which was isolated from a solar saltern [263].

7.3.3. Predominance of Deinococcus-Thermus and Spirochaetes-like phylotypes in Aghormy Lake

Family Trueperaceae, phylum Deinococcus-Thermus, represented 12.2% of Aghormy Lake sequences (Fig.7.1), which makes it the third most abundant phylogenetic group in the brine. Currently, this family is represented by the single species, *Truepera radiovictrix*, known by its extreme resistance to ionizing radiation [264]. Intriguingly, a study on radionuclides in Siwa Oasis groundwater has recorded radioactivity in levels higher than the limits set by WHO [265]. In addition, some *Deinococcus* members have shown a significant resistance to high concentrations of Cd [266], which has been previously reported in Aghormy Lake sediments [14]. These environmental records may support our findings of the thriving of polyextremophilic Deinococci-like phylotypes in Aghormy Lake.

Spirochaetes-like phylotypes were recorded only in Aghormy brine and constituted 6.8% of total reads in the collected sample (Fig.7.1). In fact, different members belonging to the family Spirochaetaceae, which is particularly abundant in the lake, have been recorded in hypersaline environments as symbionts with sulfate-reducing bacteria and Cyanobacteria [267,268,269]. It is worth noting that different species of the genus *Spirochaeta* are commonly present in H_2S-rich environments [270]. Interestingly, sequences related to cyanobacteria and sulfate-reducing Deltaproteobacteria were detected in Aghormy brine (Fig.7.4), reflecting possible microbial community interactions between the different bacterial species in the lake and the presence of an active sulfur cycle in the lake.

7.3.4. Differential abundance of Alpha- and Gammaproteobacteria in Aghormy Lake and Sebeaka saltern

Our findings have shown that Aghormy Lake was characterized by the abundance of Alphaproteobacteria (22.6%), followed by Gammaproteobacteria (14.2%) and Deltaproteobacteria (6.5%) (Fig.7.4). On the other hand, Gammaproteobacteria recorded 15.7% of Sebeaka saltern reads, followed by Alphaproteobacteria (9.7%) (Fig.7.4). Predominance of Proteobacteria has been reported in several hypersaline environments [216,251,252,271]. Rhodospirillaceae was the most abundant family of Alphaproteobacteria in Aghormy brine (9.8%) (Fig.7.4), represented mainly by the genus *Rhodovibrio* (8.16%). *Rhodovibrio* is an anoxygenic phototrophic purple non-sulfur bacterium and is known to thrive in anoxic zones of hypersaline Mg^{2+} rich environments that are exposed to light [272,273], as in Aghormy Lake [13] . It is likely that the low abundance of *Rhodovibrio*-like OTUs in Sebeaka saltern is because of the limited ability of *Rhodovibrio* to thrive in salinities higher than 240 ppt [274], which is less than that of Sebeaka saltern, where salt reaches saturation. Rhodobacteraceae, known by the presence of its halotolerant and halophilic members [272], has also shown high abundance in both sites (3.8% and 4.97% in Aghormy Lake and Sebeaka saltern, respectively).

Both sites showed differential composition of phylotypes related to the two Gammaproteobacterial families, Chromatiaceae and Ectothiorhodospiraceae (Fig.7.4). Family Chromatiaceae represented 1.2% of the total sequences in Aghormy brine. The genus *Halochromatium* constituted 0.5% of the total sequenced reads. A study on microbial mats in a hypersaline lake in Washington had found that the percentage of this genus in particular varied based on the variations in seasonal sulfur cycling [275]. Members belonging to family Chromatiaceae are known as phototrophic purple sulfur bacteria. They can grow in illuminated, anoxic and sulfide containing aquatic environments, as they can utilize sulfide or sulfur as electron donors, oxidizing them into sulfate [218]. On the other hand, although the extremely halophilic family Ectothiorhodospiraceae [276] was detected significantly in Aghormy Lake (1.75%), it showed

a considerable higher abundance in Sebeaka saltern representing 8% of the total sequenced reads there (Fig.7.4). The extremely halophilic and alkaliphilic genus *Halorhodospira* [276] in particular constituted 6.4% of the total reads. This may explain its abundance in the slightly alkaline saltern rather than the neutral Aghormy Lake.

Moreover, Aghormy Lake showed significant abundance of the family Thiohalorhabdaceae (0.9%) represented mainly by its extremely halophilic member *Thiohalorhabdus* [277] (0.85%). In case of Sebeaka saltern, order Oceonospirillales was also considerably abundant (0.8%). most members of this order are known to be halophilic or halotolerant [278].

Deltaproteobacteria-like phylotypes were mainly represented in Aghormy brine (Fig.7.4). Sequences belonging to order GMD14H09 and Desulfobacteraceae represented 4% and 1.7%, respectively, of the total sequences in Aghormy Lake (Fig7.4). As sulfate is one of the major anions in the lake [13,243], occurrence of Desulfobacteraceae-like OTUs could be expected. Desulfobacteraceae and the majority of Deltaproteobacteria are known to be sulfate-reducing bacteria that can oxidize sulfate partially or completely into sulfides [279]. This may explain co-occurrence of Desulfobacteraceae with current recorded Chromatiaceae and Spirochaetes (Fig.7. 4), which favour H_2S-rich environments [218,280].

7.3.5. Cyanobacteria-like OTUs assigned to halophilic members in both sites

Cyanobacteria-like OTUs represented 3.9% and 5.4% of the total sequenced reads in Aghormy Lake and Sebeaka saltern, respectively (Fig.7.1). Most of the Cyanobacteria-like sequences were assigned to halophilic members (Fig.7.4). Scanning microscopy and field observation studies have reported unclassified cyanobacterial mats along the margins of the Aghormy Lake [14,242]. In fact, microbial mats dominated by Cyanobacteria have been commonly observed in both thalassohaline and athalassohaline environments worldwide [281,282]. When metazoan grazers are restricted, microbial mats can prosper in hypersaline environments [283]. Typically, Cyanobacteria are the main phototrophs in these mats [279] in which they are associated with sulfur bacteria and sulfate-reducing bacteria [284], which were both detected in Aghormy Lake.

The majority of detected OTUs belonged to the family Cyanobacteriaceae, with almost all of them falling into the genus *Cyanothece* (2.4% and 3% in Aghormy brine and Sebeka salterns, respectively). Our findings support previous findings in which genus *Cyanothece* was isolated from different athalassohaline and thalassohaline habitats [285]. Although members of this genus can be isolated from freshwaters as well, halophilic members isolated from solar evaporation ponds were found to form a separate cluster from other *Cyanothece* members [286].

7.3.6. Occurrence of archaeal family, Halobacteriaceae, in Sebeaka saltern

Despite using bacterial 16S rRNA designed primers, the archaeal phylum Euryarchaeota constituted 4.83% of the total phyla distribution in Sebeaka saltern. All identified OTUs were from the family Halobacteriacaea with one abundant OTU representing 0.6% of the microbial population in the saltern. The most abundant genus was the square-cell shaped extremely halophilic archaeon *Haloquadratum* (1.3%) followed by *Halorubrum* (0.59%). Both genera are widespread in hypersaline habitats giving shades of red to these environments due to their content of red carotenoids (alpha-bacterioruberin and derivatives) [287]. This could also be a major contributor to the water red color in the saltern.

It is noteworthy that the percentage of identified Archaeal phylotypes is far from being accurate and the actual archaeal population may be much higher than the small percentage captured by the used bacterial designed primers. In fact, most studies on salterns and extremely hypersaline environments had found that archaea are the most dominant in the prokaryotic community such as in the Northern arm of Great Salt Lake [288], Dead Sea [251] and Lake Tanatar trona crystallizer [216].

Furthermore, although OTUs belonging to archaeal lineages were almost negligible in Aghormy Lake, we cannot rule out the existence of archaea there, as no archaeal 16S rRNA designed primers were used in this study.

7.3.7. Biotechnological potential of identified phylotypes in both studied brines

Aghormy Lake and Sebeaka saltern comprise diverse halophiles. These halophiles and their products have a great potential to be used in different industries, in bioremediation processes and in the production of potential antimicrobial agents. Different halophiles were found to produce excessive amounts of exopolysaccharides (EPS). EPSs can be used as gelling agents, emulsifiers, and in microbially enhanced oil recovery [289]. Different archaea from family Halobacteriaceae, significantly abundant in Sebeaka saltern, were found to produce EPSs that can be used in the emulsification of petroleum [290]. Moreover, phylotypes belonging to the cyanobacterium *Cyanothece* were found in both studied brines. It has been reported that extracted sulfated polysaccharides from *Cyanothece* spp. inhibit the adhesion of *Heliobacter pylori* to gastric epithelial cells [291]. In fact, glycosides were extracted from Aghormy microbial mats [292]. Although the type of detected glycosides was not determined, yet as these microbial mats possessed antibacterial activity against some bacterial isolates such as *Vibrio cholera* [292], a potential antimicrobial effect against different pathogens could exist. In addition, *Cyanothece* extracts have shown strong cytotoxic activity against different cancer cell lines [293,294], increasing their potential as

anticancer agents. Other antimicrobial agents were detected in halophiles. For instance, halocin which targets Na+/H+ antiporter causing cell lysis is produced by *Halurubrum* archaeon belonging to Halobacteriaceae [295].

Carotenoids are natural pigments with strong antioxidant and immune boosting activities. Thus, they have been utilized in different pharmaceuticals as antitumor and prevention agents in heart diseases. Their production from halophiles gained a considerable attention since their production from the halophilic alga *Dunaliella salina* as no fear of contamination by non-halophiles is present, in addition to the simple extraction techniques using hypoosmotic conditions for direct lysis of the cells [296]. Here, Halobacteriaceae archaea, *Salinibacter* and *Halorhodospira* spp, all identified in Sebeaka saltern, can produce carotenoids [289,297].

Furthermore, liposomes synthesized form ether-linked lipids derived from halophilic archaea such as those from family Halobacteriaceae showed higher survival rates than those synthesized from fatty acid derivatives. Hence, they can be utilized as a better alternative to conventional liposomes in delivering different compounds to their cellular target sites [290].

Moreover, osmolytes or compatible solutes can stabilize whole cells or biomolecules. They can also be used as protective agents against different stresses [290]. Different halophiles identified in both studied sites can be utilized in the production of osmolytes. *Halorhodospira* genus was found to thrive in Sebeaka saltern. Previous studies showed that expression of methyltransferase genes from *Halorhodospira halochloris* involved in the production of the compatible solute glycine betaine in *E. coli* led to its intracellular accumulation and increase in salt tolerance [290]. Ectoine, another compatible solute, was first discovered in *H. halochloris* [289]. Ectoine has been as a stabilizer for different enzymes, a moisturizer in cosmetic preparations and has a potential as a stabilizer in PCR [290]. Members of *Rhodovibrio*, which is abundant in Aghormy Lake, could be potentially used in glycine betaine and ectoine production as well [297]. In addition, Aghormy Lake showed an abundance in Chromatiaceae phylotypes and in *Halochromatium* genus in particular. *Halochromatium salexigens* was found to accumulate different osmolytes such as glycine betaine, sucrose and N-acetyl-glutaminylglutamine amide [297].

7.4. Conclusions

Common halophilic prokaryotic consortia were found to be shared in both athalassohaline Aghormy Lake in Siwa Oasis and thalassohaline Sebeaka saltern at the eastern part of Bardawil Lagoon on the north coast of the Sinai Peninsula. Yet, each brine possesses unique prokaryotic phylotypes, reflecting the differences in biogeographical and physicochemical properties for each site. The prokaryotic consortium in Aghormy Lake showed higher species richness than in Sebeaka saltern. Bacteroidetes and Proteobacteria were the main phyla recorded in both brines, with predominance of Rhodothermaceae (Bacteroidetes). Abundance of *Salisaeta* and *Salinibacter*

phylotypes was recorded in Aghormy Lake and Sebeaka saltern, respectively. An interesting consortium, including Deinococci, Spirochaetes and Desulfobacteraceae was confined to Aghormy Lake. The salt saturation and crystallization in Sebeaka saltern may account for the higher abundance of OTUs belonging to certain halophiles such as *Salinibacter*, Ectothiorhodospiraceae, class AT12OctB3 and Halobacteriaceae. Many of the identified phylotypes have a great potential to be used in different biotechnological applications.

Conclusions and future prospects

In this study, we characterized a thermostable nitrilase, NitraS-ATII isolated from the extreme hypersaline environment Atlantis II Deep brine pool in the Red Sea. In addition to its unique thermostability profile, NitraS-ATII showed tolerance to different heavy metals especially those found in the LCL. This nitrilase holds a biotechnological potential in bioremediation processes in which enzymes with higher stability profiles are needed to withstand different harsh environmental conditions.

Our analysis on integrons revealed the high abundance of integrons and CALINs in halophilic genomes and metagenomes. CALINs were more abundant than complete integrons in both genomes and metagenomes, but their higher prevalence in metagenomes could be attributed to the incomplete nature of metagenomic contigs. All identified integrons belonged to new classes of integrons, and no IntIs belonging to classes 1-4 were identified in all our datasets indicating the absence or at least the rarity of these known classes in hypersaline environments. The prevalence of integrons and CALINs in halophilic genomes (17.5%) was much greater than the very low abundance in the hypersaline metagenomes either by using the IntegronFinder software or by using a PRC screening approach. For instance, the latter just ended up with two positive clones in the created hypersaline AGH fosmid library. This could be attributed mainly to the limitations of metagenomic studies such as the small sizes of obtained contigs.

Most gene cassette ORFs encode for hypothetical proteins limiting our ability to identify the role of these cassettes in adaptation to hypersaline aquatic environments. Nevertheless, TA systems were abundant in all examined halophilic genomes and metagenomes strengthening their suggestive role in stabilizing integron systems. However, identified TA systems were not just confined to large gene cassettes as reported, but rather in most identified integrons and CALINs. Moreover, the plethora of different IS elements within or nearby integrons and CALINs may account for the high prevalence of integrons within halophiles. Many of the identified ISs can mobilize with a presumed rolling circle replication mechanism and other types were shown to be able to mobilize adjacent genomic structures with unknown mechanisms. We have also identified group II introns belonging to the integron-associated group IIC-*att*C, beside identifying CALIN-associated group IIB introns for the first time in the extreme halophile *H. halochloris* and in a hypersaline metagenome. The clustering of different MGEs, especially those that require single-stranded secondary structures for their function suggests putative interactions between these transposable elements and different cellular processes that require ssDNA structures such as replication, transcription and conjugation. The absence of IS elements in the analyzed metagenomes could be due to the coincidence of many transposable elements at contig breaks.

It appears that metagenomic studies need further sophisticated tools to optimize metagenomic assemblies and to allow further explorations of complex genomic structures that could not be fully comprehended based on studying metagenomic contigs.

In addition, unfortunately our trial to assess the excision ability of two identified IntIs from hypersaline metagenomes did not meet with success. However, we cannot yet determine whether these IntIs were truly non-functional or that the used assay was not optimized to measure their true potential. In fact, there is a necessity to develop recombination assays for different IntIs, as the developed assays were used to assess the recombination activity for few IntI classes, while the list of new IntIs is continuously growing waiting for their experimental characterization. Integron systems could have an intricate and complex regulatory mechanisms that makes their characterization even harder. Our analysis of predicted P_{intI} promoters for identified *int*I genes, revealed a high frequency of ArgR transcription factors binding sites, which may suggest a possible role in the regulation of IntI expression and recombination reaction. However, this would definitely need further experimental evidence.

Moreover, we have identified archaeal integrons within halophilic and thermophilic archaea for the first time. The high similarity between the archaeal IntI and another bacterial one, both from hypersaline environments suggests possible horizontal transfer between microbial species. We have also detected arrays of successive *att*C-sites within archaeal metagenomes and genomes, that do not resemble the typical structure of gene cassette arrays in integrons or CALINs. Thus, unraveling the role of these structures needs further investigation.

The abundance of integrons in halophiles and their association with MGEs that could allow their mobilization within genomes and between species indicates their active role in microbial adaptation to their hypersaline environments. This role appears to be regulated by fine and complex regulatory networks. However, the role of integrons could be limited in archaea due to the rarity of archaeal integrons.

Finally, we compared prokaryotic communities in two different hypersaline environments: the athalassohaline Aghormy Lake in Siwa Oasis and the thalassohaline Sebeake saltern at the vicinity of Bardawil Lagoon in North Sinai. Common halophilic phylotypes were found; yet, each brine had its own unique prokaryotic consortoum, which reflects the differences in biogeographical and physicochemical properties of each site. Bacteroidetes and Proteobacteria were the main phyla recorded in both brines, with predominance of family Rhodothermaceae. Aghormy Lake showed an interesting consortium, including Deinococci, Spirochaetes and Desulfobacteraceae. Salt saturation and crystallization in Sebeaka saltern may account for the higher abundance of OTUs belonging to certain halophiles such as *Salinibacter*, Ectothiorhodospiraceae, class AT12OctB3 and Halobacteriaceae. Different identified phylotypes in both studied sites may have a potential to be exploted in different biotechnological applications.

References

1. Ventosa A. Halophilic Microorganisms. Berlin Heidelberg: Springer-Verlag; 2004. doi:10.1007/978-3-662-07656-9.

2. Javor B. Hypersaline Environments: Microbiology and Biogeochemistry. Berlin, Heidelberg: Springer Science & Business Media; 1989.

3. Galinski EA, Trüper HG. Microbial behaviour in salt-stressed ecosystems. FEMS Microbiol Rev. 1994;15:95–108.

4. Maheshwari DK, Saraf M, editors. Halophiles: Biodiversity and Sustainable Exploitation. Springer International Publishing; 2015. doi:10.1007/978-3-319-14595-2.

5. Paul S, Bag SK, Das S, Harvill ET, Dutta C. Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. Genome Biol. 2008;9:R70.

6. 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. doi:10.1371/journal.pone.0042872.

7. Winckler G, Aeschbach-Hertig W, Kipfer R, Botz R, Rübel AP, Bayer R, et al. Constraints on origin and evolution of Red Sea brines from helium and argon isotopes. Earth Planet Sci Lett. 2001;184:671–83. doi:10.1016/S0012-821X(00)00345-9.

8. Laurila TE, Hannington MD, Leybourne M, Petersen S, Devey CW, Garbe-Schönberg D. New insights into the mineralogy of the Atlantis II Deep metalliferous sediments, Red Sea. Geochem Geophys Geosystems. 2015;16:4449–78.

9. Antunes A, Ngugi DK, Stingl U. Microbiology of the Red Sea (and other) deep-sea anoxic brine lakes. Environ Microbiol Rep. 2011;3:416–33.

10. Anschutz P, Blanc G, Monnin C, Boulègue J. Geochemical dynamics of the Atlantis II Deep (Red Sea): II. Composition of metalliferous sediment pore waters. Geochim Cosmochim Acta. 2000;64:3995–4006.

11. Swift SA, Bower AS, Schmitt RW. Vertical, horizontal, and temporal changes in temperature in the Atlantis II and Discovery hot brine pools, Red Sea. Deep Sea Res Part Oceanogr Res Pap. 2012;64:118–28. doi:10.1016/j.dsr.2012.02.006.

12. Hartmann M, Scholten JC, Stoffers P, Wehner F. Hydrographic structure of brine-filled deeps in the Red Sea—new results from the Shaban, Kebrit, Atlantis II, and Discovery Deep. Mar Geol. 1998;144:311–30. doi:10.1016/S0025-3227(97)00055-8.

13. El-Sayed SA, Allam KA, Salama MH, El Begawy H. Investigation of Chemical and Radiochemical Fingerprints of Water Resources in Siwa Oasis, Western Desert, Egypt. Arab J Nucl Sci Appl. 2017;50:158–78.

14. Abd El-Karim MS, Goher ME-S. Heavy metal concentrations in cyanobacterial mats and underlying sediments in some northern western desert lakes of Egypt. J Fish Aquat Sci. 2016;11:163–73.

15. Krumgalz BS, Hornung H, Oren OH. The study of a natural hypersaline lagoon in a desert area (the Bardawil Lagoon in Northern Sinai). Estuar Coast Mar Sci. 1980;10:403–15.

16. Shaer HME. Potential Role of Sabkhas in Egypt: An Overview. In: Ashraf M, Ozturk M, Athar HR, editors. Salinity and Water Stress. Springer Netherlands; 2009. p. 221–8. doi:10.1007/978-1-4020-9065-3_23.

17. Khalil MT, Shaltout KH. Lake Bardawil and Zaranik Protected Area. 1st edition. Egypt: National Biodiversity Unit. No. 15; 2006. https://www.researchgate.net/publication/280598623_Lake_Bardawil_and_Zaranik_Protected_Area. Accessed 24 Nov 2019.

18. Rabeh SA, Azab EA, Aly MM. Studies on bacterioplankton and inhibitory strains of aquatic actinomycetes in Lake Bardawil, Egypt. World J Microbiol Biotechnol. 2007;23:167–76.

19. Anders DE, Handford CR, Hite RJ, Kyle JR, Lowenstein TK, Posey HH, et al. Evaporites, Petroleum and Mineral Resources. 1st edition. USA: Elsevier; 1991. https://www.elsevier.com/books/evaporites-petroleum-and-mineral-resources/melvin/978-0-444-88680-4. Accessed 20 Nov 2019.

20. Embabi NS. Bardawil Lake and the Surrounding Sabkhas. In: Embabi NS, editor. Landscapes and Landforms of Egypt: Landforms and Evolution. Cham: Springer International Publishing; 2018. p. 291–303. doi:10.1007/978-3-319-65661-8_22.

21. Shaheen SMA. Geoenvironmental studies on El-Bardawil Lagoon and its surroundings, North Sinai, Egypt. Ph. D Thesis. Mansoura University, Faculty of Science; 1998.

22. Sayed A, Ghazy MA, Ferreira AJS, Setubal JC, Chambergo FS, Ouf A, et al. A novel mercuric reductase from the unique deep brine environment of Atlantis II in the Red Sea. J Biol Chem. 2014;289:1675–87.

23. Mohamed YM, Ghazy MA, Sayed A, Ouf A, El-Dorry H, Siam R. Isolation and characterization of a heavy metal-resistant, thermophilic esterase from a Red Sea brine pool. Sci Rep. 2013;3:3358.

24. Elbehery AHA, Leak DJ, Siam R. Novel thermostable antibiotic resistance enzymes from the Atlantis II Deep Red Sea brine pool. Microb Biotechnol. 2017;10:189–202.

25. Kaul P, Banerjee A, Banerjee UC. Nitrile Hydrolases. In: Polaina J, MacCabe AP, editors. Industrial Enzymes. Springer Netherlands; 2007. p. 531–47. http://link.springer.com.library.aucegypt.edu:2048/chapter/10.1007/1-4020-5377-0_30. Accessed 21 Feb 2013.

26. Nigam VK, Arfi T, Kumar V, Shukla P. Bioengineering of Nitrilases Towards Its Use as Green Catalyst: Applications and Perspectives. Indian J Microbiol. 2017;57:131–8.

27. Bayer S, Birkemeyer C, Ballschmiter M. A nitrilase from a metagenomic library acts regioselectively on aliphatic dinitriles. Appl Microbiol Biotechnol. 2011;89:91–8.

28. Liu Z-Q, Zhou M, Zhang X-H, Xu J-M, Xue Y-P, Zheng Y-G. Biosynthesis of iminodiacetic acid from iminodiacetonitrile by immobilized recombinant *Escherichia coli* harboring nitrilase. J Mol Microbiol Biotechnol. 2012;22:35–47.

29. Podar M, Eads JR, Richardson TH. Evolution of a microbial nitrilase gene family: a comparative and environmental genomics study. BMC Evol Biol. 2005;5:42.

30. Estepa J, Luque-Almagro VM, Manso I, Escribano MP, Martínez-Luque M, Castillo F, et al. The nit1C gene cluster of *Pseudomonas pseudoalcaligenes* CECT5344 involved in assimilation of nitriles is essential for growth on cyanide. Environ Microbiol Rep. 2012;4:326–34.

31. Jones LB, Wang X, Gullapalli JS, Kunz DA. Characterization of the Nit6803 nitrilase homolog from the cyanotroph *Pseudomonas fluorescens* NCIMB 11764. Biochem Biophys Rep. 2021;25:100893.

32. Gerzova L, Videnska P, Faldynova M, Sedlar K, Provaznik I, Cizek A, et al. Characterization of Microbiota Composition and Presence of Selected Antibiotic Resistance Genes in Carriage Water of Ornamental Fish. PLOS ONE. 2014;9:e103865.

33. Guo X, Xia R, Han N, Xu H. Genetic diversity analyses of class 1 integrons and their associated antimicrobial resistance genes in Enterobacteriaceae strains recovered from aquatic habitats in China. Lett Appl Microbiol. 2011;52:667–75.

34. Grindley NDF, Whiteson KL, Rice PA. Mechanisms of Site-Specific Recombination. Annu Rev Biochem. 2006;75:567–605.

35. Parks AR, Peters JE. Conservative Site-Specific Recombination. In: Bell E, editor. Molecular Life Sciences. Springer New York; 2014. p. 1–10. doi:10.1007/978-1-4614-6436-5_165-1.

36. Hirano N, Muroi T, Takahashi H, Haruki M. Site-specific recombinases as tools for heterologous gene integration. Appl Microbiol Biotechnol. 2011;92:227–39.

37. Nunes-Duby SE, Kwon HJ, Tirumalai RS, Ellenberger T, Landy A. Similarities and differences among 105 members of the Int family of site-specific recombinases. Nucleic Acids Res. 1998;26:391–406.

38. Messier N, Roy PH. Integron Integrases Possess a Unique Additional Domain Necessary for Activity. J Bacteriol. 2001;183:6699–706.

39. Cambray G, Guerout A-M, Mazel D. Integrons. Annu Rev Genet. 2010;44:141–66.

40. Stokes HW, Hall RM. A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons. Mol Microbiol. 1989;3:1669–83.

41. Gillings MR. Integrons: Past, Present, and Future. Microbiol Mol Biol Rev. 2014;78:257–77.

42. Holmes AJ, Gillings MR, Nield BS, Mabbutt BC, Nevalainen KMH, Stokes HW. The gene cassette metagenome is a basic resource for bacterial genome evolution. Environ Microbiol. 2003;5:383–94.

43. Jové T, Da Re S, Denis F, Mazel D, Ploy M-C. Inverse correlation between promoter strength and excision activity in class 1 integrons. PLoS Genet. 2010;6:e1000793.

44. Jacquier H, Zaoui C, Pors M-JS, Mazel D, Berçot B. Translation regulation of integrons gene cassette expression by the attC sites. Mol Microbiol. 2009;72:1475–86.

45. Escudero JA, Loot C, Mazel D. Integrons as Adaptive Devices. In: Rampelotto PH, editor. Molecular Mechanisms of Microbial Evolution. Cham: Springer International Publishing; 2018. p. 199–239. doi:10.1007/978-3-319-69078-0_9.

46. Nield BS, Holmes AJ, Gillings MR, Recchia GD, Mabbutt BC, Nevalainen KMH, et al. Recovery of new integron classes from environmental DNA. FEMS Microbiol Lett. 2001;195:59–65.

47. MacDonald D, Demarre G, Bouvier M, Mazel D, Gopaul DN. Structural basis for broad DNA-specificity in integron recombination. Nature. 2006;440:1157–62.

48. Partridge SR, Tsafnat G, Coiera E, Iredell JR. Gene cassettes and cassette arrays in mobile resistance integrons. FEMS Microbiol Rev. 2009;33:757–84.

49. Boucher Y, Labbate M, Koenig JE, Stokes HW. Integrons: mobilizable platforms that promote genetic diversity in bacteria. Trends Microbiol. 2007;15:301–9.

50. 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.

51. Nivina A, Escudero JA, Vit C, Mazel D, Loot C. Efficiency of integron cassette insertion in correct orientation is ensured by the interplay of the three unpaired features of attC recombination sites. Nucleic Acids Res. 2016;44:7792–803.

52. Loot C, Bikard D, Rachlin A, Mazel D. Cellular pathways controlling integron cassette site folding. EMBO J. 29:2623–34.

53. Larouche A, Roy PH. Effect of *att*C structure on cassette excision by integron integrases. Mob DNA. 2011;2:3.

54. Collis CM, Kim M-J, Stokes HW, Hall RM. Integron-encoded IntI integrases preferentially recognize the adjacent cognate *att*I site in recombination with a 59-be site. Mol Microbiol. 2002;46:1415–27.

55. Frumerie C, Ducos-Galand M, Gopaul DN, Mazel D. The relaxed requirements of the integron cleavage site allow predictable changes in integron target specificity. Nucleic Acids Res. 2010;38:559–69.

56. Labbate M, Boucher Y, Joss MJ, Michael CA, Gillings MR, Stokes HW. Use of chromosomal integron arrays as a phylogenetic typing system for *Vibrio cholerae* pandemic strains. Microbiol Read Engl. 2007;153 Pt 5:1488–98.

57. Loot C, Ducos-Galand M, Escudero JA, Bouvier M, Mazel D. Replicative resolution of integron cassette insertion. Nucleic Acids Res. 2012;40:8361–70.

58. Mazel D. Integrons: agents of bacterial evolution. Nat Rev Microbiol. 2006;4:608–20.

59. Nardelli M, Scalzo PM, Ramírez MS, Quiroga MP, Cassini MH, Centrón D. Class 1 Integrons in Environments with Different Degrees of Urbanization. PLoS ONE. 2012;7. doi:10.1371/journal.pone.0039223.

60. Gillings M, Boucher Y, Labbate M, Holmes A, Krishnan S, Holley M, et al. The Evolution of Class 1 Integrons and the Rise of Antibiotic Resistance. J Bacteriol. 2008;190:5095–100.

61. Antelo V, Romero H, Batista S. Detection of integron integrase genes on King George Island, Antarctica. Chin J Polar Sci. 2015;:30–7.

62. Makowska N, Zawierucha K, Nadobna P, Piątek-Bajan K, Krajewska A, Szwedyk J, et al. Occurrence of integrons and antibiotic resistance genes in cryoconite and ice of Svalbard, Greenland, and the Caucasus glaciers. Sci Total Environ. 2020;716:137022.

63. Abella J, Bielen A, Huang L, Delmont TO, Vujaklija D, Duran R, et al. Integron diversity in marine environments. Environ Sci Pollut Res Int. 2015;22:15360–9.

64. Koenig JE, Sharp C, Dlutek M, Curtis B, Joss M, Boucher Y, et al. Integron Gene Cassettes and Degradation of Compounds Associated with Industrial Waste: The Case of the Sydney Tar Ponds. PLOS ONE. 2009;4:e5276.

65. Melderen LV, Bast MSD. Bacterial Toxin–Antitoxin Systems: More Than Selfish Entities? PLOS Genet. 2009;5:e1000437.

66. Unterholzner SJ, Poppenberger B, Rozhon W. Toxin–antitoxin systems. Mob Genet Elem. 2013;3:e26219.

67. Gerdes K, editor. Prokaryotic Toxin-Antitoxins. Berlin Heidelberg: Springer-Verlag; 2013. doi:10.1007/978-3-642-33253-1.

68. Chan WT, Espinosa M, Yeo CC. Keeping the Wolves at Bay: Antitoxins of Prokaryotic Type II Toxin-Antitoxin Systems. Front Mol Biosci. 2016;3. doi:10.3389/fmolb.2016.00009.

69. Rowe-Magnus DA, Guerout A-M, Biskri L, Bouige P, Mazel D. Comparative analysis of superintegrons: engineering extensive genetic diversity in the Vibrionaceae. Genome Res. 2003;13:428–42.

70. Szekeres S, Dauti M, Wilde C, Mazel D, Rowe-Magnus DA. Chromosomal toxin-antitoxin loci can diminish large-scale genome reductions in the absence of selection. Mol Microbiol. 2007;63:1588–605.

71. Labbate M, Case RJ, Stokes HW. The Integron/Gene Cassette System: An Active Player in Bacterial Adaptation. In: Gogarten MB, Gogarten JP, Olendzenski LC, editors. Horizontal Gene Transfer: Genomes in Flux. Totowa, NJ: Humana Press; 2009. p. 103–25. doi:10.1007/978-1-60327-853-9_6.

72. Iqbal N, Guérout A-M, Krin E, Roux FL, Mazel D. Comprehensive Functional Analysis of the 18 *Vibrio cholerae* N16961 Toxin-Antitoxin Systems Substantiates Their Role in Stabilizing the Superintegron. J Bacteriol. 2015;197:2150–9.

73. Sberro H, Leavitt A, Kiro R, Koh E, Peleg Y, Qimron U, et al. Discovery of Functional Toxin/Antitoxin Systems in Bacteria by Shotgun Cloning. Mol Cell. 2013;50:136–48.

74. Mazel D, Dychinco B, Webb VA, Davies J. A Distinctive Class of Integron in the *Vibrio cholerae* Genome. Science. 1998;280:605–8.

75. Tansirichaiya S, Rahman MA, Antepowicz A, Mullany P, Roberts AP. Detection of Novel Integrons in the Metagenome of Human Saliva. PLOS ONE. 2016;11:e0157605.

76. Buongermino Pereira M, Österlund T, Eriksson KM, Backhaus T, Axelson-Fisk M, Kristiansson E. A comprehensive survey of integron-associated genes present in metagenomes. BMC Genomics. 2020;21:495.

77. Rosewarne CP, Pettigrove V, Stokes HW, Parsons YM. Class 1 integrons in benthic bacterial communities: abundance, association with Tn402-like transposition modules and evidence for coselection with heavy-metal resistance. FEMS Microbiol Ecol. 2010;72:35–46.

78. Stokes HW, Holmes AJ, Nield BS, Holley MP, Nevalainen KMH, Mabbutt BC, et al. Gene Cassette PCR: Sequence-Independent Recovery of Entire Genes from Environmental DNA. Appl Environ Microbiol. 2001;67:5240–6.

79. Tsafnat G, Coiera E, Partridge SR, Schaeffer J, Iredell JR. Context-driven discovery of gene cassettes in mobile integrons using a computational grammar. BMC Bioinformatics. 2009;10:281.

80. Joss MJ, Koenig JE, Labbate M, Polz MF, Gillings MR, Stokes HW, et al. ACID: annotation of cassette and integron data. BMC Bioinformatics. 2009;10:118.

81. Rowe-Magnus DA, Guerout A-M, Biskri L, Bouige P, Mazel D. Comparative Analysis of Superintegrons: Engineering Extensive Genetic Diversity in the Vibrionaceae. Genome Res. 2003;13:428–42.

82. Bouvier M, Demarre G, Mazel D. Integron cassette insertion: a recombination process involving a folded single strand substrate. EMBO J. 2005;24:4356–67.

83. Vit C, Loot C, Escudero JA, Nivina A, Mazel D. Integron Identification in Bacterial Genomes and Cassette Recombination Assays. In: de la Cruz F, editor. Horizontal Gene Transfer: Methods and Protocols. New York, NY: Springer US; 2020. p. 189–208. doi:10.1007/978-1-4939-9877-7_14.

84. Bouvier M, Ducos-Galand M, Loot C, Bikard D, Mazel D. Structural Features of Single-Stranded Integron Cassette *att*C Sites and Their Role in Strand Selection. PLoS Genet. 2009;5. doi:10.1371/journal.pgen.1000632.

85. Léon G, Roy PH. Excision and Integration of Cassettes by an Integron Integrase of Nitrosomonas europaea. J Bacteriol. 2003;185:2036–41.

86. Baharoglu Z, Bikard D, Mazel D. Conjugative DNA Transfer Induces the Bacterial SOS Response and Promotes Antibiotic Resistance Development through Integron Activation. PLoS Genet. 2010;6. doi:10.1371/journal.pgen.1001165.

87. Cambray G, Sanchez-Alberola N, Campoy S, Guerin É, Da Re S, González-Zorn B, et al. Prevalence of SOS-mediated control of integron integrase expression as an adaptive trait of chromosomal and mobile integrons. Mob DNA. 2011;2:6.

88. Guerin E, Cambray G, Sanchez-Alberola N, Campoy S, Erill I, Da Re S, et al. The SOS response controls integron recombination. Science. 2009;324:1034.

89. Strugeon E, Tilloy V, Ploy M-C, Re SD. The Stringent Response Promotes Antibiotic Resistance Dissemination by Regulating Integron Integrase Expression in Biofilms. mBio. 2016;7. doi:10.1128/mBio.00868-16.

90. Baharoglu Z, Krin E, Mazel D. Connecting environment and genome plasticity in the characterization of transformation-induced SOS regulation and carbon catabolite control of the *Vibrio cholerae* integron integrase. J Bacteriol. 2012;194:1659–67.

91. Cagle CA, Shearer JES, Summers AO. Regulation of the integrase and cassette promoters of the class 1 integron by nucleoid-associated proteins. Microbiol Read Engl. 2011;157 Pt 10:2841–53.

92. Ghaly TM, Geoghegan JL, Tetu SG, Gillings MR. The Peril and Promise of Integrons: Beyond Antibiotic Resistance. Trends Microbiol. 2020. doi:10.1016/j.tim.2019.12.002.

93. Bikard D, Julié-Galau S, Cambray G, Mazel D. The synthetic integron: an *in vivo* genetic shuffling device. Nucleic Acids Res. 2010;38:e153–e153.

94. Rowe-Magnus DA. Integrase-directed recovery of functional genes from genomic libraries. Nucleic Acids Res. 2009;37:e118.

95. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol. 2005;3:722–32.

96. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile Genetic Elements Associated with Antimicrobial Resistance. Clin Microbiol Rev. 2018;31. doi:10.1128/CMR.00088-17.

97. Hallet B, Sherratt DJ. Transposition and site-specific recombination: adapting DNA cut-and-paste mechanisms to a variety of genetic rearrangements. FEMS Microbiol Rev. 1997;21:157–78.

98. Johnson CM, Grossman AD. Integrative and Conjugative Elements (ICEs): What They Do and How They Work. Annu Rev Genet. 2015;49:577–601.

99. Siguier P, Gourbeyre E, Varani A, Ton-Hoang B, Chandler M. Everyman's Guide to Bacterial Insertion Sequences. Mob DNA III. 2015;:555–90.

100. Hickman AB, Dyda F. DNA Transposition at Work. Chem Rev. 2016;116:12758–84.

101. 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.

102. Babakhani S, Oloomi M. Transposons: the agents of antibiotic resistance in bacteria. J Basic Microbiol. 2018;58:905–17.

103. Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev. 2009;33:376–93.

104. Wozniak RAF, Waldor MK. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat Rev Microbiol. 2010;8:552–63.

105. Tourasse NJ, Kolstø A-B. Survey of group I and group II introns in 29 sequenced genomes of the *Bacillus cereus* group: insights into their spread and evolution. Nucleic Acids Res. 2008;36:4529–48.

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

107. Lambowitz AM, Belfort M. Introns as Mobile Genetic Elements. Annu Rev Biochem. 1993;62:587–622.

108. Lambowitz AM, Zimmerly S. Mobile Group II Introns. Annu Rev Genet. 2004;38:1–35.

109. Zimmerly S, Hausner G, Wu X. Phylogenetic relationships among group II intron ORFs. Nucleic Acids Res. 2001;29:1238–50.

110. Hausner G, Hafez M, Edgell DR. Bacterial group I introns: mobile RNA catalysts. Mob DNA. 2014;5:8.

111. Domingues S, da Silva GJ, Nielsen KM. Integrons: Vehicles and pathways for horizontal dissemination in bacteria. Mob Genet Elem. 2012;2:211–23.

112. Gombac F, Riccio ML, Rossolini GM, Lagatolla C, Tonin E, Monti-Bragadin C, et al. Molecular Characterization of Integrons in Epidemiologically Unrelated Clinical Isolates of *Acinetobacter baumannii* from Italian Hospitals Reveals a Limited Diversity of Gene Cassette Arrays. Antimicrob Agents Chemother. 2002;46:3665–8.

113. Moura A, Henriques I, Ribeiro R, Correia A. Prevalence and characterization of integrons from bacteria isolated from a slaughterhouse wastewater treatment plant. J Antimicrob Chemother. 2007;60:1243–50.

114. Nandi S, Maurer JJ, Hofacre C, Summers AO. Gram-positive bacteria are a major reservoir of Class 1 antibiotic resistance integrons in poultry litter. Proc Natl Acad Sci U S A. 2004;101:7118–22.

115. Doublet B, Praud K, Weill F-X, Cloeckaert A. Association of IS*26*-composite transposons and complex In4-type integrons generates novel multidrug resistance loci in *Salmonella* genomic island 1. J Antimicrob Chemother. 2009;63:282–9.

116. Roy Chowdhury P, Ingold A, Vanegas N, Martínez E, Merlino J, Merkier AK, et al. Dissemination of Multiple Drug Resistance Genes by Class 1 Integrons in *Klebsiella pneumoniae* Isolates from Four Countries: a Comparative Study ▿. Antimicrob Agents Chemother. 2011;55:3140–9.

117. Toleman MA, Bennett PM, Walsh TR. ISCR elements: novel gene-capturing systems of the 21st century? Microbiol Mol Biol Rev MMBR. 2006;70:296–316.

118. Toleman MA, Bennett PM, Walsh TR. Common regions e.g. orf513 and antibiotic resistance: IS91-like elements evolving complex class 1 integrons. J Antimicrob Chemother. 2006;58:1–6.

119. Ramírez MS, Piñeiro S, Argentinian Integron Study Group, Centrón D. Novel insights about class 2 integrons from experimental and genomic epidemiology. Antimicrob Agents Chemother. 2010;54:699–706.

120. Collis CM, Kim M-J, Partridge SR, Stokes HW, Hall RM. Characterization of the Class 3 Integron and the Site-Specific Recombination System It Determines. J Bacteriol. 2002;184:3017–26.

121. Sørum H, Roberts MC, Crosa JH. Identification and cloning of a tetracycline resistance gene from the fish pathogen *Vibrio salmonicida*. Antimicrob Agents Chemother. 1992;36:611–5.

122. Domingues S, Nielsen KM, da Silva GJ. The blaIMP-5-carrying integron in a clinical *Acinetobacter baumannii* strain is flanked by miniature inverted-repeat transposable elements (MITEs). J Antimicrob Chemother. 2011;66:2667–8.

123. Gillings MR, Labbate M, Sajjad A, Giguère NJ, Holley MP, Stokes HW. Mobilization of a Tn402-Like Class 1 Integron with a Novel Cassette Array via Flanking Miniature Inverted-Repeat Transposable Element-Like Structures. Appl Environ Microbiol. 2009;75:6002–4.

124. Poirel L, Carrër A, Pitout JD, Nordmann P. Integron Mobilization Unit as a Source of Mobility of Antibiotic Resistance Genes. Antimicrob Agents Chemother. 2009;53:2492–8.

125. Meng H, Zhang Z, Chen M, Su Y, Li L, Miyoshi S-I, et al. Characterization and horizontal transfer of class 1 integrons in *Salmonella* strains isolated from food products of animal origin. Int J Food Microbiol. 2011;149:274–7.

126. Hu Z, Zhao W-H. Identification of plasmid- and integron-borne blaIMP-1 and blaIMP-10 in clinical isolates of *Serratia marcescens*. J Med Microbiol. 2009;58 Pt 2:217–21.

127. Hochhut B, Lotfi Y, Mazel D, Faruque SM, Woodgate R, Waldor MK. Molecular Analysis of Antibiotic Resistance Gene Clusters in *Vibrio cholerae* O139 and O1 SXT Constins. Antimicrob Agents Chemother. 2001;45:2991–3000.

128. Léon G, Roy PH. Potential role of group IIC-*att*C introns in integron cassette formation. J Bacteriol. 2009;191:6040–51.

129. Quiroga C, Centrón D. Using genomic data to determine the diversity and distribution of target site motifs recognized by class C-*att*C group II introns. J Mol Evol. 2009;68:539–49.

130. Domingues S, Harms K, Fricke WF, Johnsen PJ, Silva GJ da, Nielsen KM. Natural Transformation Facilitates Transfer of Transposons, Integrons and Gene Cassettes between Bacterial Species. PLOS Pathog. 2012;8:e1002837.

131. Gestal AM, Liew EF, Coleman NV. Natural transformation with synthetic gene cassettes: new tools for integron research and biotechnology. Microbiol Read Engl. 2011;157 Pt 12:3349–60.

132. Ramírez MS, Merkier AK, Quiroga MP, Centrón D. Acinetobacter baumannii is able to gain and maintain a plasmid harbouring In35 found in Enterobacteriaceae isolates from Argentina. Curr Microbiol. 2012;64:211–3.

133. Schmieger H, Schicklmaier P. Transduction of multiple drug resistance of Salmonella enterica serovar typhimurium DT104. FEMS Microbiol Lett. 1999;170:251–6.

134. Yeom S-J, Kim H-J, Lee J-K, Kim D-E, Oh D-K. An amino acid at position 142 in nitrilase from Rhodococcus rhodochrous ATCC 33278 determines the substrate specificity for aliphatic and aromatic nitriles. Biochem J. 2008;415:401–7.

135. Robertson DE, Chaplin JA, DeSantis G, Podar M, Madden M, Chi E, et al. Exploring Nitrilase Sequence Space for Enantioselective Catalysis. Appl Environ Microbiol. 2004;70:2429–36.

136. Gupta N, Balomajumder C, Agarwal VK. Enzymatic mechanism and biochemistry for cyanide degradation: A review. J Hazard Mater. 2010;176:1–13.

137. Cowan D, Cramp R, Pereira R, Graham D, Almatawah Q. Biochemistry and biotechnology of mesophilic and thermophilic nitrile metabolizing enzymes. Extermophiles. 1998;2:207–16.

138. Kim J-S, Tiwari MK, Moon H-J, Jeya M, Ramu T, Oh D-K, et al. Identification and characterization of a novel nitrilase from *Pseudomonas fluorescens* Pf-5. Appl Microbiol Biotechnol. 2009;83:273–83.

139. Sonbol SA, Ferreira AJS, Siam R. Red Sea Atlantis II brine pool nitrilase with unique thermostability profile and heavy metal tolerance. BMC Biotechnol. 2016;16:14.

140. Wang H, Li G, Li M, Wei D, Wang X. A novel nitrilase from *Rhodobacter sphaeroides* LHS-305: cloning, heterologous expression and biochemical characterization. World J Microbiol Biotechnol. 2014;30:245–52.

141. Laemmli UK. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature. 1970;227:680–5.

142. Goyal SS, Rains DW, Huffaker RC. Determination of ammonium ion by fluorometry or spectrophotometry after on-line derivatization with o-phthalaldehyde. Anal Chem. 1988;60:175–9.

143. Banerjee A, Sharma R, Banerjee UC. A rapid and sensitive fluorometric assay method for the determination of nitrilase activity. Biotechnol Appl Biochem. 2003;37:289–93.

144. Kr zel A, Lesniak W, Jezowska-Bojczuk M, Mlynarz P, Brasuñ J, Kozlowski H, et al. Coordination of heavy metals by dithiothreitol, a commonly used thiol group protectant. J Inorg Biochem. 2001;84:77–88.

145. Lee H-J, Nam HJ, Noh D-Y. Dimeric Mercury(II) Chloride Complex of Sulfur-rich Ligand : Synthesis and X-ray Crystal Structure of trans-[{Hg(m-Cl)Cl(dPhEDT-DTT)}2] · (CH3CN)2. Bull Korean Chem Soc. 1999;20:1368–70.

146. Kelley LA, Sternberg MJE. Protein structure prediction on the web a case study using the Phyre server. Nat Protoc. 2009;4:363–71.

147. Kumar S, Nussinov R. Salt bridge stability in monomeric proteins. J Mol Biol. 1999;293:1241–55.

148. Kumar S, Tsai CJ, Ma B, Nussinov R. Contribution of salt bridges toward protein thermostability. J Biomol Struct Dyn. 2000;17 Suppl 1:79–85.

149. Kumar S, Nussinov R. Relationship between ion pair geometries and electrostatic strengths in proteins. Biophys J. 2002;83:1595–612.

150. Sarakatsannis JN, Duan Y. Statistical characterization of salt bridges in proteins. Proteins. 2005;60:732–9.

151. Podar M, Eads J, Richardson T. Evolution of a microbial nitrilase gene family: a comparative and environmental genomics study. BMC Evol Biol. 2005;5:42. doi:10.1186/1471-2148-5-42.

152. Yancey PH. Organic osmolytes as compatible, metabolic and counteracting cytoprotectants in high osmolarity and other stresses. J Exp Biol. 2005;208 Pt 15:2819–30.

153. Santos H, da Costa MS. Compatible solutes of organisms that live in hot saline environments. Environ Microbiol. 2002;4:501–9.

154. Yu H, Huang H. Engineering proteins for thermostability through rigidifying flexible sites. Biotechnol Adv. 2014;32:308–15.

155. Reed CJ, Lewis H, Trejo E, Winston V, Evilia C. Protein Adaptations in Archaeal Extremophiles. Archaea. 2013;2013:e373275.

156. Lam SY, Yeung RCY, Yu T-H, Sze K-H, Wong K-B. A Rigidifying Salt-Bridge Favors the Activity of Thermophilic Enzyme at High Temperatures at the Expense of Low-Temperature Activity. PLoS Biol. 2011;9:e1001027.

157. Bosshard HR, Marti DN, Jelesarov I. Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings. J Mol Recognit JMR. 2004;17:1–16.

158. Lee C-W, Wang H-J, Hwang J-K, Tseng C-P. Protein Thermal Stability Enhancement by Designing Salt Bridges: A Combined Computational and Experimental Study. PLoS ONE. 2014;9:e112751.

159. Purification and Characterization of Nitrilase from *Fusarium solani* IMI196840 | Protein Engineering Group. http://loschmidt.chemi.muni.cz/peg/publications/purification-and-characterization-of-nitrilase-from-fusarium-solani-imi196840/. Accessed 17 Jun 2014.

160. Zhang Z-J, Xu J-H, He Y-C, Ouyang L-M, Liu Y-Y. Cloning and biochemical properties of a highly thermostable and enantioselective nitrilase from *Alcaligenes* sp. ECU0401 and its potential for (R)-(−)-mandelic acid production. Bioprocess Biosyst Eng. 2011;34:315–22. doi:10.1007/s00449-010-0473-z.

161. Mazel D, Dychinco B, Webb VA, Davies J. Antibiotic Resistance in the ECOR Collection: Integrons and Identification of a Novel aad Gene. Antimicrob Agents Chemother. 2000;44:1568–74.

162. Elsaied H, Stokes HW, Nakamura T, Kitamura K, Fuse H, Maruyama A. Novel and diverse integron integrase genes and integron-like gene cassettes are prevalent in deep-sea hydrothermal vents. Environ Microbiol. 2007;9:2298–312.

163. Abdulaziz AM, Faid AM. Evaluation of the groundwater resources potential of Siwa Oasis using three-dimensional multilayer groundwater flow model, Mersa Matruh Governorate, Egypt. Arab J Geosci. 2015;8:659–75.

164. 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.

165. Hartmann M, Scholten JC, Stoffers P, Wehner F. Hydrographic structure of brine-filled deeps in the Red Sea—new results from the Shaban, Kebrit, Atlantis II, and Discovery Deep. Mar Geol. 1998;144:311–30.

166. Schmidt M, Botz R, Faber E, Schmitt M, Poggenburg J, Garbe-Schönberg D, et al. High-resolution methane profiles across anoxic brine–seawater boundaries in the Atlantis-II, Discovery, and Kebrit Deeps (Red Sea). Chem Geol. 2003;200:359–75.

167. 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.

168. 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;17:273.

169. Moura A, Soares M, Pereira C, Leitão N, Henriques I, Correia A. INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. Bioinformatics. 2009;25:1096–8.

170. Ferreira AJS, Siam R, Setubal JC, Moustafa A, Sayed A, Chambergo FS, et al. Core Microbial Functional Activities in Ocean Environments Revealed by Global Metagenomic Profiling Analyses. PLOS ONE. 2014;9:e97338.

171. Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. DNA Res. 2008;15:387–96.

172. Noguchi H, Park J, Takagi T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res. 2006;34:5623–30.

173. 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.

174. 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.

175. Jové T, Da Re S, Tabesse A, Gassama-Sow A, Ploy M-C. Gene Expression in Class 2 Integrons Is SOS-Independent and Involves Two Pc Promoters. Front Microbiol. 2017;8. doi:10.3389/fmicb.2017.01499.

176. Gillings MR, Holley MP, Stokes HW, Holmes AJ. Integrons in *Xanthomonas*: A source of species genome diversity. Proc Natl Acad Sci. 2005;102:4419–24.

177. Darmon E, Leach DRF. Bacterial Genome Instability. Microbiol Mol Biol Rev MMBR. 2014;78:1–39.

178. Lee H, Doak TG, Popodi E, Foster PL, Tang H. Insertion sequence-caused large-scale rearrangements in the genome of Escherichia coli. Nucleic Acids Res. 2016;44:7109–19.

179. Loukas A, Kappas I, Abatzopoulos TJ. HaloDom: a new database of halophiles across all life domains. J Biol Res-Thessalon. 2018;25:2.

180. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016;44 Database issue:D7–19.

181. Okonechnikov K, Golosova O, Fursov M, the UGENE team. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012;28:1166–7.

182. Xie Z, Tang H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics. 2017;33:3340–7.

183. 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.

184. Rapa RA, Labbate M. The function of integron-associated gene cassettes in *Vibrio* species: the tip of the iceberg. Front Microbiol. 2013;4. doi:10.3389/fmicb.2013.00385.

185. Sonbol S, Siam R. The association of group IIB intron with integrons in hypersaline environments. Mob DNA. 2021;12:8.

186. Cho S, Cho Y-B, Kang TJ, Kim SC, Palsson B, Cho B-K. The architecture of ArgR-DNA complexes at the genome-scale in Escherichia coli. Nucleic Acids Res. 2015;43:3079–88.

187. Schröder J, Tauch A. Transcriptional regulation of gene expression in Corynebacterium glutamicum: the role of global, master and local regulators in the modular and hierarchical gene regulatory network. FEMS Microbiol Rev. 2010;34:685–737.

188. Halsey CR, Lei S, Wax JK, Lehman MK, Nuxoll AS, Steinke L, et al. Amino Acid Catabolism in Staphylococcus aureus and the Function of Carbon Catabolite Repression. mBio. 2017;8. doi:10.1128/mBio.01434-16.

189. Janion C, Sikora A, Nowosielska A, Grzesiuk E. Induction of the SOS response in starved Escherichia coli. Environ Mol Mutagen. 2002;40:129–33.

190. Janion C. Some aspects of the SOS response system--a critical survey. Acta Biochim Pol. 2001;48:599–610.

191. Escudero JA, Loot C, Parissi V, Nivina A, Bouchier C, Mazel D. Unmasking the ancestral activity of integron integrases reveals a smooth evolutionary transition during functional innovation. Nat Commun. 2016;7:10937.

192. Guhathakurta A, Summers D. Involvement of ArgR and PepA in the pairing of ColE1 dimer resolution sites. Microbiol Read Engl. 1995;141 (Pt 5):1163–71.

193. 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.

194. 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.

195. 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.

196. 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.

197. Favrot L, Blanchard JS, Vergnolle O. Bacterial GCN5-Related N-Acetyltransferases: From Resistance to Regulation. Biochemistry. 2016;55:989–1002.

198. Solano-Gutierrez JS, Pino C, Robledo J. Toxin–antitoxin systems shows variability among Mycobacterium tuberculosis lineages. FEMS Microbiol Lett. 2019;366. doi:10.1093/femsle/fny276.

199. Thakur Z, Dharra R, Saini V, Kumar A, Mehta PK. Insights from the protein-protein interaction network analysis of *Mycobacterium tuberculosis* toxin-antitoxin systems. Bioinformation. 2017;13:380–7.

200. Montecillo AD, Ilag LL, Kohli A, Lantican NB, Raymundo AK. Archaeal community profiling of a mixed culture from mud and water slurry from a solfataric mudspring as revealed by a community proteome approach. Philipp Sci Lett. 2019;12:1–10.

201. Stårsta M, Hammarlöf DL, Wäneskog M, Schlegel S, Xu F, Gynnå AH, et al. RHS-elements function as type II toxin-antitoxin modules that regulate intra-macrophage replication of Salmonella Typhimurium. PLOS Genet. 2020;16:e1008607.

202. Ferrari A, Maggi S, Montanini B, Levante A, Lazzi C, Yamaguchi Y, et al. Identification and first characterization of DinJ-YafQ toxin-antitoxin systems in *Lactobacillus* species of biotechnological interest. Sci Rep. 2019;9:7645.

203. Jurėnas D, Garcia-Pino A, Van Melderen L. Novel toxins from type II toxin-antitoxin systems with acetyltransferase activity. Plasmid. 2017;93:30–5.

204. Harms A, Stanger FV, Scheu PD, de Jong IG, Goepfert A, Glatter T, et al. Adenylylation of Gyrase and Topo IV by FicT Toxins Disrupts Bacterial DNA Topology. Cell Rep. 2015;12:1497–507.

205. Castro-Roa D, Garcia-Pino A, De Gieter S, van Nuland NAJ, Loris R, Zenkin N. The Fic protein Doc uses an inverted substrate to phosphorylate and inactivate EF-Tu. Nat Chem Biol. 2013;9:811–7.

206. Saavedra De Bast M, Mine N, Van Melderen L. Chromosomal toxin-antitoxin systems may act as antiaddiction modules. J Bacteriol. 2008;190:4603–9.

207. 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.

208. Léon G, Quiroga C, Centrón D, Roy PH. Diversity and strength of internal outward-oriented promoters in group IIC-*att*C introns. Nucleic Acids Res. 2010;38:8196–207.

209. Martínez-Abarca F, Toro N. Group II introns in the bacterial world. Mol Microbiol. 2000;38:917–26.

210. 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.

211. 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.

212. 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.

213. Qin PZ, Pyle AM. The architectural organization and mechanistic function of group II intron structural elements. Curr Opin Struct Biol. 1998;8:301–8.

214. 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.

215. 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.

216. 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.

217. Sorokin DY. The Microbial Sulfur Cycle at Extremely Haloalkaline Conditions of Soda Lakes. Front Microbiol. 2011;2. doi:10.3389/fmicb.2011.00044.

218. 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.

219. 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.

220. Dai L, Toor N, Olson R, Keeping A, Zimmerly S. Database for mobile group II introns. Nucleic Acids Res. 2003;31:424–6.

221. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31:3406–15.

222. 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.

223. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.

224. Byun Y, Han K. PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots. Bioinformatics. 2009;25:1435–7.

225. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: A Sequence Logo Generator. Genome Res. 2004;14:1188–90.

226. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016;33:1870–4.

227. 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.

228. 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.

229. 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.

230. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.

231. 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;3:1–21.

232. Rodríguez-Martínez J-M, Nordmann P, Poirel L. Group IIC Intron with an Unusual Target of Integration in Enterobacter cloacae. J Bacteriol. 2012;194:150–60.

233. 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.

234. Grieb MS, Nivina A, Cheeseman BL, Hartmann A, Mazel D, Schlierf M. Dynamic stepwise opening of integron *att*C DNA hairpins by SSB prevents toxicity and ensures functionality. Nucleic Acids Res. 2017;45:10555–63.

235. 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.

236. 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.

237. Siguier P, Gourbeyre E, Chandler M. Known knowns, known unknowns and unknown unknowns in prokaryotic transposition. Curr Opin Microbiol. 2017;38:171–80.

238. Trinh TQ, Sinden RR. Preferential DNA secondary structure mutagenesis in the lagging strand of replication in E. coli. Nature. 1991;352:544–7.

239. Gerday C, Glansdorff N, editors. EXTREMOPHILES. Oxford, United Kingdom: Eolss Publishers Co Ltd; 2009.

240. Edbeib MF, Wahab RA, Huyop F. Halophiles: biology, adaptation, and their role in decontamination of hypersaline environments. World J Microbiol Biotechnol. 2016;32:135.

241. Ventosa A, Oren A, Ma Y, editors. Halophiles and Hypersaline Environments: Current Research and Future Trends. Berlin Heidelberg: Springer-Verlag; 2011. doi:10.1007/978-3-642-20198-1.

242. Taher AG, Abdel-Motelib A. Microbial stabilization of sediments in a recent Salina, Lake Aghormi, Siwa Oasis, Egypt. Facies. 2014;60:45–52.

243. Hedia MRM. Assessment of Drainage Water Quality in Siwa Oasis and its Suitability for Reuse in Agricultural Irrigation. Egypt J Soil Sci. 2015;55:501–15.

244. Levy Y. Description and mode of formation of the supratidal evaporite facies in northern Sinai coastal plain. J Sediment Res. 1977;47:463–74.

245. Javor B. The Fate of Carbon and Sulfur in Hypersaline Environments. In: Javor B, editor. Hypersaline Environments: Microbiology and Biogeochemistry. 1st edition. Berlin, Heidelberg: Springer Science & Business Media; 1989. p. 53–76. doi:10.1007/978-3-642-74370-2_5.

246. Morgan RS, El-Araby A, Ghabour TK, Abd Elwahed MS. Characterization of the Wetlands of El-Bardawil and its Surrounding Environment, North Sinai, Egypt. Am-Eurasian J Agric Environ Sci. 2011;10:207–15.

247. Zahran HH. Diversity, adaptation and activity of the bacterial flora in saline environments. Biol Fertil Soils. 1997;25:211–23.

248. Pade N, Hagemann M. Salt acclimation of cyanobacteria and their application in biotechnology. Life. 2014;5:25–49.

249. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinforma Oxf Engl. 2009;25:2078–9.

250. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7:335–6.

251. Jacob JH, Hussein EI, Shakhatreh MAK, Cornelison CT. Microbial community analysis of the hypersaline water of the Dead Sea using high-throughput amplicon sequencing. Microbiol Open. 2017;6.

252. Montoya L, Vizioli C, Rodríguez N, Rastoll MJ, Amils R, Marin I. Microbial community composition of Tirez lagoon (Spain), a highly sulfated athalassohaline environment. Aquat Biosyst. 2013;9:19.

253. Demergasso C, Escudero L, Casamayor EO, Chong G, Balagué V, Pedrós-Alió C. Novelty and spatio-temporal heterogeneity in the bacterial diversity of hypersaline Lake Tebenquiche (Salar de Atacama). Extrem Life Extreme Cond. 2008;12:491–504.

254. Sørensen KB, Canfield DE, Teske AP, Oren A. Community composition of a hypersaline endoevaporitic microbial mat. Appl Env Microbiol. 2005;71:7352–65.

255. Clementino MM, Vieira RP, Cardoso AM, Nascimento APA, Silveira CB, Riva TC, et al. Prokaryotic diversity in one of the largest hypersaline coastal lagoons in the world. Extremophiles. 2008;12:595.

256. Sahl JW, Pace NR, Spear JR. Comparative molecular analysis of endoevaporitic microbial communities. Appl Env Microbiol. 2008;74:6444–6.

257. Baati H, Guermazi S, Amdouni R, Gharsallah N, Sghir A, Ammar E. Prokaryotic diversity of a Tunisian multipond solar saltern. Extrem Life Extreme Cond. 2008;12:505–18.

258. Baati H, Guermazi S, Gharsallah N, Sghir A, Ammar E. Novel prokaryotic diversity in sediments of Tunisian multipond solar saltern. Res Microbiol. 2010;161:573–82.

259. Krieg NR, Ludwig W, Euzéby J, Whitman WB. Phylum XIV. Bacteroidetes phyl. nov. In: Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, Ward NL, et al., editors. Bergey's Manual® of Systematic Bacteriology. Springer New York; 2010. p. 25–469. doi:10.1007/978-0-387-68572-4_3.

260. Vaisman N, Oren A. *Salisaeta longa* gen. nov., sp. nov., a red, halophilic member of the Bacteroidetes. Int J Syst Evol Microbiol. 2009;59 Pt 10:2571–4.

261. Mutlu MB, Martínez-García M, Santos F, Peña A, Guven K, Antón J. Prokaryotic diversity in Tuz Lake, a hypersaline environment in Inland Turkey. FEMS Microbiol Ecol. 2008;65:474–83. doi:10.1111/j.1574-6941.2008.00510.x.

262. Zhong Z-P, Liu Y, Miao L-L, Wang F, Chu L-M, Wang J-L, et al. Prokaryotic community structure driven by salinity and ionic concentrations in plateau lakes of the Tibetan plateau. Appl Environ Microbiol. 2016;:AEM.03332-15.

263. Xia J, Ling S-K, Wang X-Q, Chen G-J, Du Z-J. Aliifodinibius halophilus sp. nov., a moderately halophilic member of the genus *Aliifodinibius*, and proposal of Balneolaceae fam. nov. Int J Syst Evol Microbiol. 2016;66:2225–33.

264. Albuquerque L, Simões C, Nobre MF, Pino NM, Battista JR, Silva MT, et al. *Truepera radiovictrix* gen. nov., sp. nov., a new radiation resistant species and the proposal of Trueperaceae fam. nov. FEMS Microbiol Lett. 2005;247:161–9.

265. Saleh IH, Elnaggar AM, Ibrahim HZ. Risk assessment of radionuclides in groundwater in Siwa Oasis, Egypt. Int J Adv Res. 2016;4:1459–66.

266. Jaafar R, Al-Sulami A, Al-Taee A, Aldoghachi F, Napes S. Biosorption and Bioaccumulation of Some Heavy Metals by *Deinococcus radiodurans* Isolated from Soil in Basra Governorate- Iraq. J Biotechnol Biomater. 2015;5:1–5.

267. Dubinina G, Grabovich M, Leshcheva N, Rainey FA, Gavrish E. *Spirochaeta perfilievii* sp. nov., an oxygen-tolerant, sulfide-oxidizing, sulfur- and thiosulfate-reducing spirochaete isolated from a saline spring. Int J Syst Evol Microbiol. 2011;61:110–7.

268. Liolos K, Abt B, Scheuner C, Teshima H, Held B, Lapidus A, et al. Complete genome sequence of the halophilic bacterium Spirochaeta africana type strain (Z-7692T) from the alkaline Lake Magadi in the East African Rift. Stand Genomic Sci. 2013;8:165–76.

269. Dubinina G, Grabovich M, Leshcheva N, Gronow S, Gavrish E, Akimov V. *Spirochaeta sinaica* sp. nov., a halophilic spirochaete isolated from a cyanobacterial mat. Int J Syst Evol Microbiol. 2015;65:3872–7.

270. Paster BJ. Spirochaetes. In: Whitman WB, Rainey F, Kämpfer P, Trujillo M, Chun J, DeVos P, et al., editors. Bergey's Manual of Systematics of Archaea and Bacteria. John Wiley & Sons, Inc.; 2015. doi:10.1002/9781118960608.pbm00023.

271. Weimer B, Rompato G, Parnell J, Gann R, Ganesan B, Navas C, et al. Microbial biodiversity of Great Salt Lake, Utah. Nat Resour Environ Issues. 2009;15. https://digitalcommons.usu.edu/nrei/vol15/iss1/3.

272. Garrity GM, Bell JA, Lilburn T. Class I. Alphaproteobacteria class. nov. In: Brenner DJ, Krieg NR, Staley JT, editors. Bergey's Manual® of Systematic Bacteriology. Springer US; 2005. p. 1–574. doi:10.1007/0-387-29298-5_1.

273. Imhoff JF. Anoxygenic Phototrophic Bacteria from Extreme Environments. Mod Top Phototrophic Prokaryotes. 2017;:427–80.

274. Oren A. Diversity of Halophiles. In: Horikoshi K, editor. Extremophiles Handbook. Tokyo: Springer Japan; 2011. p. 309–25. doi:10.1007/978-4-431-53898-1_14.

275. Lindemann SR, Moran JJ, Stegen JC, Renslow RS, Hutchison JR, Cole JK, et al. The epsomitic phototrophic microbial mat of Hot Lake, Washington: community structural responses to seasonal cycling. Front Microbiol. 2013;4:323.

276. Brenner DJ, Krieg NR, Staley JR. Bergey's Manual® of Systematic Bacteriology: Volume 2: The Proteobacteria, Part B: The Gammaproteobacteria. Springer Science & Business Media; 2007.

277. Sorokin DY, Tourova TP, Galinski EA, Muyzer G, Kuenen JG. *Thiohalorhabdus denitrificans* gen. nov., sp. nov., an extremely halophilic, sulfur-oxidizing, deep-lineage gammaproteobacterium from hypersaline habitats. Int J Syst Evol Microbiol. 2008;58 Pt 12:2890–7.

278. Brenner DJ, Krieg NR, Staley JR. Bergey's Manual® of Systematic Bacteriology: Volume 2: The Proteobacteria, Part B: The Gammaproteobacteria. Springer Science & Business Media; 2007.

279. Baumgartner LK, Reid RP, Dupraz C, Decho AW, Buckley DH, Spear JR, et al. Sulfate reducing bacteria in microbial mats: Changing paradigms, new discoveries. Sediment Geol. 2006;185:131–45.

280. Paster BJ. Phylum XV. Spirochaetes Garrity and Holt 2001. In: Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, Ward NL, et al., editors. Bergey's Manual® of Systematic Bacteriology: Volume Four The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes. New York, NY: Springer New York; 2010. p. 471–566. doi:10.1007/978-0-387-68572-4_4.

281. Javor B. Cyanobacteria. In: Javor B, editor. Hypersaline Environments: Microbiology and Biogeochemistry. 1st edition. Berlin, Heidelberg: Springer Science & Business Media; 1989. p. 134–46. doi:10.1007/978-3-642-74370-2_9.

282. Whitton BA, Potts M. The Ecology of Cyanobacteria: Their Diversity in Time and Space. Springer Science & Business Media; 2007.

283. Schieber J, Bose PK, Eriksson PG, Banerjee S, Sarkar S, Altermann W, et al. Atlas of Microbial Mat Features Preserved within the Siliciclastic Rock Record. Elsevier; 2007.

284. Seckbach J, Oren A, editors. Microbial Mats: Modern and Ancient Microorganisms in Stratified Systems. Dordrecht Heidelberg London New York: Springer Science & Business Media; 2010.

285. De Philippis R, Margheri MC, Materassi R, Vincenzini M. Potential of Unicellular Cyanobacteria from Saline Environments as Exopolysaccharide Producers. Appl Environ Microbiol. 1998;64:1130–2.

286. Castenholz RW, Wilmotte A, Herdman M, Rippka R, Waterbury JB, Iteman I, et al. Phylum BX. Cyanobacteria. In: Boone DR, Castenholz RW, Garrity GM, editors. Bergey's Manual® of Systematic Bacteriology. Springer New York; 2001. p. 473–599. doi:10.1007/978-0-387-21609-6_27.

287. Garrity GM, Holt JG. Euryarchaeota phy. nov. In: Bergey's Manual of Systematics of Archaea and Bacteria. John Wiley & Sons, Ltd; 2015. doi:10.1002/9781118960608.pbm00014.

288. Oren A. Salt Lakes, Metagenomics of. In: Highlander SK, Rodriguez-Valera F, White BA, editors. Encyclopedia of Metagenomics. Springer US; 2015. p. 577–84. doi:10.1007/978-1-4899-7475-4_42.

289. Oren A. Industrial and environmental applications of halophilic microorganisms. Environ Technol. 2010;31:825–34.

290. Margesin R, Schinner F. Potential of halotolerant and halophilic microorganisms for biotechnology. Extremophiles. 2001;5:73–83.

291. Ascencio F, Gama NL, De Philippis R, Ho B. Effectiveness of *Cyanothece* spp. and Cyanospira capsulata exocellular polysaccharides as antiadhesive agents for blocking attachment of Helicobacter pylori to human gastric cells. Folia Microbiol (Praha). 2004;49:64.

292. Abd El-Karim MS. Chemical Composition and Antimicrobial Activities of Cyanobacterial Mats from Hyper Saline Lakes, Northern Western Desert, Egypt. J Appl Sci. 2016;16:1–10.

293. El Semary NA, Fouda M. Anticancer activity of *Cyanothece* sp. strain extracts from Egypt: First record. Asian Pac J Trop Biomed. 2015;5:992–5.

294. Wa A, Na E-S, Om AE-H, ElTawill G, Dm I. Bioactivity and Cytotoxic Effect of Cyanobacterial Toxin AgainstHepatocellular Carcinoma. 2017.

295. Atanasova NS, Pietilä MK, Oksanen HM. Diverse antimicrobial interactions of halophilic archaea and bacteria extend over geographical distances and cross the domain barrier. MicrobiologyOpen. 2013;2:811–25.

296. Waditee-Sirisattha R, Kageyama H, Takabe T. Halophilic microorganism resources and their applications in industrial and environmental biotechnology. AIMS Microbiol. 2016;2:42.

297. Oren A. Halophilic Microorganisms and their Environments. Springer Netherlands; 2002. doi:10.1007/0-306-48053-0.

298. Demarre G, Frumerie C, Gopaul DN, Mazel D. Identification of key structural determinants of the IntI1 integron integrase that influence *att*C x *att*I1 recombination efficiency. Nucleic Acids Res. 2007;35:6475–89.

299. Sheik CS, Jain S, Dick GJ. Metabolic flexibility of enigmatic SAR324 revealed through metagenomics and metatranscriptomics. Environ Microbiol. 2014;16:304–17.

300. Baker BJ, Lesniewski RA, Dick GJ. Genome-enabled transcriptomics reveals archaeal populations that drive nitrification in a deep-sea hydrothermal plume. ISME J. 2012;6:2269–79.

301. Lesniewski RA, Jain S, Anantharaman K, Schloss PD, Dick GJ. The metatranscriptome of a deep-sea hydrothermal plume is dominated by water column methanotrophs and lithotrophs. ISME J. 2012;6:2257–68.

302. Tang K, Liu K, Jiao N, Zhang Y, Chen C-TA. Functional metagenomic investigations of microbial communities in a shallow-sea hydrothermal system. PloS One. 2013;8:e72958.

303. Mitchell AL, Scheremetjew M, Denise H, Potter S, Tarkowska A, Qureshi M, et al. EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. Nucleic Acids Res. 2018;46:D726–35.

Appendix A: Chapter 4 Supplementary Tables

TableS4.1 Analyzed complete and partial bacterial halophilic genomes

bacterial analyzed genomes	genome size	sequencing status	genome or WGS accession number	plasmids accession numbers if present
Acetohalobium arabaticum DSM 5501	2.4696	complete	NC_014378.1	-
Halothece sp. PCC 7418	4.17917	complete	NC_019779.1	-
Cellulosimicrobium cellulans PSBB019	4.79986	complete	NZ_CP021383.1	-
Desulfohalobium retbaense DSM 5692	2.90957	complete	NC_013223.1	NC_013224.1
Chromohalobacter salexigens DSM 3043	3.66514	complete	NC_007963.1	-
Halorhodospira halophila SL1	2.67845	complete	NC_008789.1	-
Halorhodospira halochloris DSM 1059	2.83456	complete	NZ_AP017372.2	-
Halanaerobium hydrogeniformans	2.61312	complete	NC_014654.1	-
Halanaerobium praevalens DSM 2228	2.30926	complete	NC_017455.1	-
Halobacillus halophilus DSM 2266	4.17177	complete	NC_017668.1	NC_017670.1, NC_017669.1
Halobacteroides halobius DSM 5150	2.64926	complete	NC_019978.1	-
Halomonas elongata DSM 2581	4.06182	complete	NC_014532.2	-
Halomonas titanicae ANRCS81	5.33979	complete	NZ_CP039374.1	-
Halothermothrix orenii H 168	2.57815	complete	NC_011899.1	-
Marinobacter hydrocarbonoclasticus ATCC 49840	3.98677	complete	NC_017067.1	-
Marinobacter hydrocarbonoclasticus VT8	4.77976	complete	NC_008740.1	NC_008738.1, NC_008739.1
Natranaerobius thermophilus JW/NM-WN-LF	3.19145	complete	NC_010718.1	NC_010715.1, NC_010724.1
Nitrosococcus halophilus Nc 4	4.14526	complete	NC_013960.1	NC_013958.1
Nodularia spumigena CCY9414	5.35144	complete	NZ_CP007203.1	-
Nodularia spumigena UHCC 0039	5.38661	complete	NZ_CP020114.1	NZ_CP020115.1
Oceanobacillus iheyensis HTE831	3.63053	complete	NC_004193.1	-
Oceanobacillus iheyensis CHQ24	3.86062	complete	NZ_CP020357.1	-
Salinibacter ruber DSM 13855	3.76289	complete	NC_007677.1	NC_007678.1
Spiribacter salinus M19-40	2.88033	complete	NC_021291.1	-
Celeribacter indicus strain P73	4.969388	complete	NZ_CP004393.1	NZ_CP004394.1, NZ_CP004395.1, NZ_CP004396.1,

				NZ_CP004397.1, NZ_CP004398.1
Flavobacterium arcticum strain SM1502	2.970356	complete	NZ_CP031188.1	-
Haliangium ochraceum DSM 14365	9.44631	complete	NC_013440.1	-
Halomonas aestuarii strain Hb3	3.54389	complete	NZ_CP018139.1	-
Halomonas beimenensis strain NTU-111	4.05303	complete	NZ_CP021435.1	-
Halomonas huangheensis strain BJGMM-B45	4.75814	complete	NZ_CP013106.1	-
Idiomarina loihiensis L2TR	2.83976	complete	NC_006512.1	-
Lentibacillus amyloliquefaciens strain LAM0015	3.85828	complete	NZ_CP013862.1	-
Marinobacter salinus strain Hb8	4.12101	complete	NZ_CP017715.1	-
Martelella endophytica strain YC6887	4.81733	complete	NZ_CP010803.1	-
Pseudomonas salegens strain CECT 8338	3.7961	complete	NZ_LT629787.1	-
Rhodothermus marinus DSM 4252	3.32972	complete	NC_013501.1	NC_013502.1
Salicibibacter kimchii strain NKC1-1	3.6416	complete	NZ_CP031092.1	-
Salicibibacter halophilus strain NKC3-5	3.75417	complete	NZ_CP035485.1	-
Salinicoccus halodurans strain H3B36	2.7569	complete	NZ_CP011366.1	-
Spiribacter curvatus strain UAH-SP71	1.92663	complete	NC_022664.1	-
Spiribacter roseus strain SSL50	1.96126	complete	NZ_CP016382.1	-
Tetragenococcus halophilus NBRC 12172	2.43853	complete	NC_016052.1	-
Virgibacillus dokdonensis strain 21D	4.2896	complete	NZ_CP018622.1	-
Virgibacillus halodenitrificans strain PDB-F2	3.95258	complete	NZ_CP017962.1	NZ_CP017963.1
Virgibacillus phasianinus strain LM2416	4.071214	complete	NZ_CP022315.1	-
Ectothiorhodospira haloalkaliphila A	3.46013	partial	NZ_CP007268.1	-
Alteribacillus bidgolensis DSM 25260	4.70318	partial	NJAU01	-
Alteribacillus bidgolensis P4B,CCM 7963,CECT 7998,DSM 25260,IBRC-M 10614,KCTC 13821 genome assembly	4.464	partial	FNDU01	-
Alteribacillus persepolensis DSM 21632	3.6191	partial	NZ_FNDK01000000	-
Chlorogloeopsis fritschii PCC 6912	7.75174	partial	RSCJ01	-
Chromohalobacter japonicus CJ	3.37628	partial	NZ_CDGZ01000000	-
Chromohalobacter japonicus SMB17	3.76792	partial	MSDQ01	-
Desulfovibrio oxyclinae DSM 11498	3.32458	partial	NZ_AQXE01000000	-
Ectothiorhodospira mobilis DSM 4180	2.62495	partial	NZ_FOUO00000000.1	-

Halarsenatibacter silvermanii SLAS-1	2.71864	partial	NZ_FNGO00000000.1	-
Halobacillus aidingensis CGMCC 1.3703	4.19184	partial	NZ_FNIZ00000000.1	-
Halobacillus alkaliphilus FP5	4.09253	partial	NZ_FOOG00000000.1	-
Halobacillus dabanensis CGMCC 1.3704	4.11984	partial	FOSB01	-
Halobacillus dabanensis HD-02	4.10233	partial	CCDH01	-
Halobacillus trueperi SS1	4.25856	partial	QTLC01	-
Halomonas arcis CGMCC 1.6494	4.14213	partial	NZ_FNII00000000.1	-
Halomonas halodenitrificans DSM 735	3.46409	partial	NZ_JHVH00000000.1	-
Halomonas meridiana ACAM 246	3.84974	partial	FSQY01	-
Halomonas saccharevitans CGMCC 1.6493	3.68129	partial	NZ_FPAQ00000000.1	-
Halomonas subterranea CGMCC 1.6495	3.7342	partial	NZ_FOGS00000000.1	-
Halonatronum saccharophilum DSM 13868	2.88452	partial	NZ_AZYG00000000.1	-
Microcoleus chthonoplastes PCC 7420	8.67904	partial	ABRS01	-
Nocardiopsis halotolerans DSM 44410	6.26393	partial	NZ_ANAX00000000.1	-
Pontibacillus halophilus JSM 076056 = DSM 19796	3.6014	partial	AULI01	-
Saccharomonospora halophila 8	3.68502	partial	AICX01	-
Salinovibrio costicola ATCC 33508 = LMG 11651	4.78167	partial	ASAI01	-
Salinovibrio costicola PRJEB21454	3.32115	partial	FYET01	-
Salisaeta longa DSM 21114	3.39902	partial	NZ_ATTH00000000.1	-
Sediminibacillus halophilus CGMCC 1.6199	4.147699	partial	NZ_FNHF00000000.1	-
Sediminibacillus halophilus NSP9.3	3.986	partial	AWXX01	-
Selenihalanaerobacter shriftii ATCC BAA-73	2.84058	partial	NZ_FUWM00000000.1	-
Spirulina subsalsa PCC 9445	5.3236	partial	NZ_ALVR00000000.1	-
Streptomyces radiopugnans CGMCC 4.3519	6.06712	partial	NZ_FOET00000000.1	-
Thalassobacillus cyri CCM7597	4.30083	partial	NZ_FNQR00000000.1	-

*TableS4.2*Analyzed complete and partial archaeal halophilic genomes

archaeal analyzed genomes	genome size (Mb)	sequencing status	genome or WGS accession number	plasmids accession numbers (for complete genomes)
Halalkalicoccus jeotgali B3	3.69865	complete	NC_014297.1	NC_014298.1, NC_014299.1, NC_014300.1 , NC_014300.1, NC_014302.1, NC_014303.1
Haloarcula hispanica ATCC 33960	3.89	complete	NC_015948.1 , NC_015943.1	NC_015944.1
Haloarcula marismortui ATCC 43049	4.27464	complete	NC_006396.1 , NC_006397.1	NC_006389.1, NC_006389.1, NC_006389.1, NC_006392.1, NC_006392.1, NC_006393.1 , NC_006394.1, NC_006395.1
Haloarcula sp CBA1115	4.22505	complete	NZ_CP010529.1	, NZ_CP010531.1, NZ_CP010532.1, NZ_CP010533.1, NZ_CP010534.1NZ_CP010530.1
Halobacterium salinarum NRC-1	2.57101	complete	NC_002607.1	NC_001869.1, NC_002608.1
Halobacterium walsbyi C23	3.36799	complete	NC_017459.1	NC_017460.1, NC_017460.1, NC_017457.1
Haloferax gibbonsii ARA6	3.91845	complete	NZ_CP011947.1	NZ_CP011948.1, NZ_CP011949.1, NZ_CP011950.1, NZ_CP011951.1
Haloferax mediterranei ATCC33500	3.90471	complete	NC_017941.2	NC_017942.1, NC_017943.1, NC_017944.1
Haloferax volcanii DS2	4.0129	complete	NC_013967.1	NC_013968.1 , NC_013965.1, NC_013964.1, NC_013966.1
Halogeometricum borinquense DSM 11551	3.94447	complete	NC_014729.1	NC_014735.1, NC_014731.1, NC_014736.1, NC_014732.1, NC_014732.1, NC_014737.1
Halomicrobium mukohataei DSM 12286	3.33235	complete	NC_013202.1	NC_013201.1
Halopiger xanaduensis SH-6(T)	4.35527	complete	CP002839.1	CP002840.1, CP002841.1, CP002842.1
Halorhabdus utahensis DSM 12940	3.116795	complete	CP001687.1	
Halorubrum lacusprofundi ATCC 49239	3.69258	complete	NC_012029.1, NC_012028.1	NC_012030.1
Haloterrigena turkmenica DSM 5511	5.44078	complete	NC_013743.1	NC_013744.1, NC_013745.1, NC_013746.1, NC_013747.1, NC_013748.1, NC_013749.1
Halovivax ruber XH-70	3.22388	complete	NC_019964.1	
Mathanohalobium evestigatum Z-7303	2.406232	complete	NC_014253.1	NC_014254.1
Methanohalophilus halophilus Z-7982	2.02296	complete	NZ_CP017921.1	

Methanohalophilus mahii DSM 5219	2.012424	complete	NC_014002.1	
Methanosalsum zhilinae DSM 4017	2.138444	complete	NC_015676.1	
Methanosarcina acetivorans C2A	5.75149	complete	AE010299.1	
Natrialba magadii ATCC 43099	4.44364	complete	NC_013922.1	NC_013923.1 , NC_013924.1, NC_013925.1
Natronobacterium gregoryi SP2	3.78836	complete	NC_019792.1	
Natronococcus occultus SP4	4.314118	complete	NC_019974.1	NC_019975.1, NC_019976.1
Natronomonas pharaonis DSM 2160	2.7497	complete	NC_007426.1	NC_007427.1, NC_007428.1
Natrialbaceae archaeon XQ-INN 246	3.972634	complete	NZ_CP050695.1	
Halorhabdus tiamatea SARL4B	3.14636	complete	NC_021921.1	NC_021913.1
Halorhabdus utahensis DSM 12940	3.1168	complete	NC_013158.1	
Halorubrum ezzemoulense Fb21	3.5686	complete	NZ_CP034940.1	NZ_CP034941.1, NZ_CP034942.1
Halostagnicola larsenii XH-48	4.13118	complete	NZ_CP007055.1	NZ_CP007056.1, NZ_CP007057.1, NZ_CP007058.1, NZ_CP007059.1
Natronorubrum bangense JCM 10635	4.24685	complete	NZ_CP031305.1	NZ_CP031306.1, NZ_CP031307.1, NZ_CP031308.1, NZ_CP031309.1
Haloterrigena daqingensis JX313	3.83336	complete	NZ_CP019327.1	NZ_CP019328.1, NZ_CP019329.1, NZ_CP019330.1
Natrinema versiforme BOL5-4	4.43264	complete	NZ_CP040330.1	NZ_CP040333.1, NZ_CP040329.1, NZ_CP040332.1, NZ_CP040331.1
Natrinema pellirubrum DSM 15624	4.30927	complete	NC_019962.1	NC_019967.1, NC_019963.1
Natrinema pallidum BOL6-1	3.84695	complete	NZ_CP040637.1	NZ_CP040638.1, NZ_CP040639.1
Natronomonas moolapensis 8.8.11	2.91257	complete	NC_020388.1	
Natronolimnobius aegyptiacus JW/NM-HA 15	3.93055	complete	NZ_CP019893.1	
Halanaeroarchaeum sulfurireducens HSR2	2.23162	complete	NZ_CP008874.1	NZ_CP008875.1
Halobiforma lacisalsi AJ5	4.36006	complete	NZ_CP019285.1	NZ_CP019286.1, NZ_CP019287.1

Halapricum salinum CBA1105	3.45178	complete	NZ_CP031310.1	
Haloterrigena jeotgali A29	4.9	complete	CP031303.1	CP031298.1, CP031299.1, CP031300.1, CP031301.1, CP031302.1, CP031304.1
Methanohalophilus portucalensis FDF-1T	2.08498	partial	NZ_CP017881.1	
Haloarcula amylolytica JCM 13557	4.22542	partial	NZ_AOLW00000000.1	
Haloarcula argentinensis DSM 12282	4.14711	partial	NZ_AOLX00000000.1	
Haloarcula japonica DSM 6131	4.28036	partial	NZ_AOLY00000000.1	
Haloarcula vallismortis ATCC 29715	3.90992	partial	NZ_AOLQ00000000.1	
Halobacterium jilantaiense CGMCC 1.5337	2.95279	partial	NZ_FOJA00000000.1	
Halobaculum gomorrense DSM 9297	3.20825	partial	NZ_FQWV00000000.1	
Halococcus morrhuae DSM 1307	2.99156	partial	NZ_AOMC00000000.1	
Halococcus saccharolyticus DSM 5350	3.4497	partial	NZ_AOMD00000000.1	
Halococcus sulifodinae DSM 8989	4.19978	partial	NZ_AOME00000000.1	
Haloferax denitrificans ATCC 35960	3.82597	partial	NZ_AOLP00000000.1	
Haloferax elongans ATCC BAA-1513	3.95214	partial	NZ_AOLK00000000.1	
Haloferax mucosum ATCC BAA-1512	3.36898	partial	NZ_AOLN00000000.1	
Haloferax sulfurifontis ATCC BAA-897	3.81243	partial	NZ_AOLM00000000.1	
Halorubrum coriense DSM 10284	3.64531	partial	NZ_AOJL00000000.1	
Halorubrum distributum JCM 10118	3.30613	partial	AOJN01	
Halorubrum distributum JCM 9100	3.30737	partial	AOJM01	
Halorubrum distributum E8	2.25364	partial	NHPH01	
Halorubrum saccharovorum DSM 1137	3.35304	partial	AOJE01	
Halorubrum sodomense RD 26	3.03055	partial	NZ_FOYN00000000.1	

Halosimplex carlsbadense 2-9-1	4.69489	partial	NZ_AOIU00000000.1	
Natronococcus amylolyticus DSM 10524	4.41653	partial	NZ_AOIB00000000.1	
Haloplanus vescus CGMCC 1.8712	2.77686	partial	NZ_FNQT00000000.1	
Natrialba asiatica DSM 12278	4.40418	partial	NZ_AOIO01000000.1	
Halorubrum aidingense JCM 13560	3.10853	partial	NZ_AOJI00000000.1	
Halorubrum arcis JCM 13916	3.3826	partial	NZ_AOJJ00000000.1	
Halorubrum californiense DSM 19288	3.68287	partial	NZ_AOJK00000000.1	
Halorubrum halophilum B8	3.46662	partial	GCA_000739595.1	
Halorubrum kocurii JCM 14978	3.61974	partial	NZ_AOJH00000000.1	
Halorubrum lipolyticum DSM 21995	3.42504	partial	GCA_000337375.1	
Halorubrum litoreum JCM 13561	3.13776	partial	NZ_AOJF00000000.1	
Halorubrum tebenquichense DSM 14210	3.32886	partial	NZ_AOJD00000000.1	
Halorubrum terrestre JCM 10247	3.37622	partial	GCA_000337435.1	
Halostagnicola kamekurae DSM 22427	4.10815	partial	NZ_FOZS00000000.1	
Halopiger salifodinae KCY07-B2	4.3509	partial	GCA_000784335.1	
Haloplanus natans DSM 17983	3.79793	partial	NZ_ATYM00000000.1	
Halopelagius inordinatus CGMCC 1.7739	3.52931	partial	NZ_FOOQ00000000.1	
Halopelagius longus BC12-B1	3.879	partial	GCA_003351065.1	
Haloterrigena hispanica CDM_6	3.96348	partial	GCA_900111485.1	
Haloterrigena limicola JCM 13563	3.52203	partial	NZ_AOIT00000000.1	
Haloterrigena saccharevitans AB14	3.98062	partial	NZ_LWLN00000000.1	
Haloterrigena salina JCM 13891	4.84161	partial	NZ_AOIS00000000.1	

Haloterrigena thermotolerans DSM 11552	3.89527	partial	NZ_AOIR00000000.1	
Halovenus aranensis IBRC-M10015	3.28712	partial	NZ_FNFC00000000.1	
Halovivax asiaticus JCM 14624	3.23845	partial	NZ_AOIQ00000000.1	
Natrialba aegyptiaca DSM 13077	4.61836	partial	NZ_AOIP00000000.1	
Natrialba hulunbeirensis JCM 10989	4.15961	partial	NZ_AOIM00000000.1	
Natrialba chahannaoensis JCM 10990	4.30927	partial	NZ_AOIN00000000.1	
Natrialba taiwanensis DSM 12281	4.63519	partial	NZ_AOIL00000000.1	
Natrinema altunense AJ2	3.774135	partial	GCA_000731985.1	
Natrinema gari JCM 14663	4.02369	partial	NZ_AOIJ00000000.1	
Natrinema salaciae DSM 25055	4.85702	partial	NZ_FOFD00000000.1	
Natronobacterium texcoconense DSM 24767	4.00987	partial	NZ_FNLC00000000.1	
Natronococcus jeotgali DSM 18795	4.49618	partial	NZ_AOIA00000000.1	
Natronolimnobius baerhuensis CGMCC 1.3597	3.90368	partial	GCA_002177135.1	
Natronolimnobius innermongolicus JCM 12255	4.58863	partial	NZ_AOHZ00000000.1	
Natronorubrum tibetense GA33	4.93084	partial	GCA_000383975.1	
Natronorubrum sediminis CGMCC 1.8981	3.78254	partial	NZ_FNWL00000000.1	
Natronorubrum sulfidifaciens JCM 14089	3.46029	partial	NZ_AOHX00000000.1	
Natronorubrum texcoconense B4,CECT 8067,JCM 17497	4.64179	partial	NZ_FNFE00000000.1	
Haloprofundus marisrubri SB9	3.92956	partial	NZ_LOPU00000000.1	
Haloterrigena mahii H13	3.79434	partial	JHUT00000000.2	
Halorubrum aethiopicum SAH-A6	3.32577	partial	NZ_LOAJ00000000.1	

Haloferax massiliensis Arc-Hr	4.01518	partial	GCA_001368915.1	
Halococcus sediminicola CBA1101	3.76437	partial	NZ_BBMP00000000.1;	
Halalkalicoccus paucihalophilus DSM 24557	3.98041	partial	NZ_LTAZ00000000.1	
Haladaptatus paucihalophilus DX253	4.28481	partial	GCA_900142335.1	
Haladaptatus litoreus CGMCC 1.7737	4.67171	partial	NZ_FTNO00000000.1	
Haladaptatus cibarius D43	3.92672	partial	NZ_JDTH00000000.1	
Halarchaeum acidiphilum JCM 16109	2.629	partial	GCA_000474235.1	
Haloarchaeobius iranensis EB21,IBRC-M 10013,KCTC 4048	3.76861	partial	NZ_FNIA00000000.1	
Haloarcula salaria ZP1-2	4.10167	partial	GCA_003992425.1	
Halorientalis persicus IBRC-M 10043	4.86976	partial	NZ_FOCX00000000.1	
Halobellus rufus CBA1103	3.85222	partial	NZ_BBJO00000000.1	
Halorientalis regularis IBRC-M 10760	4.0322	partial	NZ_FNBK00000000.1	
Halobellus clavatus CGMCC 1.10118	3.75498	partial	NZ_FNPB00000000.1	
Halorubrum amylolyticum ZC67	3.63076	partial	NZ_SDJP00000000.1	
Halobiforma haloterrestris DSM 13078	4.49544	partial	NZ_FOKW00000000.1	
Halopenitus malekzadehii IBRC-M10418	3.13682	partial	NZ_FNWU00000000.1	
Halobiforma nitratireducens JCM 10879	3.68875	partial	NZ_AOMA00000000.1	
Halococcus agarilyticus 197A	3.47673	partial	NZ_BAFM00000000.1	
Halococcus hamelinensis 100A6	3.40137	partial	GCA_000336675.1	
Halococcus thailandensis JCM 13552	4.05243	partial	NZ_AOMF00000000.1	
Haloferax larsenii CDM_5	3.79602	partial	GCA_900109695.1	
Haloferax lucentense DSM 14919	3.61906	partial	NZ_AOLH00000000.1	

Haloferax prahovense Arc-Hr	3.94622	partial	GCA_000723845.1	
Halogeometricum limi CGMCC 1.8711	3.61627	partial	NZ_FOYS00000000.1	
Halogeometricum pallidum JCM 14848	4.38452	partial	NZ_AOIV00000000.1	
Halogeometricum rufum CGMCC 1.7736	4.18712	partial	NZ_FOYT00000000.1	
Halogranum amylolyticum CGMCC 1.10121	5.18569	partial	NZ_FODV00000000.1	
Halogranum gelatinilyticum CGMCC 1.10119	3.77019	partial	NZ_FNHL00000000.1	
Halogranum rubrum CGMCC 1.7738	4.56668	partial	NZ_FOTC00000000.1	
Halogranum salarium B-1	4.49231	partial	NZ_ALJD00000000.1	
Halohasta litchfieldiae DSM 22187	3.28459	partial	NZ_FNYR00000000.1	
Halolamina pelagica CGMCC 1.10329	3.06658	partial	GCA_900115675.1	
Halolamina rubra CBA1107	2.955	partial	NZ_BBJN00000000.1	
Halolamina sediminis halo7	2.83586	partial	NZ_CVUA00000000.1	
Halomicrobium katesii DSM 19301	3.60777	partial	NZ_AQZY00000000.1	
Halomicrobium zhouii CGMCC 1.10457	4.25033	partial	NZ_FOZK00000000.1	
Caldivirga sp. SpSt-118	2.19	partial	DSBU00000000.1	
Caldivirga sp. EvPrim.Bin7	1.76	partial	WYEH00000000.1	
Caldivirga sp. CIS_19	1.45	partial	LOCC00000000.1	
Caldivirga sp. JCHS_4	1.35	partial	LOCD00000000.1	
Caldivirga sp. MG_3	1.6	partial	LOCB00000000.1	
Caldivirga sp. MU80	2.26	partial	LCTF00000000.1	
Caldivirga sp. UBA161	1.86	partial	DAXS00000000.1	

TableS4.3 ArgR binding sites in P_{intI} promoters of different integron classes (1-5)

Microorganism	**Integron class**	**Accession number**	**Presence (+) or absence (-) of ArgR binding site in P_{intI}**
Aeromonas hydrophila	1	GU295656.1	-
Achromobacter xylosoxidans	1	AY686225.1	-
Aeromonas veronii bv. *sobria*	1	FJ460183.2	-
Citrobacter freundii	1	AY162283.2	-
Corynebacterium diphtheriae	1	FR822749.1	-
Escherichia coli	1	KC417377.1	-
Enterobacter aerogenes pBWH301	1	U13880.2	-
Enterobacter cloacae	1	DQ023222.1	-
Klebsiella pneumoniae	1	DQ143913.1	-
Morganella morganii subsp. *morganii*	1	AJ621187.1	-
Pseudomonas aeruginosa	1	KM210290.1	-
Pseudomonas alcaligenes	1	GQ281702.1	-
Pseudomonas aeruginosa pVS1	1	U49101.1	-
Salmonella typhimurium IncF1 plasmid	1	AJ310778.1	-
Salmonella enterica subsp. *enterica* serovar *Agona*	1	AY289608.1	-
Serratia marcescens plasmid	1	KP177456.1	-
Shigella sp. ER.1.23	1	FJ460182.2	-
Vibrio cholerae	1	GQ214169	-
Vibrio cholerae	2	GU570570.1	+
Shigella sonnei	2	AY639870.1	-
Salmonella typhimurium plasmid incFI	2	AJ009819.1	-
Proteus mirabilis	2	JX867128.1	-
Escherichia coli	2	EU780012.1	+
Acinetobacter johnsonii	3	LN877969.1	-
Aeromonas sobria plasmid	3	KT736121.13	-
Citrobacter freundii	3	KT984195.1	-
Escherichia coli plasmid R388	3	U12441.2	-
Klebsiella oxytoca pJF-707	3	KX946994.1	-
Klebsiella pneumoniae	3	AY219651.1	-
Serratia marcescens	3	AF416297	-
Vibrio cholera strain JX20062026	4	KF680548.1	+
Vibrio cholerae	4 (intIA)	AF055586.1	+

Vibrio cholerae mega-integron MInVc	4	AF179592.1	+
Vibrio metschnikovii	4	ACZO01000007.1	+
Vibrio mimicus mega-integron MInVm	5	AF179595	-
Vibrio cholerae O139 plasmid pVC1699	-	MT127634.1	-

TableS4.4 Genetic elements of identified complete integrons and CALINs within studied genomes of halophilic microorganisms and thermophilic archaea. Only CALINs with toxin-antitoxin (TA) systems, insertion sequences (IS) or known antibiotic resistance genes (ARG) are shown.

site	Accession no.	Genetic element	Annotation (description)	position
Salinibacter ruber DSM 13855	NC_007677.1	*int*I gene-A	Integron integrase, 261 amino-acid-residues, 14.56% acidic residues.	c(77736..78521)
		Promoter	Putative P_C, LDF 0.97 Binding site for transcription factor: purR	78461..78488
		*att*I	Putative primary recombination site CCTGATCGAGGAGTTTGGG	78733..78751
		Promoter	Putative P*int*I, LDF 3.57 Binding sites for transcription factors: torR and nagC	c(78835..78866)
		Gene cassette ORF	Hypothetical protein	79006..79932
		Promoter	Promoter for the TA operon, LDF 4.96 Binding sites for transcription factors: nagC, nagC, rpoD16, crp and arcA	79911..79943
		*att*C	Putative Cassette-associated recombination site-predicted manually and by MFOLD	79954..80067
		Gene cassette ORF	VapC toxin of TA system	80198...80590
		Promoter	Promoter for the antitoxin within the toxin gene, LDF 0.26 No binding sites for transcriptional factors	80439..80467
		Gene cassette ORF	Phd/YefM antitoxin of TA system	80593..80820
		*att*C	Putative Cassette-associated recombination site-predicted manually and by MFOLD	80847..80995

		*int*I gene-B	Integron integrase, pseudogene: missing patch II and K174 (active site residue [39]), 225 amino-acid-residues, 14.22% acidic residues.	1134981..1135658 [1134525..1135677 is a duplication of c(77718..78978) with a deletion within the 2nd *int*I]
		Promoter	Putative P_{intI}, LDF 2.31 Binding sites for transcription factors: torR and rpoD17	1134636..1134667
		*att*I	Putative primary recombination site CCTGATCGAGGAGTTTGGG	c(1134751..134769)
		Promoter	Putative P_C, LDF 0.97 Binding site for transcription factor: purR	c(1135014..1135041) ,
Marinobacter hydrocarbonoclasticus VT8	NC_008740.1	*int*I gene-A	*Int*egron *int*egrase, 329 amino-acid-residues, 10.33% acidic residues.	c(1114421..1115410)
		Promoter	Putative P_{intI}, LDF 6.27 Binding sites for transcription factors: lexA, crp and argR	c(1115437..1115469)
		*att*I	Putative primary recombination site TGCTAACCTTCTGATAAGT	1115536..1115554
		Promoter	Putative P_C, LDF 4.66 Binding sites for transcription factors: purR, rpoD16, rpoD17	1115728..1115753
		Promoter	Putative P_{intI}, LDF 1.76 Binding sites for transcription factors: crp, arcA	c(1115764..1115792)
		IS*91*	IS*Mahy2*	1115766..1118105
		Gene cassette ORF	Tyr recombinase	1115873..1116724
		Gene cassette ORF	IS*91* transposase	1116724..1117848
		Gene cassette ORF	hypothetical protein	1118174..1118674
		*att*C	Cassette-associated recombination site	1118675..1118761
		Gene cassette ORF	hypothetical protein	1119098..1119202
		IS*1182*	IS*Mahy3*	1119398..1120982
		*att*C	Cassette-associated recombination site- within abovementioned ORF	1119115..1119204

		Gene cassette ORF	IS*1182* transposase	1119465..1120790
		IS*1380*	IS*Maq3* isoform	1120983..1122550
		Gene cassette ORF	IS*1380* transposase	1121083..1122384
		*att*C	Cassette-associated recombination site	1123078..1123178
		Gene cassette ORF	hypothetical protein	1123193..1123495
		*att*C	Cassette-associated recombination site	1123490..1123576
		IS*1182*	IS*Mahy5*	1123573..1125402
		Gene cassette ORF	IS*1182* transposase	1123644..1125215
		Gene cassette ORF	Hypothetical protein	1125215..1125418
		Gene cassette ORF	Hypothetical protein	1125428..1125742
		*att*C	Cassette-associated recombination site	1125745..1125848
		Gene cassette ORF	Hypothetical protein	1125862..1126434
		*att*C	Cassette-associated recombination site	1126453..1126529
		Promoter	Promoter for the TA system, LDF 1.11 No binding sites for transcription factors	1126544..1126573
		Gene cassette ORF	RelB antitoxin of TA system	1126604..1126828
		Promoter	Promoter for the toxin gene within antitoxin ORF, LDF 2.76 Binding sites for transcription factors: rpoS18, rpoD16 and metR	1126794..1126826
		Gene cassette ORF	ParE toxin of TA system	1126825..1126938
		IS*21*	IS*Spu5* isoform	c(1126938_1129503)

		Gene cassette ORF	IstB ATP binding domain protein	c(1127028..1127783)
		Gene cassette ORF	IS*21* transposase	c(1127795..1129315)
		*att*C	Cassette-associated recombination site	1129715..1129792
		*att*C	Cassette-associated recombination site	1130126..1130226
		*att*C	Cassette-associated recombination site	1130529..1130617
		*att*C	Cassette-associated recombination site	1130921..1131010
		Gene cassette ORF	hypothetical protein	c(1131347..1131460)
		Gene cassette ORF	VOC family protein	1131524..1131883
		*att*C	Cassette-associated recombination site	1132264..1132363
		IS*110*	IS*Mahy7*	c(1134238..1135620)
		ORF	IS*110* transposase	c(1134502..1135527)
		IS*21*	IS*Mahy8*	1138676..1141315
		ORF	IS*21* transposase	1138908..1140452
		ORF	IS*21*-like element helper ATPase IstB	1140467..1141222
		IS*256*	IS*Mahy14*	c(3544903..3546285)
		ORF	IS*256* transposase	c(3544937..3546164)
		*att*C	Cassette-associated recombination site	c(3546472..3546573)
		Gene cassette ORF	Immunity protein (Imm70 Superfamily)	c(3546568..3546573)
		*att*C	Cassette-associated recombination site	c(3546998..3547102)
		Gene cassette ORF	Hypothetical protein	c(3547114..3547398)
		Promoter	Putative P*int*I, LDF 5 Binding sites for transcription factors: argR and argR2	3547685..3547711
		Promoter	Putative P_C, LDF 5.68	c(3547707..3547736)

			Binding sites for transcription factors: fis and arcA	
		*att*I	Putative primary recombination site TATGTACGTACAGTTATAA	c(3547708..3547726)
		*int*I gene-B	Integron integrase, 322 amino-acid-residues, 8.7% acidic residues	3547752..3548720
Marinobacter hydrocarbonoclastic us ATCC 49840	NC_017067.1	*att*C	Cassette-associated recombination site	1157150..1157239
		Gene cassette ORF	ABC transporter (permease)	c(1157520..1157693)
		*att*C	Cassette-associated recombination site	1157897..1157968
		Gene cassette ORF	Hypothetical protein	1157984..1158337
		*att*C	Cassette-associated recombination site	1158346..1158429
		Gene cassette ORF	Hypothetical protein	1158453..1158908
		*att*C	Cassette-associated recombination site	1158909..1158956
		Gene cassette ORF	Hypothetical protein	1158976..1159272
		*att*C	Cassette-associated recombination site	1159283..1159367
		Gene cassette ORF	Hypothetical protein	1159385..1159717
		Gene cassette ORF	Hypothetical protein	1159792..1160250
		Gene cassette ORF	Hypothetical protein	1160284..1160634
		IS*3*	IS*Maq2* isoform	1160989..1162307
		Gene cassette ORF	IS*3* transposase combined orfAB	1161069.. 1162276-frameshift
		*att*C	Cassette-associated recombination site	1162450..1162525

		Gene cassette ORF	Hypothetical protein	1163294..1163590
		Gene cassette ORF	Hypothetical protein (Ypar14 Super-integron cassette)	1163708..1164067
		*att*C	Cassette-associated recombination site	1164069..1164151
		Gene cassette ORF	Hypothetical protein	1164148..1164399
		*att*C	Cassette-associated recombination site	1164401..1164498
		Gene cassette ORF	Hypothetical protein	1164678..1165019
		*att*C	Cassette-associated recombination site	1164986..1165045
		Gene cassette ORF	Hypothetical protein	1165498..1165800
		*att*C	Cassette-associated recombination site	1165713..1165802
		Gene cassette ORF	Hypothetical protein	1165809..1166132
		Gene cassette ORF	Hypothetical protein	1166218..1166631
		*att*C	Cassette-associated recombination site	1166843..1166913
		Gene cassette ORF	DUF4279 domain containing protein	1166947..1167354
		*att*C	Cassette-associated recombination site	1167349..1167426
		ORF	Nuclear transport actor 2 family protein	1167448..1167822
		IS*3*	IS*3* with a frameshift in ORFB	1168407..1169677
		ORF	IS*3* transposase combined OrfAB	1168464.. 1169645-frameshift
		*att*C	Cassette-associated recombination site	c(1969904..1969998)

		Gene cassette ORF	phosphopantetheine adenyltransferase (2.7.7.3)	c(1969993..1970373)
		*att*C	Cassette-associated recombination site	c(1970393..1970450)
		Gene cassette ORF	GNAT family N-acetyltransferase (2.3.-)	c(1970445..1970873)
		*att*C	Cassette-associated recombination site	c(1970889..1970966)
		Gene cassette ORF	VOC family protein	c(1970989..1971396)
		*att*C	Cassette-associated recombination site	c(1971421..1971510)
		*att*C	Cassette-associated recombination site	c(1971845..1971922)
		Gene cassette ORF	Antibiotic biosynthesis monooxygenase	c(1972336..1972671)
		*att*C	Cassette-associated recombination site	c(1972685)..1972785)
		Gene cassette ORF	hypothetical protein	c(1972780..1973316)
		Gene cassette ORF	DUF4145 domain containing protein	c(1973409..1974110)
		*att*I	Putative primary recombination site CGCTAATAAGCTGTTAGGA	c(1974117..1974135)
		Promoter	Putative P_{intI}, LDF 2.97 Binding sites for transcription factors: fis, lrp and metR	1974169..1974197
		Promoter	Putative P_C, LDF 4.4 Binding sites for transcription factors: metR, rpoD17	c(1974419..1974451)
		*Int*I gene	Integron integrase, complete, pseudogene: with frameshift and interrupted by IS*Maq2*, 257 amino-acid-residues, 9.3% acidic residues.	1974463..1976754- Frameshift at 1976152
		IS*3*	IS*Maq2* isoform	1974661..1975979
		ORF	IS*3* transposase combined ORFAB	1974741..1975948 - frameshift

		*att*C	Cassette-associated recombination site	c(2335846..2335932)
		Gene cassette ORF	Hypothetical protein	c(2335927..2336214)
		*att*C	Cassette-associated recombination site	c(2336234..2336320)
		Gene cassette ORF	Hypothetical protein	c(2336325..2337125)
		*att*C	Cassette-associated recombination site	c(2337142..2337226)
		Gene cassette ORF	Hypothetical protein	c(2337221..2337505)
		Gene cassette ORF	Hypothetical protein	c(2337625..2337900)
		Gene cassette ORF	Methyldantoinase immunity protein (Imm32 domain)	c(2337999..2338277)
		Gene cassette ORF	Hypothetical protein	c(2338284..2338424)
		*att*C	Cassette-associated recombination site	c(2338291..2338380)
		Gene cassette ORF	GNAT family N-acetyltransferase	c(2338375..2338761)
		Gene cassette ORF	Hypothetical protein	c(2338947..2339330)
		*att*C	Cassette-associated recombination site	c(2339345..2339429)
		Gene cassette ORF	Hypothetical protein	c(2339431..2339808)
		Gene cassette ORF	Hypothetical protein	c(2339894..2340580)
		Gene cassette ORF	AbiV family abortive infection protein (phage resistance)	c(2340678..2341331)
		Gene cassette ORF	Hypothetical protein	c(2341427..2342380)
		Gene cassette ORF	Hypothetical protein	c(2342471..2342833)
		*att*C	Cassette-associated recombination site	c(2342852..2342923)

		Gene cassette ORF	Hypothetical protein	c(2342918..2343388)
		Gene cassette ORF	Hypothetical protein	c(2343488..2344015)
		*att*C	Cassette-associated recombination site	c(2344338..2344395)
		Gene cassette ORF	Hypothetical protein	c(2344390..2344908)
		*att*C	Cassette-associated recombination site	c(2344924..2345007)
		Gene cassette ORF	Hypothetical protein	c(2345002..2345439)
		*att*C	Cassette-associated recombination site	c(2345458..2345515)
		Gene cassette ORF	Hypothetical protein	c(2345524..2345751)
		*att*C	Cassette-associated recombination site	c(3913681..3913752)
		Gene cassette ORF	Hypothetical protein YexB family (uncharacterized transmembrane protein)	c(3913747..3914286)
		IS*110*	IS*Mahy7* isoform	3914323..3915689
		Gene cassette ORF	IS*110* transposase	3914404..3915429
		*att*C	Cassette-associated recombination site	c(3916067..3916156)
		Gene cassette ORF	SEC-C domain containing protein	c(3916151..3917269)
		*att*C	Cassette-associated recombination site	c(3917287..3917358)
		Gene cassette ORF	Hypothetical protein (FRG domain)	c(3917353..3918066)
		IS*110*	IS*Mahy12*	3918806..3920387
		ORF	IS*110* transposase	3918885..3919928
		IS*91*	IS*Mahy2*	c(3921905..3924570)
		ORF	IS*91* transposase	c(3922342..3923466)
		ORF	Site specific integrase	c(3923466..3924317)
Nitrosococcus halophilus Nc4	NC_013960.1	ORF	Txe/YoeB family toxin of TA system	c(131629..131799)

		ORF	Phd/YefM family antitoxin of TA system	c(131883..132125)
		Promoter	Promoter for the TA operon, LDF 2.17 Binding site for the transcription factor: rpoS18	c(132155..132182)
		*att*C	Cassette-associated recombination site	c(132199..132262)
		Gene cassette ORF	BrnA antitoxin of TA system	c(132229..132534)
		Gene cassette ORF	BrnT family toxin of TA system	c(132531..132809)
		Promoter	Promoter for the TA operon, LDF 4.44 Binding site for the transcription factor: metJ	c(132817..132846)
		Gene cassette ORF	hypothetical protein	c(132956..133438)
		Gene cassette ORF	DUF2442 domain containing protein	c(133537..133797)
		Gene cassette ORF	DUF4160 domain containing protein	c(133757..133996)
		Gene cassette ORF	DUF2442 domain containing protein	c(134132..134398)
		Gene cassette ORF	DUF4160 domain containing protein	c(134406..134603)
		Gene cassette ORF	PEP-CTERM sorting domain containing protein	c(135231..136055)
		*att*I	Putative primary recombination site AGTCT*ATT*TCATGTTAAGC	c(136345..136363)
		Promoter	Putative P_C, LDF 2.87 Binding site for transcription factor: lexA	c(136560..136593)
		Promoter	Putative P_{intI}, LDF 5.07 Binding sites for transcription factors: lrp, rpoH2, lexA and argR	136713..136739
		*Int*I gene-A	Integron integrase, 275 amino-acid-residues, 9.1% acidic residues	136761..137585

		*att*C	Cassette-associated recombination site	c(1267034..1267109)
		Promoter	Promoter for TA operon, LDF 4.75 Binding sites for transcription factors: rpoD17, lrp,metR, soxS, rpoD17, rpoD16 and argR2	1267021..1267042
		Gene cassette ORF	Phd/YefM family antitoxin of TA system	1267160..1267432
		Gene cassette ORF	Txe/YoeB family toxin	1267413..1267682
		*att*C	Cassette-associated recombination site	c(1267686..1267757)
		Gene cassette ORF	Transposase	c(1268094..1268792)
		Gene cassette ORF	DUF3047 domain containing protein	126425..1270192
		Gene cassette ORF	RelE toxin (no domain detected by blastx)	c(1270954..1271562)
		Gene cassette ORF	Hypothetical protein	c(1270954..1271562)
		*att*C	Cassette-associated recombination site	c(1271605..1271692)
		Gene cassette ORF	hypothetical protein	c(1271718..1271945)
		Gene cassette ORF	zinc ribbon domain containing protein	c(1271949..1272215)
		*att*I	Putative primary recombination site TATCTACTCAATGTTAGGC	c(1272242..1272260)
		Promoter	Putative P_{intI}, LDF 12.44 Binding sites for transcription factors: fis, rpoD17, lexA, rpoD16, nagC, rpoD18, deoR, rpoD17, rpoD17, crp, crp, arcA and arcA	1272312..1272344
		Promoter	Putative P_C, LDF 11.24 Binding sites for transcription factors: phoB, arcA, crp, dnamino-acid-residues, rpoD16, cpxR, cpxR, fis, lrp, argR2, rpoD16 and farR	c(1272335..1272358)

		*Int*I gene-B	Integron integrase, 321 amino-acid-residues, 10% acidic residues.	1272427..1273392
		IS*4*	IS*Nhal1*	c(1273422_1274699)
		ORF	IS*4* transposase	c(1273443_1274639)
		*att*C	Cassette-associated recombination site	2373457..2373543
		Gene cassette ORF	Hypothetical protein	2373692..2375932
		*att*C	Cassette-associated recombination site	2375911..237971
		Gene cassette ORF	Hypothetical protein	2375989..2376306
		*att*C	Cassette-associated recombination site	2376301..2376377
		Gene cassette ORF	HAD hydrolase-like protein	2376396..2377013
		Gene cassette ORF	Amidophosphoribosyltransferase	2377010..2377633
		Gene cassette ORF	DNA binding protein	2377633..2378607
		Gene cassette ORF	Hypothetical protein	2378722..2378922
		Gene cassette ORF	Hypothetical protein	2378919..2379173
		*att*C	Cassette-associated recombination site	2379180..2379312
		Gene cassette ORF	Hypothetical protein	2379327..2379653
		*att*C	Cassette-associated recombination site	2379648..2379707
		Gene cassette ORF	Hypothetical protein	2379723..2380229
		*att*C	Cassette-associated recombination site	2380233..2380327
		Gene cassette ORF	Class I SAM dependent methyltransferase	2380369..2381115
		*att*C	Cassette-associated recombination site	2381253..2381311

Salinovibrio costicola ATCC 33508	ASAI01000027.1	*Int*I	Integron integrase, complete, 322 amino-acid-residues, 9.01% acidic residues	c(37320..38288)
		promoter	Putative P_C, LDF 5.68 Binding sites for transcription factors: fis and arcA	38304..38333
		Promoter	Putative P_{intI}, LDF 5 Binding sites for transcription factors: argR and argR2	c(38329..38355)
		*att*I	Putative primary recombination site AACTAATAAGCTGTTATAT	38618..38636
		Gene cassette ORF	hypothetical protein	38652..39044
		*att*C	Cassette-associated recombination site	39045..39126
		Gene cassette ORF	methyl accepting chemotaxis protein	c(39508..41595)
		Gene cassette ORF	EAL domain containing protein	c(42030..44711)
		*att*C	Cassette-associated recombination site (within abovementioned ORF)	44024..44088
Salinovibrio costicola ATCC 33508	ASAI01000046.1	IS*1634*	IS*Saco1* isoform-with a frameshift within transposase sequence-probably non-functional	27110..28875
		ORF	IS*1634* transposase	27242.. 28856-frameshift at 28537
		*att*C	Cassette-associated recombination site	c(28973..29098)
		Gene cassette ORF	Hypothetical protein	c(29101..29652)
		*att*C	Cassette-associated recombination site	c(29676..29747)
		Gene cassette ORF	DUF3800 domain containing protein	c(29751..30866)
		*att*C	Cassette-associated recombination site	c(30895..30972)
		Gene cassette ORF	Hypothetical protein	c(30974..31267)

Salinovibrio costicola PRJEB21454	LT897828.1	*int*I gene	Integron integrase, 319 amino-acid-residues, 6.9% acidic residues.	c(375211..376170)
		Promoter	Putative P_C, LDF 5.26 Binding sites for transcription factors: phoB and arcA	376383..376411
		Promoter	Putative P_{intI}, LDF 4.26 Binding sites for transcription factors: arcA and arcA	c(376410..376443)
		*att*I	Putative primary recombination site CACTAATACAATGTTAGCC	376466..376486
		Gene cassette ORF	hypothetical protein	376653..377180
		*att*C	Cassette-associated recombination site	377175..377300
		Gene cassette ORF	hypothetical protein	377325..377795
		*att*C	Cassette-associated recombination site	377869..377929
		Gene cassette ORF	hypothetical protein	377946..378614
		*att*C	Cassette-associated recombination site	378609..378711
		IS*91*	IS*Saco2*	c(379277..381473)
		ORF	IS*91* transposase	c(379467..380513)
		ORF	site specific integrase	c(380510..381391)
Salinovibrio costicola PRJEB21454	LT897834.1	*att*C	Cassette-associated recombination site	c(114946..115015)
		Gene cassette ORF	hypothetical protein	c(115004..115510)
		Gene cassette ORF	IS*110* transposase-partial	115610..115792
		Gene cassette ORF	RelB/DinJ antitoxin of TA system-very short sequence, most probably partial, no domains detected by blastx	115793..115948
		Gene cassette ORF	RelE/ParE toxin of TA system-partial missing C terminus-incomplete ParE domain	115945..116136

		Gene cassette ORF	Hypothetical protein	c(116440..116736)
		Gene cassette ORF	HigA antitoxin of TA system	c(116833..117123)
		Gene cassette ORF	ParE toxin of TA system	c(117133..117411)
		Promoter	Promoter for the antitoxin gene within the toxin ORF, LDF 1.02 Binding site for the transcription factor: rpoD17	c(117302..117335)
		Promoter	Promoter for the TA operon, LDF 3.5 Binding sites for transcription factors: dnaA and lexA	c(117418..117446)
		*att*C	Cassette-associated recombination site	c(117450..117527)
		Gene cassette ORF	GrpB family protein probable nucleotidyltransferase	c(117522..117929)
		ORF	IS*As1* transposase, partial (end of contig)	c(117952..118755)
Salinovibrio costicola subsp. alcaliphilus strain DSM 19052	MUFR01000057.1	*att*C	Cassette-associated recombination site	1342..1425
		Gene cassette ORF	Hypothetical protein	1443..1943
		*att*C	Cassette-associated recombination site	1938..2021
		Gene cassette ORF	Hypothetical protein	2036..3061
		Gene cassette ORF	Hypothetical protein	3063..3674
		*att*C	Cassette-associated recombination site	3677..3751
		Gene cassette ORF	Glutathione dependent-formaldehyde activating enzyme	3785..4186
		*att*C	Cassette-associated recombination site	4196..4265
		Gene cassette ORF	Peptidase S51/Glutamine amidotransferase	4300..4923

		*att*C	Cassette-associated recombination site	4918..4989
		Gene cassette ORF	Hypothetical protein	5009..5335
		*att*C	Cassette-associated recombination site	5404..5512
		Gene cassette ORF	Hypothetical protein	5526..6092
		*att*C	Cassette-associated recombination site	6100..6171
		Gene cassette ORF	YafQ toxin (ParE family) of TA system with *int*ernal stop codon	c(6172..6445)
		Promoter	Promoter for the toxin gene within the antitoxin gene, LDF 3.47 Binding sites for transcription factors: argR and fnr	c(6435..6467)
		Gene cassette ORF	RelB/DinJ family antitoxin of TA system	c(6438..6716)
		Promoter	Promoter for TA operon, LDF 6.19 Binding sites for transcription factors: ihf and glpR	c(6759..6787)
		*att*C	Cassette-associated recombination site	6773..6856
		Gene cassette ORF	Hypothetical protein (SIR2 superfamily)	6866..8287
		Gene cassette ORF	ATPase	8284..10347
		*att*C	Cassette-associated recombination site	10342..10413
		Gene cassette ORF	Hypothetical protein	10483..10887
		*att*C	Cassette-associated recombination site	10882..10954
		Gene cassette ORF	O-methyltransferase	10971..11546
		*att*C	Cassette-associated recombination site	11541..11653
		Gene cassette ORF	Hypothetical protein	11670..12353
		*att*C	Cassette-associated recombination site	12305..12363

Chromohalobacter japonicus CJ	LN651368.1	IS*30*	IS*Chja3*	500932..502018
		ORF	IS*30* transposase	500980..501996
		*att*C	Cassette-associated recombination site	502166..502249
		*att*C	Cassette-associated recombination site	502677..502751
		Gene cassette ORF	SHOCT domain containing protein	502778..503518
		IS*30*	IS*Chja2*	503861..504958
		Gene cassette ORF	IS*30* transposase	503913..50٤٩٣٢
		*att*C	Cassette-associated recombination site	505766..505879
		IS*30*	IS*Chja4*	505899..507008
		ORF	IS*30* transposase	506005..506997
		IS*1380*	IS*Chja1*	c(618728..620445)
		ORF	IS*1380* transposase	c(618961..620340)
		IS*30*	IS*CHja2*	620874..621971
		ORF	IS*30* transposase	620926..621945
		*att*C	Cassette-associated recombination site	c(624468..624541)
		Gene cassette ORF	Hypothetical protein	c(624996..625268)
		Gene cassette ORF	DUF4062 domain containing protein	c(625394..626233)
		*att*C	Cassette-associated recombination site	c(626253..626330)
		*att*C	Cassette-associated recombination site	c(626737..626821)
		Gene cassette ORF	RDD family protein	c(626816..627232)
		*att*C	Cassette-associated recombination site	c(627251..627342)
		*att*C	Cassette-associated recombination site	c(627675..627773)
		Gene cassette ORF	PH domain containing protein	c(627776..628186)

		*att*I	Putative primary recombination site TGAAATCAATGAGTTAGGT	c(628320..628338)
		Promoter	Putative P_{intI}, LDF 5.26 Binding sites for transcription factors: lexA, argR, argR2 and crp	628390..628416
		Promoter	Putative P_C, LDF 4.83 Binding sites for transcription factors: oxyR, rpoD16, lexA	c(628402..628431)
		*int*I gene	Integron integrase, 321 amino-acid-residues, 6.54% acidic residues	628438..629403
Chromohalobacter japonicus SMB17	MSDQ01000006	*att*C	Cassette-associated recombination site	c(452811..452915)
		Gene cassette ORF	Hypothetical protein	c(452903..453341)
		*att*C	Cassette-associated recombination site	c(453258..453335)
		*att*C	Cassette-associated recombination site	c(453820..453891)
		Gene cassette ORF	Hypothetical protein	c(453886..454782)
		*att*C	Cassette-associated recombination site	c(454814..454881)
		Gene cassette ORF	Hypothetical protein	c(454883..455605)
		Gene cassette ORF	Hypothetical protein	c(455708..456028)
		Gene cassette ORF	FRG domain containing protein	c(456116..456898)
		Gene cassette ORF	Hypothetical protein	c(456986..457711)
		*att*C	Cassette-associated recombination site	c(457722..457800)
		Gene cassette ORF	HEAT repeat domain containing protein	c(457801..458238)
		*att*C	Cassette-associated recombination site	c(458272..458374)
		Gene cassette ORF	nitronate monooxygenase	c(458376..459329)

		Promoter	Putative P_C, LDF 4.72 (typical to the one in CJ strain) Binding sites for transcription factors: oxyR, rpoD16 and lexA	c(459411..459540)
		*att*I	Putative primary recombination site CGAAATCAATGGGTTAGGT	c(459426..459444)
		Promoter	Putative P_{intI}, LDF 5.44 Binding sites for transcription factors: lexA, argR, argR2 and crp	459481..459509
		*int*I gene	Integron integrase, 346 amino-acid-residues, 6.65% acidic residues.	459547..460584
Halomonas titanicae ANRCS81	NZ_CP039374.1	IS*3*	IS*Hati1*	4076844..4078113
		ORF	IS*3* transposase combined ORFAB	4076901..4078081-frameshift
		Gene cassette ORF	CPBP family *int*ramembrane metalloprotease	4080038..4080239
		*att*C	Cassette-associated recombination site	4080233..4080314
		Gene cassette ORF	Hypothetical protein	4080339..4080878
		*att*C	Cassette-associated recombination site	4080873..4080949
		Gene cassette ORF	Restriction endonuclease	4081053..4081967
		*att*C	Cassette-associated recombination site	4081962..4082033
		Promoter	Promoter for the TA operon, LDF 3.63 Binding site for transcription factor: hns	4082051..4082079
		ORF	ParE toxin of TA system	4082087..4082
		ORF	HigA antitoxin of TA system	4082383..4082698
		IS*1182*	IS*1182* with internal deletion in transposase	4083242..4084259
		ORF	IS*1182* transposase	4083312..4084166
		*att*C	Cassette-associated recombination site	c(4735139..4735186)

		Gene cassette ORF	Hypothetical protein	c(4735187..4735738)
		Gene cassette ORF	Hypothetical protein	c(4735853..4736383)
		Gene cassette ORF	Hypothetical protein	c(4736485..4736838)
		IS*91*	IS*Hati3*	c(4736861..4739273)
		Gene cassette ORF	IS*91* transposase	c(4737165..4738316)
		Gene cassette ORF	IS*91* integrase/resolvase	c(4738309..4739166)
		Gene cassette ORF	transposase	c(4739666..4741711)
		*att*C	Cassette-associated recombination site	c(4741937..4741998)
		Gene cassette ORF	Hypothetical protein	c(4742008..4742343)
Halomonas halodenitrificans DSM 735	NZ_JHVH01000020.1	*int*I gene	Integron integrase, 321 amino-acid-residues, 7.79% acidic residues.	c(71074..72039)
		Promoter	Putative P_{intI}, LDF 3.03 Binding sites for transcription factors: rpoD17 and rpoD19	c(72155..72183)
		Promoter	Putative P_C, LDF 3.79 Binding sites for transcription factors: lexA and pdhR	72158..72186
		*att*I	Putative primary recombination site GTCTAASTACCTGTTAGAT	72230..72247
		Gene cassette ORF	DUF1837 domain containing protein	72280..73194
		Gene cassette ORF	DEAD/DEAH box helicase	73197..75275
		*att*C	Cassette-associated recombination site	75305..75385
		Gene cassette ORF	TIR domain containing protein	75436..77934

		*att*C	Cassette-associated recombination site	77929..78000
		ORF	IS*1380* transposase, partial (end of contig)	78458..78653
Halomonas elongata DSM 2581	NC_014532.2	*att*C	Cassette-associated recombination site	c(2275491..2275563)
		Gene cassette ORF	Hypothetical protein	c(2276034..2276924)
		Gene cassette ORF	DUF4062 domain containing protein	c(2277007..2278185)
		*att*C	Cassette-associated recombination site	c(2278204..2278259)
		Gene cassette ORF	GNAT-family N-acetyltransferase	c(2278214..2278720)
		Gene cassette ORF	Spectinomycin adenyltransferase	c(2278807..2279586)
		Gene cassette ORF	Hypothetical protein	c(2279673..2280305)
		Gene cassette ORF	DUF3800-domain containing protein	c(2280328..2281521)
		*att*C	Cassette-associated recombination site	c(2281541..2281630)
		Gene cassette ORF	Polygammaglutamate hydrolase-family protein	c(2281625..2282239)
		*att*C	Cassette-associated recombination site	c(2692417..2692507)
		Gene cassette ORF	Hypothetical protein	c(2692502..2693395)
		*att*C	Cassette-associated recombination site	c(2693415..2693474)
		*att*C	Cassette-associated recombination site	c(2693856..2693927)
		Gene cassette ORF	Hypothetical protein	c(2693938..2694930)
		*att*C	Cassette-associated recombination site	c(2694955..2694930)
		Gene cassette ORF	Hypothetical protein	c(2695016..2696530)
		*att*C	Cassette-associated recombination site	c(2696545..2696604)

		Gene cassette ORF	AAA-family ATPase	c(2696616..2698142)
Halomonas arcis CGMCC 1.6494	NZ_FNII01000009.1	Gene cassette ORF	AAA-family ATPase	79248..81143
		*att*C	Cassette-associated recombination site	81168..81266
		Gene cassette ORF	Transposase	81334..82904-frameshift at 82248
		Gene cassette ORF	Hypothetical protein	83129..83392
		*att*C	Cassette-associated recombination site	83387..83487
		Gene cassette ORF	Urea carboxylase-associated family protein	83519..84139
		*att*C	Cassette-associated recombination site	84186..84258
		Gene cassette ORF	DUF3703-domain containing protein	84277..84645
		*att*C	Cassette-associated recombination site	84640..84717
		Gene cassette ORF	Hypothetical protein	84733..85206
		*att*C	Cassette-associated recombination site	85201..85260
		Gene cassette ORF	Hypothetical protein	85273..85833
		*att*C	Cassette-associated recombination site	85828..85905
		Gene cassette ORF	Hypothetical protein	866645..86673
		Gene cassette ORF	Hypothetical protein	86645.. 86803
		Gene cassette ORF	Hypothetical protein	86772..87161
		*att*C	Cassette-associated recombination site	87156..87240
		Gene cassette ORF	Hypothetical protein	c(87222..87374)

		*att*C	Cassette-associated recombination site	87662..87748
		Promoter	Promoter for the TA operon, LDF 4.9 Binding sites for transcription factors: rpoD17 and narL	87753..87781
		ORF	ParE toxin of TA system	87881..88066
		ORF	HigA antitoxin of TA system	88072..88359
Halomonas meridiana strain ACAM 246	FSQY01000001.1	*int*I	Integron integrase, 314 amino-acid-residues, 6.69% acidic residues	c(2813775..2814719)
		Promoter	Putative $P_{intI,}$ LDF 1.16 No binding sites for transcription factors	c(2814791..2814823)
		Promoter	Putative P_C, LDF 3.35 Binding site for transcription factor: lrp	2814818..2814847
		*att*I	Putative primary recombination site GTAGAASTCAATGAGGTAGAT	2814827..2814846
		Gene cassette ORF	VOC family protein	2814903..2815349
		Promoter	Promoter for the TA operon, LDF 5.71 Binding sites for the transcription factors: lrp, argR and tyrR	2815492..2815519
		Gene cassette ORF	BrnA family toxin of TA system	2815526..2815819
		Gene cassette ORF	BrnA antitoxin of TA system	2815812..2816021
		Promoter	Promoter for the TA operon, LDF 4.29 No binding sites for transcription factors	2816083..2816111
		Gene cassette ORF	Putative antitoxin for TA system (ribbon-helix-helix protein CopG family)	2816177..2816449
		Gene cassette ORF	ParE family toxin of TA system	2816446..2816727
		Gene cassette ORF	Antibiotic biosynthesis monooxygenase	2816861..2817163
		*att*C	Cassette-associated recombination site	2817167..2817582

		Gene cassette ORF	hypothetical protein	2817295..2817582
		*att*C	Cassette-associated recombination site	2817585..2817673
		ORF	RelE/ParE toxin of TA system	c(2817668..2817982)
		ORF	Antitoxin of TA system	c(2817985..2818221)
		Promoter	Promoter for TA operon, LDF 2.18 No binding sites for transcription factors	c(2818246..2818277)
		ORF	IS66 transposase (it was difficult to determine the peripheries of the IS element)	c(2818271..2819560)
		IS*256*	IS*Hame1*	3325612..3326994
		ORF	IS*256* transposase	332511..3326964-
		Gene cassette ORF	VapC toxin (PIN domain)	c(3327182..3327577)
		Gene cassette ORF	VapB antitoxin	c(3327580..3327774)-
		Promoter	Promoter for TA operon, LDF 4.57 Binding sites for transcription factors: rpoS17 and soxS	3327812..3327840
		*att*C	Cassette-associated recombination site	3327832..3327909
		IS*5*	IS*Hame2*	3327908..3329133
		Gene cassette ORF	IS*5* transposase	3327999..3328997
		*att*C	Cassette-associated recombination site	3329397..3329497
		ORF	ParE toxin of TA system	c(3329492..3329785)
		ORF	ParD antitoxin of TA system	c(3329785..3330024)
		Promoter	Promoter for TA operon, LDF 3.45 Binding site for transcription factor: narL	3330053..3330082
Halomonas saccharevitans CGMCC 1.6493	NZ_FPAQ01000028.1	*int*I	Integron integrase, 321 amino-acid-residues, 7.17% acidic residues	c(43903..44868)

		Promoter	Putative P_{intI}, LDF 2.74 Binding sites for transcription actors: lexA, argR, argR2 and nagC	c(44903..44937)
		*att*I	Putative primary recombination site TGCTATCAATGGGTTATAC	44973..44991
		Promoter	Putative P_C, LDF 4.37 Binding site for transcription factor: rpoD15	45034..45064
		*att*C	Cassette-associated recombination site	45468..45552
		*att*C	Cassette-associated recombination site	46014..46115
		Gene cassette ORF	Methylcytosine-specific restriction endonuclease HNH family	46132..46929
		*att*C	Cassette-associated recombination site	46899..46957
Halomonas subterranea CGMCC 1.6495	NZ_FOGS01000004.1	*int*I gene	Integron integrase, 314 amino-acid-residues, 7.96% acidic residues	c(292680..293624)
		promoter	Putative P_C, LDF 4.73 Binding sites for transcription factors: rpoD16, lexA and, lexA	293646..293671
		promoter	Putative P_{intI}, LDF 6.5 Binding sites for transcription factors: lexA, cysB, lexA and lexA	c(293653..293683
		*att*I	Putative primary recombination site AGTCAAAATGGTTGGCT	293755.. 293771
		Gene cassette ORF	LysE family translocator (L-lysine exporter)	293860..294480
		*att*C	Cassette-associated recombination site	294480..294557
		ORF	IS*1182* transposase, partial (end of contig)	294625..294870
Chlorogloeopsis fritschii PCC 6912	RSCJ01000013.1	*att*C	Cassette-associated recombination site	56489..56548
		Gene cassette ORF	VOC family protein	56688..57143

		*att*C	Cassette-associated recombination site	57150..57236
		Gene cassette ORF	SDR family oxidoreductase	57512..58201
		*att*C	Cassette-associated recombination site	58244..58327
		Promoter	Promoter for TA system, LDF 3.61 Binding sites for transcription factors: lrp and rpoD15	58374..58407
		ORF	HicB antitoxin	58436..58654
		Promoter	Promoter for putative TA genes, LDF 4.32 Binding sites for transcription factors: rpoD16 and arcA	58689..58715
		Promoter	Promoter for putative TA genes, LDF 1.42 No Binding sites for transcription factors	59003..59028
		ORF	DUF344 family protein (putative antitoxin)	58759..59079
		ORF	DUF5615 family PIN-like protein (putative toxin)	59063..59399
Chlorogloeopsis fritschii PCC 6912	RSCJ01000029.1	*att*C	Cassette-associated recombination site	c(78168..78255)
		Gene cassette ORF	XRE family transcriptional regulator	78363..78686
		*att*C	Cassette-associated recombination site	c(78830..78916)
		Gene cassette ORF	VapC toxin (PIN domain) of TA system	c(78921..79334)
		Promoter	Promoter for toxin gene within antitoxin gene, LDF 0.66 No binding sites for transcription factors	c(79337..79368)
		Gene cassette ORF	DUF2281 domain protein (37% coverage and 75% similarity to CopG transcription factor Putative antitoxin)	c(79331..79615)
		Promoter	Promoter for TA operon, LDF 3.88 Binding sites for transcription factors: fis and soxS	c(79646..79673)
		*att*C	Cassette-associated recombination site	c(79747..79811)

Chlorogloeopsis fritschii PCC 6912	RSCJ01000042.1	Gene cassette ORF	class I SAM dependent methyltransferase	43846..44562
		*att*C	Cassette-associated recombination site	44532..44618
		Gene cassette ORF	Hypothetical protein	44779..45000
		Gene cassette ORF	Hypothetical protein	c(45202..45444)
		Gene cassette ORF	Hypothetical protein	45703..45978
		Gene cassette ORF	Nucleotidyltransferase	46158..47015
		Promoter	Promoter for TA system, LDF 0.99 No binding site for transcription factors	47584..47616
		Gene cassette ORF	BnrT toxin of TA system	47651..47929
		Promoter	Promoter for the antitoxin gene within the upstream toxin gene, LDF 3.3 Binding site for transcription factor: fis	47829..47856
		Gene cassette ORF	BrnA antitoxin of TA system	47892..48170
		*att*C	Cassette-associated recombination site	48165..48250
Halorhodospira halochloris DSM 1059	NZ_AP017372.2	Gene cassette ORF	Hypothetical protein	1184958..1185458
		*att*C	Cassette-associated recombination site	1185442..1185516
		Promoter	promoter for TA operon, LDF 3.41, Binding site for transcription factor: rpoD18	1185541..1185569
		Gene cassette ORF	BrnT family toxin	1185576..1185863
		Gene cassette ORF	BrnA family antitoxin	1185860..1186075
		*att*C	Cassette-associated recombination site	1186078..1186137

		Gene cassette ORF	NgoF VII family restriction endonuclease	1186155..1186583
		Gene cassette ORF	Hypothetical protein	1186587..1187339
		*att*C-like	Putative Cassette-associated recombination site- CAC and GTG instead of the conserved triad (AAC and GTT) in the R" and R' sites, respectively, no unpaired spacer between R & L boxes	1187352..1187425
		H.ha.F1	5' truncated IIB group II Intron	1187659..1188795
		Gene cassette ORF	Bacterial class E intron encoded protein, internal deletion causing a frameshift and an internal stop	1187659..1188685
		Promoter	Putative promoter for the TA system, LDF 2.48 No binding sites for transcription factors	1188819..1188846
		Gene cassette ORF	RelE/ParE family toxin	1188869..1189147
		Gene cassette ORF	HigA family antitoxin	1189158..118472
		*att*C	Cassette-associated recombination site	1189467..1189526
		Gene cassette ORF	DUF1643 domain-containing protein	1189540..1190004
		*att*C	Cassette-associated recombination site	1190007..1190078
		Gene cassette ORF	DUF3800 domain-containing protein	1190085..1190897
		*att*C	Cassette-associated recombination site	1190899..1190970
		Gene cassette ORF	SIR2 family hypothetical protein	1191028..1192185
		Gene cassette ORF	DUF4160 domain-containing protein	1192152..1192376
		Gene cassette ORF	DUF2442 domain-containing protein	1192479..1192751
		Gene cassette ORF	HNH endonuclease	1192791..1193168
		Promoter	promoter for TA operon within upstream ORF, LDF 2.9,	1193084..1193116

			Binding sites for transcription factors: rpoD16, rpoD15 and purR	
		Gene cassette ORF	Antitoxin	1193223..1193453
		Gene cassette ORF	RelE/ParE family toxin	1193453..1193749
		*att*C	Cassette-associated recombination site	1193745..1193828
		Gene cassette ORF	DUF4160 domain-containing protein	1193885..1194151
		Gene cassette ORF	DUF2442 domain-containing protein	1194160..1194408
		Promoter	promoter for TA operon within upstream gene cassette ORF, LDF 0.24, Binding site for transcription factor: rpoD16	1194272..1194303
		*att*C	Cassette-associated recombination site	1194472..1194535
		Gene cassette ORF	BrnT family toxin	1194537..1194854
		Gene cassette ORF	BrnA family antitoxin	1194851..1195123
		Promoter	Putative promoter for TA operon, LDF 3.13, Binding sites for transcription factors: rpoD16 and rpoD17	1195221..1195254
		Gene cassette ORF	BrnT family toxin	1195502..1195870
		Promoter	promoter for antitoxin gene within toxin ORF, LDF 1.78, No binding sites for transcription factors	1195568..1195596
		Gene cassette ORF	BrnA family antitoxin	1195863..1196114
		*att*C	Cassette-associated recombination site predicted by bs folding using MFOLD	1196117..1196210
		H.ha.F2	5’ truncated IIB group II intron	1196335..1197101
		Gene cassette ORF	Bacterial class E intron encoded protein, 5’ deletion	1196335..1196964

		promoter	Promoter for TA operon, LDF 1.05, No binding sites for transcription factors	1196722..1196747
		Promoter	promoter for antitoxin gene within upstream toxin gene, LDF 0.89, No binding sites for transcription factors	1197092..1197120
		Gene cassette ORF	HicA family toxin-frame-shift due to 1 nucleotide deletion at 1197266 position	1197111.. 1197370
		Promoter	promoter for antitoxin gene within upstream toxin gene, LDF 2.37, No binding sites for transcription factors	1197185..1197210
		Gene cassette ORF	HicB family antitoxin	1197360..1197572
		*att*C	Cassette-associated recombination site	1197567..1197626
		IS*200/605*	IS*Hahl*1	1197652..1199464
		ORF	IS*Hahl1* TnpA (transposase)	c(1197731..1198045)
		ORF	IS*Hahl1* TnpB (accessory protein)	1198170..1199444
Marinobacter salinus strain Hb8	NZ_CP017715.1	*att*C	Cassette-associated recombination site	c(1657629..1657727)
		Gene cassette ORF	Uracil DNA glycosylase	c(1657722..1658327)
		Gene cassette ORF	M23 family metallopeptidase	c(1658297..1658857)
		Gene cassette ORF	Ferritin	c(1658997..1659284)
		*att*C	Cassette-associated recombination site	c(1659321..1659398)
		Gene cassette ORF	Txe/YoeB family toxin	c(1659400..1659636)
		Gene cassette ORF	Hypothetical protein	c(1659678..1659971)
		Gene cassette ORF	GNAT N-acetyltransferase	c(1659952..1660488)
		Gene cassette ORF	Hypothetical protein	c(1660543..1661178)

		Gene cassette ORF	RelE/ParE family toxin	c(1661270..1661551)
		Gene cassette ORF	Antitoxin (RHH-CopG family)	c(1661548..1661820)
		Promoter	Promoter for TA toxin within antitoxin gene, LDF 1.58 Binding site for transcription factor: oxyR	c(1661753..1661786)
		Promoter	Promoter for TA operon, LDF 2.15 No binding sites for transcription factors	c(1661853..166q882)
		Gene cassette ORF	Acetyltransferase	c(1661837..1661965)
		*att*C	Cassette-associated recombination site	c(1661887..1661963)
		Gene cassette ORF	Txe/YoeB family toxin	c(1661965..1662228)
		Gene cassette ORF	Phd/YefM family antitoxin	c(1662225..1662467)
		IS*1182*	IS*1182* with indels and frameshifts-probably non-functional	c(1665285..1666460)
		ORF	IS*1182* transposase	c(1665402..1666387)-frameshift at 1665611 & 2 indels
		IS*1182*	IS*Masa1*	c(1667140..1668969)
		ORF	IS*1182* transposase	c(1667325..1668896)
		IS*1182*	IS*Masa1*	c(1669539..1671380)
		ORF	IS*1182* transposase	c(1669724..1671295)
		*att*C	Cassette-associated recombination site	c(1673240..1673328)
		Gene cassette ORF	Hypothetical protein	c(1673334..1673702)
		Gene cassette ORF	Hypothetical protein	c(1673712..1673906)
		IS*1182*	IS*Masa2*	c(1673712..1675625)
		Gene cassette ORF	IS*1182* transposase	c(1673903..1675474)
		Gene cassette ORF	AAA family ATPase	c(1675629..1677299)

		IS*1182*	IS*Masa1*	c(1677335..1679164)
		Gene cassette ORF	IS*1182* transposase	c(1677520..1679091)
		Gene cassette ORF	Hypothetical protein	c(1679253..1679585)
		*att*C	Cassette-associated recombination site	c(1679603..1679704)
		Gene cassette ORF	Hypothetical protein	c(1679699..1680049)
		Gene cassette ORF	Bacterial class C group II intron encoded protein	c(1680170..1681510)
		Gene cassette ORF	Hypothetical protein	c(1682016..1682237)
		Promoter	Promoter for the TA operon, LDF 4.36 Binding site for transcription factor: pdhR	1682322..1682350
		Gene cassette ORF	Phd/YefM family antitoxin	1682385..1682636
		Gene cassette ORF	Txe/YoeB family toxin	1682633..1682887
		Gene cassette ORF	Hypothetical protein	1682847..1683320
		*att*I	Putative primary recombination site GTTTCACCGTAGGTTAGCG	c(1683265..1683283)
		*att*I	Putative primary recombination site GTATAATTAGCTGTTAAAG	c(1683315..1683333)
		Promoter	Putative P_C, LDF 4.57 Binding site for transcription factor: rpoD16	c(1683327..1683352)
		Promoter	Putative P_{intI}, LDF 6.93 Binding sites for transcription factors: ihf, lexA, cysB, lexA, lexA and lexA	1683618..1683646
		Promoter	Putative P_C, LDF 4.09	c(1683630..1683658)

			Binding sites for transcription factors: rpoD16, lexA and lexA	
		*int*I gene	Integron integrase, complete, 328 amino-acid-residues, 10.06% acidic residues	1683671..1684657
Pseudomonas salegens strain CECT 8338	NZ_LT629787.1	Gene cassette ORF	Hypothetical protein	819662..820237
		*att*C	Cassette-associated recombination site	820253..820330
		*att*C	Cassette-associated recombination site	820690..820737
		Gene cassette ORF	trypsin	821187..821579
		Gene cassette ORF	Hypothetical protein	821687..822322
		IS*1182*	IS*1182* with a frameshift within its transposase sequence-probably non-functional	822400..823479
		Gene cassette ORF	IS*1182* transposase	822459..823324-frameshift at 822929
		Gene cassette ORF	Hypothetical protein	823499..823807
		Gene cassette ORF	Toll/interleukin-1 receptor domain-containing protein (TIR domain)	823909..824865
		*att*C	Cassette-associated recombination site	824869..824916
		IS*1182*	IS*1182* with a frameshift within its transposase sequence-probably non-functional	824913..825732
		Gene cassette ORF	IS*1182* transposase	824972..825578-frameshift at 825256
		Gene cassette ORF	GIY-YIG nuclease family protein	825769..826704
		Gene cassette ORF	GFA family protein	826801..827202
		Gene cassette ORF	Antibiotic biosynthesis monooxygenase	827300..827608
		Gene cassette ORF	Hypothetical protein	827717..828421

		Gene cassette ORF	Hypothetical protein	828518..829042
		Gene cassette ORF	Hypothetical protein	c(829057..829155)
		Gene cassette ORF	Hypothetical protein	829382..829777
		Promoter	Promoter for TA operon, LDF 2.36 Binding site for transcription factor: rpoS17	830190..830221
		Gene cassette ORF	Phd/YefM family antitoxin	830284..830553
		Gene cassette ORF	RelE/ParE family toxin	830554..830853
		*att*C	Cassette-associated recombination site	830848..830925
		Gene cassette ORF	Hypothetical protein	830944..831201
		*att*C	Cassette-associated recombination site	831222..831299
		*att*C	Cassette-associated recombination site	c(2318309..2318377)
		Gene cassette ORF	Hypothetical protein	c(2318381..2318863)
		Gene cassette ORF	Hypothetical protein	c(2318960..2319241)
		Gene cassette ORF	DUF 2570 domain-containing protein	c(2319398..2319793)
		*att*C	Cassette-associated recombination site	c(2319809..2319886)
		Gene cassette ORF	Hypothetical protein	c(2319778..2319888)
		Gene cassette ORF	Txe/YoeB family toxin	c(2319888..2320151)
		Gene cassette ORF	Phd/YefM family toxin	c(2320148..2320390)
		Promoter	Promoter for TA operon, LDF 2.74 No binding sites for transcription factors	c(2320423..2320450)

		*att*C	Cassette-associated recombination site	c(2320463..2320540)
		Gene cassette ORF	DUF 1272 domain-containing protein	c(2320535..2320783)
		*att*C	Cassette-associated recombination site	c(2320802..2320879)
		Gene cassette ORF	Hypothetical protein	c(2320874..2321227)
		Gene cassette ORF	IS*1182* transposase	c(2321861..2323429)
		IS*1182*	IS*Pssa1*	c(2321708..2323488)
		Gene cassette ORF	Hypothetical protein	c(2323579..2324097)
		*att*C	Cassette-associated recombination site	c(2324102..2324179)
		*att*I	Putative primary recombination site GCCCAAAAGCAAGGTTAAAT	c(2324717..2324736)
		Promoter	Putative P_C, LDF 3.53 Binding sites for transcription factors: fur, rpoS17 and nagC	c(2324748..2324776)
		Promoter	Putative P_{intI}, LDF 3.77 Binding sites for transcription factors: argR2 and phoB	2324887..2324915
		*int*I gene	Integron integrase-A, 328 amino-acid-residues,9.45 % acidic residues	2324945..2325931
		Promoter	Putative P_C, LDF 0.62 Binding site for transcription factor: metR	c(2325069..2325097)
		*att*C	Cassette-associated recombination site	c(2639505..2639582)
		Gene cassette ORF	DUF 1993 domain-containing protein	c(2639641..2640207)
		Gene cassette ORF	Demethoxyubiquinone hydroxylase family protein	c(2640264..2640773)
		Gene cassette ORF	Hypothetical protein	c(2640881..2641450)
		Gene cassette ORF	Txe/YoeB family toxin	c(2641450..2641713)

		Gene cassette ORF	Phd/YefM family antitoxin	c(2641710..2641952)
		Promoter	Promoter for TA operon, LDF 3.73 Binding site for transcription factor: ihf	c(2641984..2642011)
		Gene cassette ORF	RelE/ParE family toxin	c(2642154..2642450)
		Gene cassette ORF	ParD family antitoxin	c(2642450..2642704)
		Promoter	Promoter for TA operon, LDF 4.34 Binding site for transcription factor: narL	c(2642741..2642768)
		Gene cassette ORF	HigA family antitoxin	c(2642857..2643171)
		Gene cassette ORF	RelE/ParE family toxin	c(2643191..2643466)
		Promoter	Promoter for TA operon, LDF 1.74 Binding site for transcription factor: rpoD19	c(2643489..2643515)
		Promoter	Promoter for TA operon, LDF 3.89 Binding sites for transcription factors: arcA and lexA	264359..2643618
		Gene cassette ORF	VapB family antitoxin	2643657..2643851
		Gene cassette ORF	VapC toxin family (PIN domain)	2643848..2644249
		Gene cassette ORF	IS*1182* transposase	c(2644406..2645738)-frameshift at 2644906
		IS*1182*	IS*Pssa1*-isoform with a deletion and a frameshift within transposase sequence-probably non-functional	c(2644251..2645797)
		Gene cassette ORF	DUF 4262-domain containing protein	c(2645866..2646507)
		*att*C	Cassette-associated recombination site	c(2646527..2646604)
		Gene cassette ORF	Hypothetical protein	c(2646611..2646961)
		*att*C	Cassette-associated recombination site	c(2646975..2647085)

		Gene cassette ORF	Hypothetical protein	c(2647080..2647874)
		Gene cassette ORF	Hypothetical protein	c(2647876..2648511)
		*att*C	Cassette-associated recombination site	c(2648525..2648602)
		Gene cassette ORF	Hypothetical protein	c(2648604..2648849)
		*att*C	Cassette-associated recombination site	c(2648871..2648943)
		Gene cassette ORF	DUF 2971 domain-containing protein	c(2648946..2649839)
		Gene cassette ORF	Nucleotidyltransferase	c(2649948..2651345)
		*att*C	Cassette-associated recombination site	c(2651400..2651477)
		Gene cassette ORF	DUF 1493 domain-containing protein	c(2651472..2651810)
		IS*21*	IS*Pssa2*	2651814..2653797
		Gene cassette ORF	IS*21* transposase	2651933..2652931
		Gene cassette ORF	IS*21*-like element helper ATPase IstB	2652928..2653725
		Gene cassette ORF	Hypothetical protein	c(2653746..2653883)
		*att*C	Cassette-associated recombination site	c(2653804..2653881)
		Gene cassette ORF	Hypothetical protein	c(2653885..2654211)
		*att*C	Cassette-associated recombination site	c(2654230..2654301)
		Gene cassette ORF	Hypothetical protein	c(2654304..2654876)
		*att*C	Cassette-associated recombination site	c(2654898..2654975)
		Gene cassette ORF	Hypothetical protein	c(2654983..2655198)
		*att*C	Cassette-associated recombination site	c(2655250..2655327)

		*att*I	Putative primary recombination site GTCTAATCACTGTTATGT	c(2655646..2655663)
		Promoter	Putative P_C, LDF 3.06 No binding sites for transcription factors	c(2655707..2655731)
		Promoter	Putative P_{intI}, LDF 4.06 Binding sites for transcription factors: OmpR, lexA and lexA	2655789..2655818
		*int*I gene	Integron integrase-B, 321 amino-acid-residues, 8.1 % acidic residues	2655849..2656814
Natrialbaceae archaeon XQ-INN 246 strain 2447	NZ_CP050695.1	IS*66*	IS*Narch2*	c(918828..921142)
		ORF	IS*66* transposase	c(918850..920454)
		ORF	IS*66* TnpB accessory protein	c(920500..920856)
		ORF	Hypothetical accessory gene	c(920853..921059)
		ORF	AAA family ATPase	c(921914..923797)
		*att*C	Cassette-associated recombination site	c(923871..923945)
		Gene cassette ORF	Hypothetical protein	c(923940..924713)
		*att*I	Putative primary recombination site GATC**CAT**TCACT**GTT**AGAC	c(924729..924747)
		Promoter	Putative P_{intI}, LDF 3.07 No binding sites for transcription factors	924871..924899
		Promoter	Putative P_C, LDF 1.88 Binding sites for transcription factor: rpoD18	c(924895..924923)
		*int*I gene	Integron integrase, 390 amino-acid-residues, 8% acidic residues.	924992..926164
		IS*21*	IS*Nacrch3*	c(930308..933272)
		ORF	IS*21* helper accessory protein	c(930730..931479)
		ORF	IS*21* transposase Tnp	c(931503..933038)
Euryarchaeota archaeon isolate J059 k99_253731	RFHV01000337.1	Gene cassette ORF	Hypothetical protein	c(43..798)
		*att*C	Cassette-associated recombination site	c(824..1010)

		Gene cassette ORF	OsmC family peroxiredoxin (1.11.1.-) (oxidoreductase)	c(1111..1620)
		Gene cassette ORF	Hypothetical protein	c(2197..2871)
		Promoter	Putative P_{intI}, LDF 5.11 Binding sites for transcription factors: rpoD15, tyrR, metR, phoB, rpoD19, hipB, rpoD16 and rpoH2	2867..2898
		*att*I	Putative primary recombination site ATAA**AAA**GGACT**GTT**CGGT	2897..2915
		Promoter	Putative P_C, LDF 6.61 Binding sites for transcription factors: rpoD18, rpoH3 and cpxR	c(2948..2976)
		*int*I gene	Integron integrase, partial (end of contig).	3075..4178
Euryarchaeota archaeon isolate J059 k99_312182	RFHV01000400.1		Integron integrase, complete, 321 amino-acid-residues, 9.66 % acidic residues.	30..995
Candidatus Aenigmarchaeota archaeon isolate B34_G1 B34_Guay1_scaffold_69367	QMZW01000251.1	*int*I gene	Integron integrase, complete, 227 amino-acid-residues, 10.13 % acidic residues.	c(385..1068)

Appendix B: Chapter 5 Supplementary Tables

TableS5.1 Analyzed *Caldivirga* spp genomes

***Caldivirga* analyzed genomes**	**genome size (Mb)**	**sequencing status**	**genome or WGS accession number**
Caldivirga maquilingensis IC-167	2.07757	complete	NC_009954.1
Caldivirga sp. SpSt-118	2.19	partial	DSBU00000000.1
Caldivirga sp. EvPrim.Bin7	1.76	partial	WYEH00000000.1
Caldivirga sp. CIS_19	1.45	partial	LOCC00000000.1
Caldivirga sp. JCHS_4	1.35	partial	LOCD00000000.1
Caldivirga sp. MG_3	1.6	partial	LOCB00000000.1
Caldivirga sp. MU80	2.26	partial	LCTF00000000.1
Caldivirga sp. UBA161	1.86	partial	DAXS00000000.1

TableS5.2 Genetic elements of identified complete integrons and CALINs within studied metagenomes of hypersaline environments and genomes of different *Caldivirga* spp. Only CALINs with toxin-antitoxin (TA) systems, insertion sequences (IS) or known antibiotic resistance genes (ARG) are shown.

site	**Accession no.**	**Genetic element**	**Annotation (description)**	**position**
Th	AGBJ01000022.1	*int*I gene	Integron integrase, complete, 322 amino-acid-residues, 10.87% acidic residues.	c(19612..20580)
		Promoter	Putative P_C, LDF 5.41 Binding sites for transcription factors: modE, rpoD17 and rpoD15	20494..20525
		Promoter	Putative P_{intI}, LDF 3.2 Binding sites for transcription factors: argR2, rpoD15, tyrR, lexA and ada	c(20692..20720)
		Promoter	Putative P_C, LDF 13.55 Binding sites for transcription factors: lexA, rpoD17, lrp, lrp, lexA, argR, argR2, ihf, argR2, rpoD16, fnr and arcA	20894..20920
		Promoter	Putative P_{intI}, LDF 10.82 Binding sites for transcription factors: lrp, rpoH2, rpoH2, lexA, argR, tyrR, rpoD18 and cpxR	c(20989..21014)

Th	AGBJ01001039.1	*int*I gene	Integron integrase, complete, 323 amino-acid-residues, 9.9% acidic residues	c(26..997)
		Promoter	Putative Pc, LDF 4.46 Binding sites for transcription factors: farR and oxyR	1015..1047
		Promoter	Putative P_{intI}, LDF 324 Binding sites for transcription factors: ompR, lrp, lrp and rpoD17	c(1073..1100)
		*att*C	Secondary integration site	1265..1321
Th	AGBJ01007148	*int*I gene	Integron integrase, partial from both ends: missing patch I and very short sequence from box II (very short contig).	c(2..532)
Th	AGBJ01001366.1	Promoter	Putative promoter for the TA system, LDF 2.34 Binding sites for transcription factors: OmpR, ihf and crp	47..75
		*att*C	Cassette-associated recombination site	99..234
		Promoter	Putative promoter for the antitoxin in TA system, LDF 5.06 Binding site for transcription factor: fnr	252..280
		Gene cassette ORF	BrnT toxin in TA system	279..575
		Gene cassette ORF	Hypothetical protein: Putative antitoxin in TA system	572..778
		*att*C	Cassette-associated recombination site	788..912
GNM2	ABPQ01003014.1	*att*C	Putative Cassette-associated recombination site-AAT instead of AAC in R''	c(251..328)
		*att*C	Cassette-associated recombination site	c(584..642)
		Gene cassette ORF	HicA toxin-partial (end of contig)	c(663..797)
GNM2	ABPQ01006959.1	*att*C	Cassette-associated recombination site	c(99..172)
		Gene cassette ORF	HTH protein (putative antitoxin of TA system)	c(196..597)
		*att*C	Cassette-associated recombination site	c(220..262)
		Gene cassette ORF	HigB toxin of TA system, partial (end of contig)	c(573..866)

GNM3	ABPQ01007625.1	*int*I gene	Integron integrase, partial: missing box II (end of contig), pseudogene: missing patch I (perhaps due to an indel)	74..733
		Promoter	Putative P_C within *int*I gene, LDF 0.82 Binding sites for transcription factors: araC and soxS	c(319..348)
GNM3	ABPQ01010372.1	ORF	HicB family antitoxin	c(1..141)
		*int*I gene	Integron integrase, partial: missing N terminus, pseudogene: frameshift, missing box II due to possible deletion at the C terminus.	c(87..861)-frameshift at 156
GNM4	ABPS01005223.1	*int*I gene	Integron integrase, partial: missing few residues but all necessary domains are detected.	c(44..745)
GNM5	ABPT01000232.1	*int*I gene	Integron integrase, partial: missing patch I, box I and box II (short contig), pseudogene	2..640
GNM6	ABPU01005246.1	Promoter	Putative P_C, LDF 2.29 Binding sites for transcription factors: argR and fruR	c(55..84)
		Promoter	Putative $P_{intI,}$ LDF 3.47 Binding sites for transcription factors: cysB and OmpR	84..117
		*int*I gene	Integron integrase, partial: missing some residues at C terminus	130..307
		Promoter	Putative P_C, LDF 3.59 Binding site for transcription factor: rpoD17	c(185..217)
		Promoter	Putative P_C, LDF 4.31 Binding site for transcription factors: rpoD17, purR, rpoD18 and fnr	c(467..495)
GNM7	ABPV01008848.1	*int*I gene	Integron integrase. Partial: short contig, but all necessary domains are present. (Y220L: unknown effect on activity)	3..815
GNM7	ABPV01012279.1	*int*I gene	Integron integrase, partial: missing patch I, pseudogene: frameshift and	3..838 -frameshift at 626
GNM9	ABPX01006760.1	*int*I gene	Integron integrase, partial: missing most of patch I (short contig).	3..626
GNM10	ABPY01004164.1	*int*I gene	Integron integrase, complete, pseudogene: frameshift. 231amino-acid-residues, 9.1% acidic residues	116.. 712 -frameshift at 189
BSL	LFCJ01003999.1	*att*C	Cassette-associated recombination site	53..182
		*att*C	Cassette-associated recombination site	563..694
		Promoter	Promoter for TA operon, LDF 4.77 Binding sites for transcription factors: fur, metJ, ompR. tus	969..997

		Gene cassette ORF	RelE/ParE toxin	1004..1285
		Gene cassette ORF	HigA family antitoxin	1308..1604
		*att*C	Cassette-associated recombination site	1635..1762
		Promoter	Promoter for TA system and upstream ORF, LDF 2.19 Binding site for transcription factor: rpoD19	1642..1675
		Gene cassette ORF	GIY-YIG nuclease	1783..2091
		Promoter	Promoter for TA system and upstream ORF, LDF 1.63 No binding sites for transcription factors	2093..2121
		Gene cassette ORF	BrnT family toxin	2130..2396
		Gene cassette ORF	BrnA family antitoxin	2393..2698
		*att*C	Cassette-associated recombination site	2693..2821
		Promoter	Promoter for TA operon, LDF 2.68 No binding sites for transcription factors	2970..2999
		Gene cassette ORF	Antitoxin (CopG TR-RHH)	3030..3278
		Gene cassette ORF	Toxin (PIN domain)	3259..3675
		*att*C	Cassette-associated recombination site	3763..3891
		Gene cassette ORF	Fic family protein	4050..5150
		*att*C	Cassette-associated recombination site	5151..5279
		Gene cassette ORF	AAA family ATPase	5352..6536
		*att*C	Cassette-associated recombination site	6531..6658
		Promoter	Promoter for TA operon, LDF 3.28 Binding site for transcription factor: lexA	6853..6881
		Gene cassette ORF	HicB antitoxin	6892..7134
		*att*C	Cassette-associated recombination site	7054..7173
		Gene cassette ORF	HicA family toxin	7134..7358
		*att*C	Cassette-associated recombination site	7395..7523

		Promoter	Promoter for TA operon, LDF 2.83 No binding sites for transcription factors	7570..7601
		ORF	YefM antitoxin	7626..7880
		ORF	Txe/YoeB toxin	7877..8131
TTCSL	LFFM01001065.1	*att*C	Cassette-associated recombination site	9541..9655
		Promoter	Promoter for the TA operon, LDF 0.76 No binding sites for transcription factors	9652..9684
		Promoter	Promoter for the TA operon, LDF 4.62 No binding sites for transcription factors	9806..9833
		Gene cassette ORF	DUF433 protein-putative antitoxin	9748..9960
		Gene cassette ORF	DUF5615 protein (PIN-like domain)-putative toxin	9957..10280
		*att*C	Cassette-associated recombination site	10291..10407
		Promoter	Promoter for the TA operon, LDF 0.43 No binding sites for transcription factors	10442..10471
		Gene cassette ORF	DUF433 protein-putative antitoxin	10477..10698
		Gene cassette ORF	DUF5615 protein (PIN-like domain) -putative toxin	10695..11039
		*att*C	Putative Cassette-associated recombination site-GCT instead of GTT in R’	11044..11166
		Promoter	Promoter for the TA operon, LDF 0.45 Binding site for transcription factor: rpoD17	11075..11103
		Gene cassette ORF	DUF433 protein-putative antitoxin	11242..11550
		Promoter	Promoter for the toxin gene within upstream antitoxin, LDF 0.63 No binding sites for transcription factors	11372..11397
		Gene cassette ORF	DUF5615 protein (PIN-like domain) -putative toxin	11551..11880
		*att*C	Cassette-associated recombination site	11905..12035
		Gene cassette ORF	Hypothetical protein	12254..12370
		Promoter	Promoter for the TA operon, LDF 2.11 Binding site for transcription factor: rpoD18	12538..12560
		*att*C	Cassette-associated recombination site	12697..12810

		Promoter	Promoter for the TA operon, LDF 3.49 Binding sites for transcription factors: rhaS, rpoD18	12867..12894
		Gene cassette ORF	Hypothetical protein	12893..13114
		Gene cassette ORF	PIN domain containing nuclease	13118..13534
		*att*C	Putative Cassette-associated recombination site-ACC instead of AAC in R''	13546..13665
TTCSL	LFFM01001574.1	ORF	Tyrosyl-DNA-phosphodiesterase	6538..8658
		ORF	DNA helicase UvrD	8664..10724
		*att*C	Cassette-associated recombination site	8666..8708
		ORF	nuclease	10721..11626
		ORF	ATP dependant helicase	11623..15084
		*att*C	Cassette-associated recombination site	12214..12263
TTCSL	LFFM01002330.1	ORF	ABC transporter	690..2627
		*att*C	Cassette-associated recombination site	1065..1223
		ORF	Hypothetical protein	c(2638..2988)
		*att*C	Cassette-associated recombination site	2722..2769
TTCSL	LFFM01004875.1	ORF	Hypothetical protein, partial	3..1193
		*att*C	Cassette-associated recombination site	c(1101..1154)
		ORF	glycosyltransferase	c(1243..2412)
		*att*C	Cassette-associated recombination site	c(1666..1729)
PSL	LKMJ01007318.1	*int*I gene	Integron integrase, complete. 466amino-acid-residues, 11.8% acidic residues	c(8202..9599)
		Promoter	Putative Pc, LDF 4.04. Binding site for transcription factor: rpoD15	9582..9609
		Promoter	Putative P_{intI}, LDF 4.8 Binding site for transcription factor: farR	c(9636..9659)
		*att*I	Putative primary recombination site TCTAAACCACTGTTATGC	9723..9744
		Gene cassette ORF	Hypothetical protein	9751..10035
		Gene cassette ORF	Hypothetical protein, partial:-end of contig before any possible *att*C sites	10155..10406

PSL	LKMJ01017989.1	*int*I gene	Integron integrase, complete, pseudogene: part of boxII is missing, with no catalytic tyrosine and no frameshifts detected. 283amino-acid-residues, 8.13% acidic residues.	c(464..1315)
		Promoter	Putative P_C, LDF 6.67 Binding sites for transcription factors: ihf, phoB3, fis, ihf and ihf	1338..1371
		Promoter	Putative P_{intI}, LDF5.55 Binding sites for transcription factors: ihf, crp, deoR, ihf, argR2 and lrpa	c(1369..1400)
		*att*I	Putative primary recombination site AAAATGATACGTTGGTT	1551..1571
		Gene cassette ORF	ArcDNA binding protein (putative antitoxin in TA system)	1590..1832
		Gene cassette ORF	vapC toxin (PIN domain)	1829..2230
		*att*C	Cassette-associated recombination site	2244..2333
TSL	LFIK01004686.1	*att*I	Putative primary recombination site (incomplete as it is at the contig periphery) TGCTAATTATATGTTA.	c(1..16)
		Promoter	Putative P_C, LDF 2.91 Binding sites for transcription factors: phoB, phoB3, rpoD1 and, arcA	c(37..66)
		*int*I gene	Integron integrase, complete, 314amino-acid-residues, 10.19% acidic residues.	97..1041
TSL	LFIK01005835.1	*int*I gene	Integron integrase, complete. 431 amino-acid-residues, 10% acidic residues	c(295..1590)
		Promoter	Putative P_{intI}, LDF 2.58 No transcription factors binding sites	c(1721..1750)
		Promoter	Putative P_C, LDF 2.51 Binding site for transcription factor: crp	1789..1822
		*att*I	Primary recombination site GCCCAATATACGTTAAAT	1811..1829
		Gene cassette ORF	Hypothetical protein	1844..2503
		*att*C	Cassette-associated recombination site	2498..2580
		Promoter	Promoter for TA system, LDF 4.27 Binding site for transcription factor: rpoD19	2576..2604
		Gene cassette ORF	Antitoxin of TA system	2631..2855
		Gene cassette ORF	ParE toxin of TA system	2900..3151

		*att*C	Cassette-associated recombination site	3146..3216
		Gene cassette ORF	Hypothetical protein	3236..3589
		*att*C	Cassette-associated recombination site with AAT instead of AAC in R"	3581..3664
		ORF	VapC toxin (PIN domain) of TA system	c(3682..3957)
		ORF	vapB antitoxin	c(4054..4249)
		promoter	Promoter for TA system, LDF 2.45 Binding sites for transcription factors: crp and dnaA	4281..4308
TSL	LFIK01006738.1	*int*I gene	Integron integrase, complete., 423amino-acid-residues, 10.64% acidic residues.	c(7333..8601)
		Promoter	Putative P_C, LDF 1.57 Binding sites for transcription factors: rpoD15 and phoB	8210..8243
		Promoter	Putative $P_{C,}$ LDF 3.88 Binding sites for transcription factors: hns, fis, ihf and arcA	8597..8620
		Promoter	Putative P_{intI}, LDF 7.19 Binding sites for transcription factors: crp, lexA, argR, argR2, ihf, argR2 and metR	c(8616..8646)
		*att*I	Putative primary recombination site CCCTAACAGAGGCGTTAGGG	8782..8801
		Promoter	Putative P_C, LDF 1.93 Binding site for transcription factor: rpoD17	8901..8928
		Promoter	Putative $P_{intI,}$ LDF 2.63 No transcription factors binding sites	c(8944..8972)
TSL	LFIK01016104.1	*Int*I gene	Integron integrase, complete, 279 amino-acid-residues, 8.6% acidic residues.	c(2436..3272)
		Promoter	Putative P_C, LDF 3.61 Binding sites for transcription factors: ompR and lexA	3419..3446
		Promoter	Putative P_{intI}, LDF 2.49 Binding sites for transcription factors: rpoD15, lexA and metJ	c(3439..3470)
		Gene cassette ORF	Hypothetical protein	3622..4362
		*att*C	Cassette-associated recombination site	4315..4370
		Gene cassette ORF	Hypothetical protein	4384..4860
		*att*C	Cassette-associated recombination site	4855..4921
TSL	LFIK01007609.1	ORF	ParE toxin in a TA system	c(213..503)
		ORF	antitoxin in a TA system	c(500..733)
		ORF	Hypothetical protein	c(825..1112)

		Promoter	Promoter for TA system, LDF 8.44 Binding sites for transcription factors: rpoD17, araC, lrp, fnr, crp, purR, lexA and purR	c(1193..1220
		Promoter	Promoter for TA system, LDF 1.47 Binding site for transcription factor: rpoD18	c(1501..1530)
		Promoter	Putative P_{intI} , LDF 4.09 Binding sites for transcription factors: rpoD15, argR2, lexA and purR	1803..1832
		Promoter	Promoter for TA system, LDF 4.8 Binding sites for transcription factors: arcA, soxS, rpoD16, rpoD16, phoB3, ihf and rpoH3	c(1808..1832)
		*int*I gene	Integron integrase, complete, 320 amino-acid-residues, 8.44% acidic residues	1828..2790
		Promoter	Putative P_C, LDF 7.01 Binding sites for transcription factors: rpoD17, purR, nagC, ihf, lexA and argR2	2927..2953
		*att*I	Putative primary recombination site (inverted integron) GCGTAAAAAGCCCGTTGGAC	2996..3015
		Gene cassette ORF	S9 Family Peptidase	c(3214..5293)
		Gene cassette ORF	(HAD) hydrolase-like protein	5479..6307
		*att*C	Cassette-associated recombination site (within previous ORF)	5947..5996
		*att*C	Cassette-associated recombination site	6317..6363
TSL	LFIK01017073.1	Gene cassette ORF	Hypothetical protein (at the contig periphery thus no *att*C detected downstream)	c(1..2190)
		*att*C	Cassette-associated recombination site	c(2215..2304)
		Gene cassette ORF	Hypothetical protein	c(2299..2973)
		*att*C	Cassette-associated recombination site, within ORF	c(2550..2596)
		Gene cassette ORF	Hypothetical protein	c(2986..3462)
		*att*C	Cassette-associated recombination site	c(3504..3548)
		Promoter	Putative $P_{intI,}$ LDF 5.65 Binding sites for transcription factors: carP, purR, cytR, ihf and cpxR	3657..3686
		Promoter	Putative P_C, LDF 8.05 Binding sites for transcription factors: rpoD16, phoB, cpxR, rpoD19, rpoD19, rpoD18 and phoB3	c(3685..3718)

		*int*I gene	Integron integrase, complete, 317 amino-acid-residues, 8.52% acidic residues	3723..4673
TSL	LFIK01005867.1	Promoter	Putative promoter for the toxin ORF, LDF 4.22, Binding sites for transcription factors rpoD16, ihf and phoB	c(2787..2817)
		Promoter	Putative promoter for the toxin ORF, LDF 3.34, Binding site for transcription factor: rpoD16	c(2818..2847)
		Promoter	Putative promoter for the TA operon, LDF 1.79, No transcription factors binding sites	c(2862..2890)
		Promoter	Putative promoter for the TA operon, LDF 0.72, No transcription factors binding sites	c(2890..2918)
		*att*C	Cassette-associated recombination site	c(2897..3020)
		Gene cassette ORF	Hypothetical protein	c(3178..3603)
		UHB.F1 ORF	Intron encoded protein (group II reverse transcriptase/maturase), 411 amino-acid-residues	3872..5111
		UHB.F1	5’ truncated group IIC intron	3872..5204
		Gene cassette ORF	Hypothetical protein	c(5183..5524)
		Gene cassette ORF	Serine hydrolase (betalactamase transpeptidase)	c(5676..6734)
		*att*C	Cassette-associated recombination site	c(6840..6925)
		Gene cassette ORF	No significant similarity	c(6893..7672)
		*att*C	Cassette-associated recombination site	c(7679..7748)
		Gene cassette ORF	Hypothetical protein	c(7726..8007)
		*att*C	Cassette-associated recombination site	c(8025..8092)
		Gene cassette ORF	Txe/YoeB family addiction module toxin	c(8092..8346)
		Gene cassette ORF	YoeB-YefM toxin-antitoxin system antitoxin YefM	c(8343..8594)

		*att*C	Cassette-associated recombination site	c(8641..8714)
		Gene cassette ORF	Hypothetical protein	c(8709..8987)
		*att*C	Cassette-associated recombination site	c(8999..9068)
		Gene cassette ORF	Hypothetical protein	c(9192..9608)
TSL	LFIK01005957.1	ORF	RelE/ParE family toxin	c(3923..4213)
		ORF	ParD-like antitoxin	c(4200..4418)
		Promoter	Putative promoter for the toxin gene, LDF 0.97, No transcription factors binding sites	c(4405..4435)
		*att*C	Cassette-associated recombination site	c(4463..4591)
		Promoter	Putative promoter for TA operon, LDF 2.17, Binding sites for transcription factors crp and rpoD19	c(4709..4737)
		Gene cassette ORF	Hypothetical protein	c(4525..5097)
		IEP	Chloroplast-like 1(CL1) intron encoded protein, 500 amino-acid-residues	c(5223..6725)
		UHB.I2	Group IIB Intron	c(5096.. 7296)
		*att*C	Cassette-associated recombination site	c(7295..7364)
		Gene cassette ORF	PH domain-containing protein	c(7420..7884)
		*att*C	Cassette-associated recombination site	c(7890..7959)
		Gene cassette ORF	HNH endonuclease	c(7930..8415)
		*att*C	Cassette-associated recombination site	c(8427..8512)
		Gene cassette ORF	Putative GNAT N-acetyltransferase (30% identity)	c(8588..8971)
		*att*C	Cassette-associated recombination site	c(8978..9047)
		Gene cassette ORF	Hypothetical protein	c(9053..9580)

		*att*C	Cassette-associated recombination site	c(9589..9658)
		Gene cassette ORF	Hypothetical protein, partial	c(9627..9770)
KD UINF	contig00306, [168], [164] SRX352368	*int*I gene	Integron integrase, complete, 343 amino-acid-residues, 9.6% acidic residues	c(16..1047)
		Promoter	Putative P_{intI}, LDF 5.78 Binding sites for transcription factors: rpoD16, rpoD16, rpoD17 and ilvY	c(1085..1118)
		Promoter	Putative P_C, LDF 10.34 Binding sites for transcription factors: rpoD17, lexA, lexA, rpoD17, arcA, tus, fnr	1096..1127
		*att*I	Putative primary recombination site GTTTAAATGTTGTTCAAC	1149..1167
		Gene cassette ORF	Crp/Fnr family transcriptional regulator	1226..1972
		Gene cassette ORF	methylmalonyl CoA mutase associated GTPase MeaB	1991..2926
		Gene cassette ORF	Twin Argnine translocase TatC subunit	2919..3641
		Gene cassette ORF	polyprenyl synthetase family protein	3651..4640
		Gene cassette ORF	glycosyltransferase 9	, 4643..5665
		Gene cassette ORF	3deoxyD-manno-octulosonic acid transferase domain containing protein	5662..5642
		Gene cassette ORF	ATP-binding cassette transporter	6935..8251
		Gene cassette ORF	alcohol phosphatidyl transferase-frameshift	8244.. 8806 Insertion 8373..8386
		Gene cassette ORF	signal recognition particle docking protein FtsY	8810..9206
KD UINF	contig01002, [168], [164], SRX352368	*int*I	Integron integrase, partial (small contig). (D161E: similar mutations showed increase in recombination activity [298])	3..998
KD UINF	contig06491, [168], [164], SRX352368	*att*I	Putative primary recombination site ATTCAACGTAGCCGTTCTGT	59..78
		Promoter	Putative P_{intI}, LDF 5.38 Binding sites for transcription factors: lexA, nagC and soxS	97..125
		Promoter	Putative P_C, LDF 0.66 No binding sites for transcription factors	c(131..157)
		Promoter	Putative P_C, LDF 2.16 Binding site for transcription factor: rpoD17	c(224..257)

		*int*I	Integron integrase, complete, 293 amino-acid-residues, 9.6% acidic residues. (D161G: similar mutations caused slight increase in *int*egration and decrease in excision [298], Y220L: unknown effect on activity)	375..1256
		Promoter	Putative P_C, LDF 2.74 Binding site for transcription factor argR	c(388..416)
		Promoter	Putative P_C, LDF 3.5 Binding sites for transcription factors: tyrR, ihf, arcA, purr and rpoD15	c(487..511)
KD UINF	contig12234, [168], [164], SRX352368	*Int*I gene	Integron integrase, partial: missing part of the C terminus (end of contig), pseudogene: 2 frameshifts	c(1..799)-frameshifts at 408 and 608
		*att*I	Putative primary recombination site ATCCAGCAATTTGGGTTGGGA	815..835
		Promoter	Putative P_C, LDF 5.4 Binding sites for transcription factors: rpoS18, lrp and deoR	842..874
		Promoter	Putative P_{intI}, LDF 4.61 Binding sites for transcription factors: argR, arcA, phoB and rpoD16	c(957..980)
KD UINF	contig17426, [168], [164], SRX352368	*int*I	Integron integrase, partial: missing patch I and part of patch II	1..561
KD UINF	contig20623, [168], [164], SRX352368	*int*I	Integron integrase, partial: missing both termini including box II (small contig)	c(1..636)
KD UINF	contig00958, [168], [164], SRX352368	Gene cassette ORF	Type II deoxyribonuclease	1..276
		*att*C	Cassette-associated recombination site	280..349a
		Promoter	Promoter for TA operon, LDF 5.71 Binding sites for transcriptional factors: gcvA, nagC and phoB	355..383
		Gene cassette ORF	BrnT family toxin of TA system	390..668
		Gene cassette ORF	BrnA family antitoxin of TA system	671..952
		*att*C	Cassette-associated recombination site	947..1070
KD UINF	contig01316, [168], [164], SRX352368	*att*C	Cassette-associated recombination site	c(1399..1470)
		Gene cassette ORF	Diguanylate cyclase	c(1465..2343)
		*att*C	Cassette-associated recombination site	c(2663..2778)
		Gene cassette ORF	VapC toxin of TA system	c(2773..3171)
		Gene cassette ORF	hypothetical protein (probably an antitoxin)	c(3168..3332)

		Promoter	Promoter for the TA operon, LDF 4.82 Binding sites for transcriptional factors: fnr and rpoD17	3377..3404
KD UINF	contig03241, [168], [164], SRX352368	*att*C	Cassette-associated recombination site	c(846..973)
		Gene cassette ORF	Hypothetical protein	c(981..1283)
		*att*C	Cassette-associated recombination site	c(1306..1430)
		Gene cassette ORF	HigA antitoxin of TA system	c(1425.. 1714)-frameshift at 1577
		Gene cassette ORF	RelE/ParE toxin of TA system	c(1714..1992)
		Promoter	Promoter for the antitoxin gene within the toxin gene, LDF 3.11 Binding sites for transcription factors: araC, ihf and rpoD17	c(1893..1921)
		Promoter	Promoter for the TA operon, LDF 0.55 Binding sites for transcription factors: argR2, fnr and deoR	c(2008..2041)
KD UINF	contig04157, [168], [164], SRX352368	Promoter	Promoter for the TA operon, LDF 6.26 Binding sites for transcription factors: soxS, ihf, ihf, modE and arcA	350..380
		Promoter	Promoter for the toxin gene within the antitoxin gene, LDF 3.42 Binding sites for transcription factors: rpoD16, argR and arcA	661..695
		Gene cassette ORF	HicB antitoxin of TA system	502..726
		Gene cassette ORF	HicA toxin of TA system	723..908
		*att*C	Cassette-associated recombination site	903..1027
		Gene cassette ORF	Hypothetical protein	1043..1519
		*att*C	Cassette-associated recombination site	1638..1753
GR	OFEH01000041.1	ORF	Hypothetical protein	18932..19303
		ORF	Alanine glycosylate aminotransferase	19404..20558
		*att*C	Cassette-associated recombination site	19405..19478
		ORF	APC family permease	20555..22069
		ORF	Thymidylate kinase	22226..228443
		ORF	Hypothetical protein	c(22757..24925)
		*att*C	Cassette-associated recombination site	24464..24548
GR	OFEH01000073.1	ORF	Hypothetical protein	24265..25941
		*att*C	Cassette-associated recombination site	24599..24712

		*att*C	Cassette-associated recombination site	24789..24852
		ORF	Peptidase	c(26041..26910)
		ORF	Hypothetical protein	26960..27826
		ORF	Pyridoxal phosphate-dependant aminotransferase	c(28173..29399)
		ORF	Serine hydrolase (betalactamase)	29506..30867
		ORF	Thiaminase II	c(30884..31543)
		ORF	NCS2 family permease	31708..33147
		*att*C	Cassette-associated recombination site	31778..31822
GR *Caldivirga* sp.SpST18	OFEH01000190.1 DSBU01000026.1	ORF	Aldehyde ferrodoxin oxidoreductase	c(20829..22691) 5809..7671
		*att*C	Cassette-associated recombination site	21029..21100 c(7400..7471)
		ORF	Hypothetical protein	23587..24663 c(3848..4919)
		*att*C	Cassette-associated recombination site	23682..23799 c(4707..4824)
GR *Caldivirga* sp. Strain EvPrim.Bin7	OFEH01000320.1 WYEH01000170.1	ORF	ATPase	11494..12633 8824..9963
		*att*C	Cassette-associated recombination site	12089..12148 9419..9478
		ORF	VWA containing CoxE family protein	12617..13963 9947..11293
		*att*C	Cassette-associated recombination site	12894..12941 10224..10271
Caldivirga sp. Strain EvPrim.Bin7	WYEH01000170.1	ORF	aminotransferase	c(11298..12497)
		ORF	Arginine deiminase	12553..13800
		*att*C	Cassette-associated recombination site	13861..13908
		ORF	50S ribosomal protein L14e	13869..14183
GR	OFEH01000355.1	ORF	Hypothetical protein	15079..16755
		*att*C	Cassette-associated recombination site	c(16545..16586)
		*att*C	Cassette-associated recombination site	c(16679..16726)
GR	OFEH01000555.1	ORF	Beta-N-acetylhexosaminidase	293..1960
		*att*C	Cassette-associated recombination site	c(1503..1577)
		ORF	Xylose isomerase	1996..2817
		ORF	MFS transporter	c(2834..3949)
		ORF	peptidase	4033..5028
		ORF	Hypothetical protein	c(5033..5410)
		ORF	adenosylhomocysteinase	c(5533..6852)

		*att*C	Cassette-associated recombination site	c(6930..6976)
		ORF	NBD sugar kinase	c(6936..7919)
GR	OFEH01000586.1	*att*C	Cassette-associated recombination site	c(378..461)
		ORF	Hypothetical protein	452..616
		ORF	No significant similarity	597..791
		ORF	Hypothetical protein	784..993
		ORF	DUF 1156 domain containing protein	990..4091
		ORF	DUF 499 domain containing protein	4088..7246
		*att*C	Cassette-associated recombination site	c(5007..5119)
		*att*C	Putative Cassette-associated recombination site_ AAG instead of AAC at R" and CTT instead of GTT at R'	c(6133..5119)
GR	OFEH01000598.1	ORF	Radical SAM domain containing protein	c(3985..5052)
		*att*C	Cassette-associated recombination site	c(4941..5068)
		ORF	Hypothetical protein	c(5102..5473)
		*att*C	Cassette-associated recombination site	c(5413..5563)
		ORF	DEAD/DEAH box helicase	5521..8301
		*att*C	Cassette-associated recombination site	c(7385..7528)
GR	OFEH01000602.1	ORF	tRNA methyltransferase	2930..3544
		*att*C	Cassette-associated recombination site	c(3220..3263)
		*att*C	Cassette-associated recombination site	c(3539..3671)
		ORF	4-demethylwyosine synthase	3596..4753
		*att*C	Cassette-associated recombination site	c(4041..4118)
GR *Caldivirga* sp. Strain SPST18	OFEH01001317.1 DSBU01000138.1	ORF	ABC transporter	c(3..2027) c(3..1625)
		*att*C	Cassette-associated recombination site	c(417..459)
		*att*C	Cassette-associated recombination site	c(1031..1134) c(15..57)
		*att*C	Cassette-associated recombination site	c(1374..1420) c(972..1018)
		ORF	Hypothetical protein	2171..2848 1768..2445
		*att*C	Cassette-associated recombination site	c(2457..2501) c(2054..2098)
		ORF	glycosidase	c(2862..3911) c(2459..3508)
		*att*C	Cassette-associated recombination site	c(3134..3177) c(2731..2774)
Caldivirga sp. Strain SPST18	DSBU01000138.1	ORF	Hypothetical protein	c(3570..4583)

		ORF	Hypothetical protein	4897..7392
		*att*C	Cassette-associated recombination site	c(5684..5791)
		*att*C	Cassette-associated recombination site	c(6405..6456)
		*att*C	Cassette-associated recombination site	c(6596..6643)
GR	OFEH01001470.1	ORF	Hypothetical protein	c(2854..4722)
		*att*C	Cassette-associated recombination site	c(2955..3017)
		*att*C	Cassette-associated recombination site	c(3174..3233)
		*att*C	Cassette-associated recombination site	c(3246..3287)
		*att*C	Cassette-associated recombination site	c(3322..3365)
		*att*C	Cassette-associated recombination site	c(4026..4067)
		*att*C	Cassette-associated recombination site	c(4569..4625)
Caldivirga sp. Strain MU80	LCTF01000038.1	ORF	ABC transporter substrate-binding protein	10149..12578
		*att*C	Cassette-associated recombination site	12119..12207
		ORF	DMT family transporter	12653..13561
		ORF	Hypothetical protein	c(13558..13707)
		ORF	dienelactone hydrolase family protein	c(13704..14492)
		ORF	Beta-glucosidase	c(14581..17478)
GR *Caldivirga* sp. Strain MU80	OFEH01001501.1 LCTF01000038.1	ORF	Thermopsin family protease-partial in GR contig	1..1227 17693..20521
		*att*C	Cassette-associated recombination site	19293..19340
		*att*C	Cassette-associated recombination site	215..262 19509..19556
		*att*C	Cassette-associated recombination site	597..691 19891..19985
		*att*C	Cassette-associated recombination site	1062..1147 20356..20441
GR	OFEH01001730.1	ORF	Hypothetical protein	c(3..1541)
		*att*C	Cassette-associated recombination site	c(1094..1164)
		*att*C	Cassette-associated recombination site	c(1301..1370)
		ORF	Hypothetical protein	c(1541..1768)
GR	OFEH01002492.1	ORF	tRNA guanine transglycosylase	c(2..592)
		*att*C	Cassette-associated recombination site	217..275
		ORF	Hypothetical protein	c(595..1545)
		ORF	Hypothetical protein	c(1547..2203)
		*att*C	Cassette-associated recombination site	1902..1956

GR	OFEH01002707.1	ORF	Hypothetical protein	c(3..2210)
		*att*C	Cassette-associated recombination site	1614..1661
		*att*C	Cassette-associated recombination site	1936..1984
		*att*C	Cassette-associated recombination site	2227..2275
GR	OFEH01002918.1	ORF	Lichenysin non-ribosomal peptide synthetase	402..674
		*att*C	Cassette-associated recombination site	521..594
		ORF	3- dehydroquinate dehydratase	745..951
		ORF	No significant similarity	965..1078
		ORF	No significant similarity	1075..1191
		*att*C	Cassette-associated recombination site	1200..1246
GR	OFEH01003023.1	*att*C	Cassette-associated recombination site	c(1122..1206)
		ORF	No significant similarity	1024..1566
		ORF	No significant similarity	1572..2066
		*att*C	Cassette-associated recombination site	c(1925..2001)
GR	OFEH01004332.1	ORF	Phosphoribosylformylglycinamidine synthase subunit PurL	c(1..1491)
		*att*C	Cassette-associated recombination site	1215..1316
		*att*C	Cassette-associated recombination site	1320..1387
GR	OFEH01004580.1	ORF	ATP binding protein	c(2..394)
		*att*C	Cassette-associated recombination site	c(262..320)
		ORF	hypothetical protein	c(583..1014)
		ORF	Hypothetical protein	c(953..1441)
		*att*C	Cassette-associated recombination site	c(1174..1230)
		*att*C	Cassette-associated recombination site	c(1231..1295)
GR	OFEH01005341.1	ORF	Hypothetical protein	c(23..655)
		ORF	glycosyltransferase	c(660..1235)
		*att*C	Cassette-associated recombination site	684..750
		*att*C	Cassette-associated recombination site	1007..1057
GR	OFEH01009322.1	ORF	Hypothetical protein	1..402
		*att*C	Cassette-associated recombination site	237..288
		*att*C	Cassette-associated recombination site	305..358
GR	OFEH01009495.1	ORF	Hypothetical protein	c(2..610)

		*att*C	Cassette-associated recombination site	166..237
		*att*C	Cassette-associated recombination site	289..332
		*att*C	Cassette-associated recombination site	430..471

Appendix C: Chapter 6 Supplementary Tables

TableS6.1 Analyzed complete and partial bacterial halophilic genomes

bacterial analyzed genomes	genome size	sequencing status	genome or WGS accession number	plasmids accession numbers if present
Acetohalobium arabaticum DSM 5501	2.4696	complete	NC_014378.1	-
Halothece sp. PCC 7418	4.17917	complete	NC_019779.1	-
Cellulosimicrobium cellulans PSBB019	4.79986	complete	NZ_CP021383.1	-
Desulfohalobium retbaense DSM 5692	2.90957	complete	NC_013223.1	NC_013224.1
Chromohalobacter salexigens DSM 3043	3.66514	complete	NC_007963.1	-
Halorhodospira halophila SL1	2.67845	complete	NC_008789.1	-
Halorhodospira halochloris DSM 1059	2.83456	complete	NZ_AP017372.2	-
Halanaerobium hydrogeniformans	2.61312	complete	NC_014654.1	-
Halanaerobium praevalens DSM 2228	2.30926	complete	NC_017455.1	-
Halobacillus halophilus DSM 2266	4.17177	complete	NC_017668.1	NC_017670.1, NC_017669.1
Halobacteroides halobius DSM 5150	2.64926	complete	NC_019978.1	-
Halomonas elongata DSM 2581	4.06182	complete	NC_014532.2	-
Halomonas titanicae ANRCS81	5.33979	complete	NZ_CP039374.1	-
Halothermothrix orenii H 168	2.57815	complete	NC_011899.1	-
Marinobacter hydrocarbonoclasticus ATCC 49840	3.98677	complete	NC_017067.1	-
Marinobacter hydrocarbonoclasticus VT8	4.77976	complete	NC_008740.1	NC_008738.1, NC_008739.1
Natranaerobius thermophilus JW/NM-WN-LF	3.19145	complete	NC_010718.1	NC_010715.1, NC_010724.1
Nitrosococcus halophilus Nc 4	4.14526	complete	NC_013960.1	NC_013958.1
Nodularia spumigena CCY9414	5.35144	complete	NZ_CP007203.1	-
Nodularia spumigena UHCC 0039	5.38661	complete	NZ_CP020114.1	NZ_CP020115.1
Oceanobacillus iheyensis HTE831	3.63053	complete	NC_004193.1	-
Oceanobacillus iheyensis CHQ24	3.86062	complete	NZ_CP020357.1	-
Salinibacter ruber DSM 13855	3.76289	complete	NC_007677.1	NC_007678.1
Spiribacter salinus M19-40	2.88033	complete	NC_021291.1	-
Ectothiorhodospira haloalkaliphila A	3.46013	partial	NZ_CP007268.1	-
Alteribacillus bidgolensis DSM 25260	4.70318	partial	NJAU01	-
Alteribacillus bidgolensis P4B,CCM 7963,CECT 7998,DSM	4.464	partial	FNDU01	-

25260,IBRC-M 10614,KCTC 13821 genome assembly				
Alteribacillus persepolensis DSM 21632	3.6191	partial	NZ_FNDK01000000	-
Chlorogloea fritschii PCC 6912	7.75174	partial	RSCJ01	-
Chromohalobacter japonicus CJ	3.37628	partial	NZ_CDGZ01000000	-
Chromohalobacter japonicus SMB17	3.76792	partial	MSDQ01	-
Desulfovibrio oxyclinae DSM 11498	3.32458	partial	NZ_AQXE01000000	-
Ectothiorhodospira mobilis DSM 4180	2.62495	partial	NZ_FOUO00000000.1	-
Halarsenatibacter silvermanii SLAS-1	2.71864	partial	NZ_FNGO00000000.1	-
Halobacillus aidingensis CGMCC 1.3703	4.19184	partial	NZ_FNIZ00000000.1	-
Halobacillus alkaliphilus FP5	4.09253	partial	NZ_FOOG00000000.1	-
Halobacillus dabanensis CGMCC 1.3704	4.11984	partial	FOSB01	-
Halobacillus dabanensis HD-02	4.10233	partial	CCDH01	-
Halobacillus trueperi SS1	4.25856	partial	QTLC01	-
Halomonas arcis CGMCC 1.6494	4.14213	partial	NZ_FNII00000000.1	-
Halomonas halodenitrificans DSM 735	3.46409	partial	NZ_JHVH00000000.1	-
Halomonas meridiana ACAM 246	3.84974	partial	FSQY01	-
Halomonas saccharevitans CGMCC 1.6493	3.68129	partial	NZ_FPAQ00000000.1	-
Halomonas subterranea CGMCC 1.6495	3.7342	partial	NZ_FOGS00000000.1	-
Halonatronum saccharophilum DSM 13868	2.88452	partial	NZ_AZYG00000000.1	-
Microcoleus chthonoplastes PCC 7420	8.67904	partial	ABRS01	-
Nocardiopsis halotolerans DSM 44410	6.26393	partial	NZ_ANAX00000000.1	-
Pontibacillus halophilus JSM 076056 = DSM 19796	3.6014	partial	AULI01	-
Saccharomonospora halophila 8	3.68502	partial	AICX01	-
Salinovibrio costicola ATCC 33508 = LMG 11651	4.78167	partial	ASAI01	-
Salinovibrio costicola PRJEB21454	3.32115	partial	FYET01	-
Salisaeta longa DSM 21114	3.39902	partial	NZ_ATTH00000000.1	-
Sediminibacillus halophilus CGMCC 1.6199	4.147699	partial	NZ_FNHF00000000.1	-

Sediminibacillus halophilus NSP9.3	3.986	partial	AWXX01	-
Selenihalanaerobacter shriftii ATCC BAA-73	2.84058	partial	NZ_FUWM00000000.1	-
Spirulina subsalsa PCC 9445	5.3236	partial	NZ_ALVR00000000.1	-
Streptomyces radiopugnans CGMCC 4.3519	6.06712	partial	NZ_FOET00000000.1	-
Thalassobacillus cyri CCM7597	4.30083	partial	NZ_FNQR00000000.1	-

TableS6.2 Analyzed complete and partial archaeal halophilic genomes

archaeal analyzed genomes	genome size (Mb)	sequencing status	genome or WGS accession number	plasmids accession numbers if present (for complete genomes)
Halalkalicoccus jeotgali B3	3.69865	complete	NC_014297.1	NC_014298.1, NC_014299.1, NC_014300.1 , NC_014300.1, NC_014302.1, NC_014303.1
Haloarcula hispanica ATCC 33960	3.89	complete	NC_015948.1 , NC_015943.1	NC_015944.1
Haloarcula marismortui ATCC 43049	4.27464	complete	NC_006396.1 , NC_006397.1	NC_006389.1, NC_006389.1, NC_006389.1, NC_006392.1, NC_006392.1, NC_006393.1 , NC_006394.1, NC_006395.1
Haloarcula sp CBA1115	4.22505	complete	NZ_CP010529.1	, NZ_CP010531.1, NZ_CP010532.1, NZ_CP010533.1, NZ_CP010534.1NZ_CP010530.1
Halobacterium salinarum NRC-1	2.57101	complete	NC_002607.1	NC_001869.1, NC_002608.1
Halobacterium walsbyi C23	3.36799	complete	NC_017459.1	NC_017460.1, NC_017460.1, NC_017457.1
Haloferax gibbonsii ARA6	3.91845	complete	NZ_CP011947.1	NZ_CP011948.1, NZ_CP011949.1, NZ_CP011950.1, NZ_CP011951.1
Haloferax mediterranei ATCC33500	3.90471	complete	NC_017941.2	NC_017942.1, NC_017943.1, NC_017944.1
Haloferax volcanii DS2	4.0129	complete	NC_013967.1	NC_013968.1 , NC_013965.1, NC_013964.1, NC_013966.1
Halogeometricum borinquense DSM 11551	3.94447	complete	NC_014729.1	NC_014735.1, NC_014731.1, NC_014736.1, NC_014732.1, NC_014732.1, NC_014737.1
Halomicrobium mukohataei DSM 12286	3.33235	complete	NC_013202.1	NC_013201.1
Halopiger xanaduensis SH-6(T)	4.35527	complete	CP002839.1	CP002840.1, CP002841.1, CP002842.1
Halorhabdus utahensis DSM 12940	3.116795	complete	CP001687.1	-
Halorubrum lacusprofundi ATCC 49239	3.69258	complete	NC_012029.1, NC_012028.1	NC_012030.1
Haloterrigena turkmenica DSM 5511	5.44078	complete	NC_013743.1	NC_013744.1, NC_013745.1, NC_013746.1, NC_013747.1, NC_013748.1, NC_013749.1
Halovivax ruber XH-70	3.22388	complete	NC_019964.1	-

Methanohalobium evestigatum Z-7303	2.406232	complete	NC_014253.1	NC_014254.1
Methanohalophilus halophilus Z-7982	2.02296	complete	NZ_CP017921.1	-
Methanohalophilus mahii DSM 5219	2.012424	complete	NC_014002.1	-
Methanosalsum zhilinae DSM 4017	2.138444	complete	NC_015676.1	-
Methanosarcina acetivorans C2A	5.75149	complete	AE010299.1	-
Natrialba magadii ATCC 43099	4.44364	complete	NC_013922.1	NC_013923.1 , NC_013924.1, NC_013925.1
Natronobacterium gregoryi SP2	3.78836	complete	NC_019792.1	-
Natronococcus occultus SP4	4.314118	complete	NC_019974.1	NC_019975.1, NC_019976.1
Natronomonas pharaonis DSM 2160	2.7497	complete	NC_007426.1	NC_007427.1, NC_007428.1
Methanohalophilus portucalensis FDF-1T	2.08498	partial	NZ_CP017881.1	-
Haloarcula amylolytica JCM 13557	4.22542	partial	NZ_AOLW00000000.1	-
Haloarcula argentinensis DSM 12282	4.14711	partial	NZ_AOLX00000000.1	-
Haloarcula japonica DSM 6131	4.28036	partial	NZ_AOLY00000000.1	-
Haloarcula vallismortis ATCC 29715	3.90992	partial	NZ_AOLQ00000000.1	-
Halobacterium jilantaiense CGMCC 1.5337	2.95279	partial	NZ_FOJA00000000.1	-
Halobaculum gomorrense DSM 9297	3.20825	partial	NZ_FQWV00000000.1	-
Halococcus morrhuae DSM 1307	2.99156	partial	NZ_AOMC00000000.1	-
Halococcus saccharolyticus DSM 5350	3.4497	partial	NZ_AOMD00000000.1	-
Halococcus sulifodinae DSM 8989	4.19978	partial	NZ_AOME00000000.1	-
Haloferax denitrificans ATCC 35960	3.82597	partial	NZ_AOLP00000000.1	-
Haloferax elongans ATCC BAA-1513	3.95214	partial	NZ_AOLK00000000.1	-
Haloferax mucosum ATCC BAA-1512	3.36898	partial	NZ_AOLN00000000.1	-
Haloferax sulfurifontis ATCC BAA-897	3.81243	partial	NZ_AOLM00000000.1	-
Halorubrum coriense DSM 10284	3.64531	partial	NZ_AOJL00000000.1	-

Halorubrum distributum JCM 10118	3.30613	partial	AOJN01	-
Halorubrum distributum JCM 9100	3.30737	partial	AOJM01	-
Halorubrum distributum E8	2.25364	partial	NHPH01	-
Halorubrum saccharovorum DSM 1137	3.35304	partial	AOJE01	-
Halorubrum sodomense RD 26	3.03055	partial	NZ_FOYN00000000.1	-
Halosimplex carlsbadense 2-9-1	4.69489	partial	NZ_AOIU00000000.1	-
Natronococcus amylolyticus DSM 10524	4.41653	partial	NZ_AOIB00000000.1	-

TableS6.3 Analyzed metagenomic assemblies from different marine, freshwater and hydrothermal vents environments

	Site	**Description**	**Assembly Accession number or reference**	**Total assembled sequence length**	**Number of contigs**
Marine	ADR	North Adriatic Sea, Italy, depth 1m	GCA_900205615.1	24428552	29430
	ARC	Arctic Ocean, station 54, depth 40.3m	GCA_900247125.1	6551393	10186
	PAC	Pacific Ocean, depth 100m	GCA_002896035.2	201472418	193946
	Red10	Red Sea water column Station 192 - depth 10m	GCA_001626065.1	97729439	57007 scaffolds
	Red25	Red Sea water column Station 192 - depth 25m	GCA_001629045.1	57846509	34483 scaffolds
	Red50	Red Sea water column Station 192 - depth 50m	GCA_001629095.1	86416103	47563 scaffolds
	Red100	Red Sea water column Station 192 - depth 100m	GCA_001629115.1	50269729	34015 scaffolds
	Red200	Red Sea water column Station 192 - depth 200m	GCA_001629075.1	45247809	30314 scaffolds
	Red500	Red Sea water column Station 192 - depth 500m	GCA_001629135.1	72981833	44066 scaffolds
	SOCo	Metagenomic co-assembly of South Ocean 3 biosamples: SAMEA2621487, SAMEA2621509, SAMEA2621536, depth 5m	GCA_001757065.1	185494017	19160
	TIB	Trindade and Martin Vaz Islands, Eastern Brazil, depth 5m	GCA_001371195.1	110278656	116750
	WIO	Western Indian Ocean, Fiji islands and Western and Northern Madagascar, depth 5m	GCA_001370375.1	199208958	216738

	CIOI	Central Indian Ocean Islands, depth 5m	GCA_001370295.1	70690895	62491
	WSIS	West and South Indian Shelf, depth 5m	GCA_001370155.1	53135041	47352
	MED	Mediterranean Sea (Tunisian Plateau/Gulf of Sidra & Ionian Sea), depth 5m	GCA_001369555.1	99812943	73799
	ATII 50	Atlantis II 50 m water column, Red Sea	[170], [168]	53647835	78510
	ATII 200	Atlantis II 200 m water column, Red Sea	[170], [168]	49971663	72359
	ATII 700	Atlantis II 700 m water column, Red Sea	[170], [168]	51443487	64636
	ATII 1500	Atlantis II 1500 m water column, Red Sea	[170], [168]	32542975	39190
Marine hydrothermal vents	GB VNT	Guaymas Basin deep-sea hydrothermal vent plume water, Deep Gulf of California	[168], [299], [300], [301]	10092836	12928
	K VNT	Kueishantao shallow-sea hydrothermal vent, Taiwan	[168], [302]	4724790	4235
	LC MM	Loki's Castle deep-sea vent biofilm (microbial mat)	[168], [302]	13324405	11319
Fresh water	RG	River Ganga, Varanasi, India	GCA_004348215.1	18532629	24721
	LL	Lansing Lake, Michigan, USA	GCA_009467185.1	23646022	7443
	MSUL	MSU3 Lake, Michigan, USA	GCA_009467265.1	5647297	1981
	LEN	Lake Erie, Niagara, Canada	GCA_900249105.1	15494770	15973
	SWS	The surface of water catchment in Singapore, WC_Site 4c	GCA_900258585.1, [303]	40691729	46218
	TLB	Taihu Lake water bloom, China	GCA_001515565.1	60186787	46225
	WG	Wintergreen Lake, Michigan, USA	GCA_009469485.1	8769751	1460

TableS6.4 Genetic elements description and position within gene cassette arrays in examined sites

site	Genetic element	Annotation (description)	position
TSL1	Gene cassette ORF	Hypothetical protein	c(9192..9608)
TSL1	*att*C	Integron Finder prediction	c(8999..9068)
TSL1	Gene cassette ORF	Hypothetical protein	c(8709..8987)
TSL1	*att*C	Integron Finder prediction	c(8641..8714)
TSL1	Gene cassette ORF	YoeB-YefM toxin-antitoxin system antitoxin YefM	c(8343..8594)
TSL1	Gene cassette ORF	Txe/YoeB family addiction module toxin	c(8092..8346)
TSL1	*att*C	Integron Finder prediction	c(8025..8092)
TSL1	Gene cassette ORF	Hypothetical protein	c(7726..8007)
TSL1	*att*C	Integron Finder prediction	c(7679..7748)
TSL1	Gene cassette ORF	No significant similarity	c(6893..7672)
TSL1	*att*C	Integron Finder prediction	c(6840..6925)
TSL1	Gene cassette ORF	Serine hydrolase (betalactamase transpeptidase)	c(5676..6734)
TSL1	Gene cassette ORF	Hypothetical protein	c(5183..5524)
TSL1	UHB.F1	5' truncated group IIC intron	3872..5204
TSL1	UHB.F1 ORF	Intron encoded protein (group II reverse transcriptase/maturase), 411 aa	3872..5111
TSL1	Putative internal promoter	LDF score 1.38, -10: CGGTAATCT, -35: TCGAGA no transcription factors binding sites detected	c(4692..4721)
TSL1	Putative internal promoter	LDF score 1.81, -10: GTTTACCAT, -35: CTGACG no transcription factors binding sites detected	c(4238..4267)
TSL1	Putative promoter	LDF score 3.04, -10: TTGTAGTTT, -35: TTGCCA Binding sites for transcription factors soxS and fis	c(3731..3763)
TSL1	Putative promoter for IEP-ORF	LDF score 2.04, -10: CGTTGTAAT, -35: TTGTGT Binding sites for transcription factor rpoD17	3675..3701
TSL1	Putative promoter	LDF score 1.05, -10: AGGTAGAAA, -35: TTTCCG Binding sites for transcription factor rpoD15	c(3399..3426)

TSL1	Putative promoter for IEP-ORF	LDF score 1.0, -10: TCCGATATT, -35: TTGGCG Binding sites for transcription factor rpoD16	3328..3356
TSL1	Gene cassette ORF	Hypothetical protein	c(3178..3603)
TSL1	*att*C	Integron Finder prediction	c(2897..3020)
TSL1	Putative promoter for the TA operon	LDF score 0.72, -10: CGGGAAAAT, -35: GCGCCT no transcription factors binding sites detected	c(2890..2918)
TSL1	Putative promoter for the TA operon	LDF score 1.79, -10: CGTTATGAC, -35: TTTCAA no transcription factors binding sites detected	c(2862..2890)
TSL1	Putative promoter for the toxin ORF	LDF score 3.34, -10: CAGTATATT, -35: TTGAGG Binding sites for transcription factor rpoD16	c(2818..2847)
TSL1	Putative promoter for the toxin ORF	LDF score 4.22, -10:, ATTGAAAAT, -35: TTGATG Binding sites for transcription factors rpoD16, ihf and phoB	c(2787..2817)
TSL1	Gene cassette ORF	Antitoxin	c(2618..2839)
TSL1	Gene cassette ORF	RelE/ParE family toxin	c(2371..2631)
TSL1	*att*C	Integron Finder prediction	c(2300..2367)
TSL2	Gene cassette ORF	Hypothetical protein, partial	c(9627..9770)
TSL2	*att*C	Integron Finder prediction	c(9589..9658)
TSL2	Gene cassette ORF	Hypothetical protein	c(9053..9580)
TSL2	*att*C	Integron Finder prediction	c(8978..9047)
TSL2	Gene cassette ORF	Putative GNAT N-acetyltransferase (30% identity)	c(8588..8971)
TSL2	*att*C	Integron Finder prediction	c(8427..8512)
TSL2	Gene cassette ORF	HNH endonuclease	c(7930..8415)
TSL2	*att*C	Integron Finder prediction	c(7890..7959)
TSL2	Gene cassette ORF	PH domain-containing protein	c(7420..7884)
TSL2	*att*C and stem loop	Integron Finder prediction	c(7295..7364)
TSL2	Putative internal IEP-ORF promoter	LDF score 3.92, -10: TGATATAAT, -35: CTGATT Binding sites for transcription factor rpoD16	c(7246..7271)
TSL2	UHB.I2	IIB1 group II Intron	c(5096.. 7296)
TSL2	Putative internal IEP-ORF promoter	LDF score 1.04, -10: TGATAAACC, -35: TTTCTT Binding sites for transcription factor crp	c(6745..6771)
TSL2	IEP	Chloroplast-like 1(CL1) IEP, 500 aa	c(5223..6725)
TSL2	Putative internal promoter	LDF score 2.69, -10: GCGTAGAAT, -35: CTACCG	c(6397..6423)

		Binding sites for transcription factor narL	
TSL2	Putative internal promoter	LDF score 1.27, -10: TGTTAACGT, -35: GTCCCG Binding sites for transcription factor rpoD16	c(5932..5960)
TSL2	Putative internal promoter	LDF score 1.08, -10: GTCTACTAT, -35: TCGAAA no transcription factors binding sites detected	c(5406..5438)
TSL2	Gene cassette ORF	Hypothetical protein	c(4525..5097)
TSL2	Putative promoter for TA operon	LDF score 2.17, -10: CGGAATATT, -35: GTGATG Binding sites for transcription factors crp and rpoD19	c(4709..4737)
TSL2	*att*C	Integron Finder prediction	c(4463..4591)
TSL2	Putative promoter for the toxin gene	LDF score 0.97, -10: TACTGTAAT, -35: ATGCTA no transcription factors binding sites detected	c(4405..4435)
TSL2	ORF downstream the gene cassette array	ParD-like antitoxin	c(4200..4418)
TSL2	ORF downstream the gene cassette array	RelE/ParE family toxin	c(3923..4213)
Halorhodospira halochloris DSM 1059	IS*200/605* element	70% coverage to IS*Hahl*1 with 98% identity. A deletion in the middle, thus missing 5'end of both *tnp*A and *tnp*B genes. Y1 transposase domain in *tnp*A is complete	449664..450944
Halorhodospira halochloris DSM 1059	IS*200/605* element	80% coverage to IS*Hahl*1 with 98% identity. A deletion in the 5' end, thus having a truncated *tnp*A gene and a complete *tnp*B gene.	460538..461995
Halorhodospira halochloris DSM 1059	IS*200/605* element	57% identity to IS*Hahl*1. Frameshifts in both *tnp*A and *tnp*B genes most probably rendering them inactive	c(690007..691779)
Halorhodospira halochloris DSM 1059	Partial IS*200/605* element	17% coverage to IS*Hahl*1 with 95% identity. No transposase genes detected	767946..769352
Halorhodospira halochloris DSM 1059	Putative CALIN promoter	LDF score 2.52, -10: CCTTATAAA, -35: CTGCTT Binding sites for transcription factors metR, rpoD17, rpoD16	1184472..1184504
Halorhodospira halochloris DSM 1059	Putative CALIN promoter	LDF score 1.22, -10: CAGTATCCT, -35: CTGCGA Binding sites for transcription factor rpoD16	1184783..1184809
Halorhodospira halochloris DSM 1059	Gene cassette ORF	Hypothetical protein	1184958..1185458
Halorhodospira halochloris DSM 1059	*att*C	Integron Finder prediction	1185442..1185516
Halorhodospira halochloris DSM 1059	Putative promoter for TA operon	LDF score 3.41, -10: GCATACAAT, -35: TTGACC Binding sites for transcription factor rpoD18	1185541..1185569
Halorhodospira halochloris DSM 1059	Gene cassette ORF	BrnT family toxin	1185576..1185863
Halorhodospira halochloris DSM 1059	Gene cassette ORF	BrnA family antitoxin	1185860..1186075

Halorhodospira halochloris DSM 1059	*att*C	Integron Finder prediction	1186078..1186137
Halorhodospira halochloris DSM 1059	Gene cassette ORF	NgoFVII family restriction endonuclease	1186155..1186583
Halorhodospira halochloris DSM 1059	Gene cassette ORF	Hypothetical protein	1186587..1187339
Halorhodospira halochloris DSM 1059	*att*C-like	CAC and GTG instead of the conserved triad (AAC and GTT) in the R box, no unpaired spacer between R & L boxes,	1187352..1187425
Halorhodospira halochloris DSM 1059	Putative promoter for IEP ORF	LDF score 0.9, -10: GGTTAAGCG, -35: GTGAGG Binding sites for transcription factors rpoD18 & ihf	1187532..1187562
Halorhodospira halochloris DSM 1059	H.ha.F1	5' truncated IIB group II Intron	1187659..1188795
Halorhodospira halochloris DSM 1059	IEP	Bacterial class E IEP, 342 aa, internal deletion causing a frameshift at 133 and an internal stop at 300	1187659..1188685
Halorhodospira halochloris DSM 1059	Putative internal promoter	LDF score 1.05, -10: TCGTAGACT, -35: TTTATC no transcription factors binding sites detected	1188443..1188468
Halorhodospira halochloris DSM 1059	Putative internal promoter	LDF score 0.91, -10: CGGTATGCC, -35: TTGTCG no transcription factors binding sites detected	1188106..1188139
Halorhodospira halochloris DSM 1059	Putative promoter for the TA system	LDF score 2.38, -10: CGTTATTAA, -35: TTGCCA no transcription factors binding sites detected	1188819..1188846
Halorhodospira halochloris DSM 1059	Gene cassette ORF	RelE/ParE family toxin	1188869..1189147
Halorhodospira halochloris DSM 1059	Gene cassette ORF	HigA family antitoxin	1189158..118472
Halorhodospira halochloris DSM 1059	*att*C	Integron Finder prediction	1189467..1189526
Halorhodospira halochloris DSM 1059	Gene cassette ORF	DUF1643 domain-containing protein	1189540..1190004
Halorhodospira halochloris DSM 1059	*att*C	Integron Finder prediction	1190007..1190078
Halorhodospira halochloris DSM 1059	Gene cassette ORF	DUF3800 domain-containing protein	1190085..1190897
Halorhodospira halochloris DSM 1059	*att*C	Integron Finder prediction	1190899..1190970
Halorhodospira halochloris DSM 1059	Gene cassette ORF	SIR2 family protein	1191028..1192185
Halorhodospira halochloris DSM 1059	Gene cassette ORF	DUF4160 domain-containing protein	1192152..1192376
Halorhodospira halochloris DSM 1059	Gene cassette ORF	DUF2442 domain-containing protein	1192479..1192751
Halorhodospira halochloris DSM 1059	Gene cassette ORF	HNH endonuclease	1192791..1193168

Halorhodospira halochloris DSM 1059	Putative promoter for TA operon within upstream ORF	LDF score 2.9, -10: GAGTATAAG, -35: GTCATA Binding sites for transcription factors rpoD16, rpoD15 & purR	1193084..1193116
Halorhodospira halochloris DSM 1059	Gene cassette ORF	antitoxin	1193223..1193453
Halorhodospira halochloris DSM 1059	Gene cassette ORF	RelE/ParE family toxin	1193453..1193749
Halorhodospira halochloris DSM 1059	*att*C	Integron Finder prediction	1193745..1193828
Halorhodospira halochloris DSM 1059	Gene cassette ORF	DUF4160 domain-containing protein	1193885..1194151
Halorhodospira halochloris DSM 1059	Gene cassette ORF	DUF2442 domain-containing protein	1194160..1194408
Halorhodospira halochloris DSM 1059	*att*C	Integron Finder prediction	1194472..1194535
Halorhodospira halochloris DSM 1059	Putative promoter for TA operon within upstream gene cassette ORF	LDF score 0.24, -10: TCGTACTTT, -35 TTTTTA Binding sites for transcription factor rpoD16	1194272..1194303
Halorhodospira halochloris DSM 1059	Gene cassette ORF	BrnT family toxin	1194537..1194854
Halorhodospira halochloris DSM 1059	Gene cassette ORF	BrnA family antitoxin	1194851..1195123
Halorhodospira halochloris DSM 1059	Putative promoter for TA operon	LDF score 3.13, -10: CGGCATTTT, -35: TTGACA Binding sites for transcription factors rpoD16 & rpoD17	1195221..1195254
Halorhodospira halochloris DSM 1059	Gene cassette ORF	BrnT family toxin	1195502..1195870
Halorhodospira halochloris DSM 1059	Putative promoter for antitoxin gene within toxin ORF	LDF score 1.78, -10: ATGCATACT, -35: TTGGCT no transcription factors binding sites detected	1195568..1195596
Halorhodospira halochloris DSM 1059	Gene cassette ORF	BrnA family antitoxin	1195863..1196114
Halorhodospira halochloris DSM 1059	*att*C	Predicted by bs folding using MFOLD	1196117..1196210
Halorhodospira halochloris DSM 1059	Putative promoter within detected attC site	LDF score 1.16, -10: GCTTAGCAT, -35: TTGGTT no transcription factors binding sites detected	1196172..1196199
Halorhodospira halochloris DSM 1059	H.ha.F2	5' truncated IIB group II Intron	1196335..1197101
Halorhodospira halochloris DSM 1059	IEP	Bacterial class E IEP, 210 aa, 5' deletion	1196335..1196964
Halorhodospira halochloris DSM 1059	Putative internal promoter	LDF score 1.20, -10: CGGTATGCC, -35: TTGCCG no transcription factors binding sites detected	1196390..1196418
Halorhodospira halochloris DSM 1059	Putative internal promoter	LDF score 1.05, -10: TCGTAGACT, -35: TTTATC no transcription factors binding sites detected	1196722..1196747

Halorhodospira halochloris DSM 1059	Gene cassette ORF-frame shift	HicA family toxin-frame-shift due to 1 nucleotide deletion at 1197266 position	1197111.. 1197370
Halorhodospira halochloris DSM 1059	Putative promoter for antitoxin gene within upstream toxin gene and 3' end of the intron	LDF score 0.89, -10: TGAGAAAAT, -35: TTACAA no transcription factors binding sites detected	1197092..1197120
Halorhodospira halochloris DSM 1059	Putative promoter for antitoxin gene within upstream toxin gene	LDF score 2.37, -10: GGCTAGGAT, -35: TTGTCA no transcription factors binding sites detected	1197185..1197210
Halorhodospira halochloris DSM 1059	Gene cassette ORF	HicB family antitoxin	1197360..1197572
Halorhodospira halochloris DSM 1059	*att*C	Integron Finder prediction	1197567..1197626
Halorhodospira halochloris DSM 1059	IS*Hahl*1	IS200/605 family (IS605 group) insertion sequence	1197652..1199464
Halorhodospira halochloris DSM 1059	IS*Hahl*1-LE	IS left end forming hairpin structure	1197652..1197730
Halorhodospira halochloris DSM 1059	TnpA	ISHahl1 TnpA (transposase)	c(1197731..1198045)
Halorhodospira halochloris DSM 1059	TnpB	ISHahl1 TnpB (accessory protein)	1198170..1199444
Halorhodospira halochloris DSM 1059	IS*Hahl*1-RE	IS right end forming hairpin structure	1199445..1199464
Halorhodospira halochloris DSM 1059	IS*Hahl*1 isoform	IS200/605 family (IS605 group) insertion sequence	1269635..1271447
Halorhodospira halochloris DSM 1059	IS*200/605* element	57% coverage to IS*Hahl*1 with 97% identity. A deletion in the middle, thus missing N-termini of both *tnp*A and *tnp*B genes.	c(1472999..1474045)
Halorhodospira halochloris DSM 1059	*Ori*C	Predicted OriC by γBORIS	2787842..2789091

Appendix D: Chapter 6 Supplementary Figures

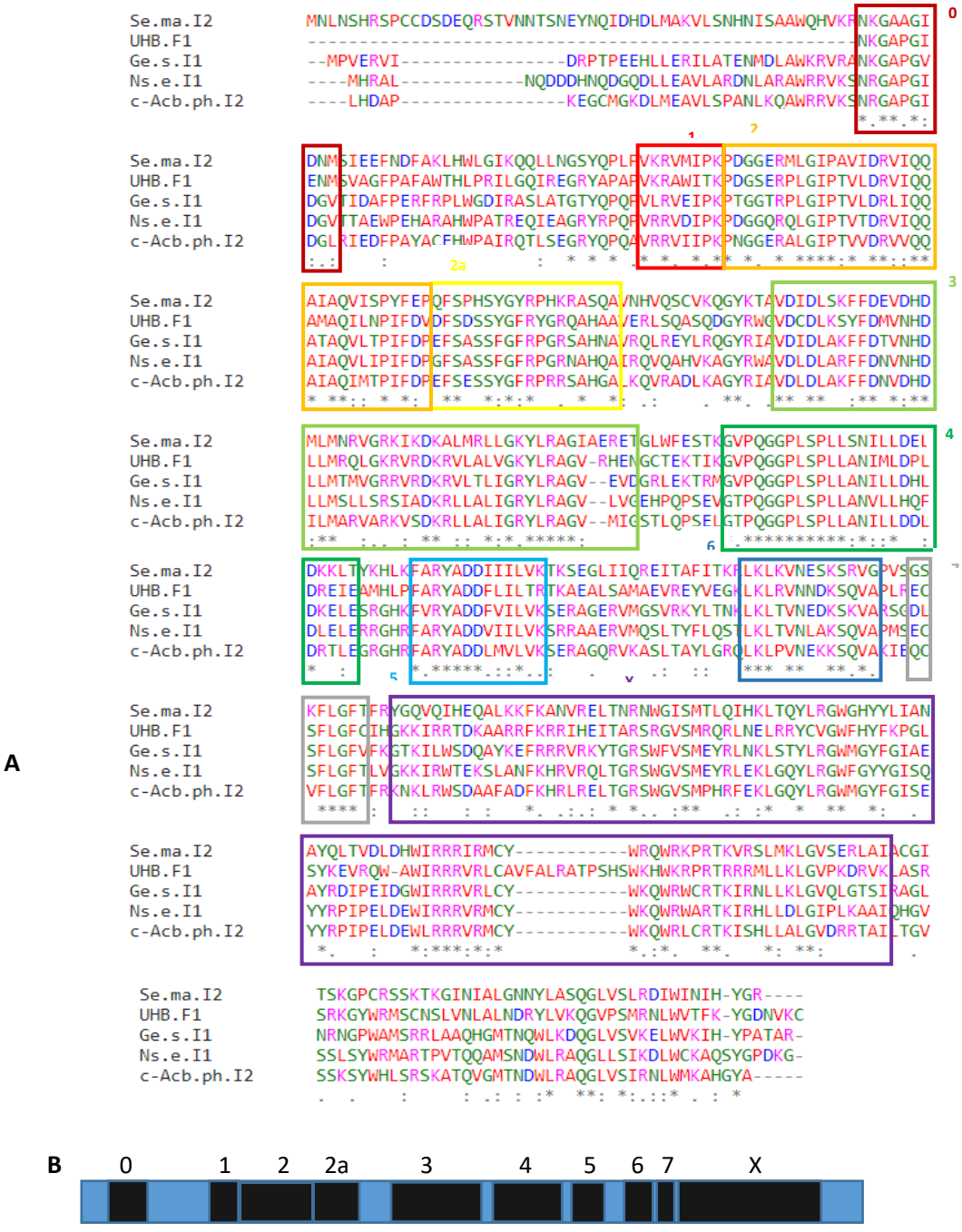


Fig.S6.1 A: Multiple sequence alignment of UHB.F1 with closely related IEPs showing RT domains (RT0-7) and X domain and the highly conserved YADD motif within RT5 domain. B. Schematic representation of UHB.F1 IEP showing relative positions of its RT domains (0-7) and X domain.

>IEP- UHB.F1-3872..5111

NKGAPGIENMSVAGFPAFAWTHLPRILGQIREGRYAPAPVKRAWITKPDGSERPLGIPTVLDRVIQ
QAMAQILNPIFDVDFSDSSYGFRYGRQAHAAVERLSQASQDGYRWGVDCDLKSYFDMVNHDLL
MRQLGKRVRDKRVLALVGKYLRAGVRHENGCTEKTIKGVPQGGPLSPLLANIMLDPLDREIEA
MHLPFARYADDFLILTRTKAEALSAMAEVREYVEGKLKLRVNNDKSQVAPLRECSFLGFCIHGK
KIRRTDKAARRFKRRIHEITARSRGVSMRQRLNELRRYCVGWFHYFKPGLSYKEVRQWA/WIRR
RVRLC/AVFALRATPSHSWKHWKRPRTRRRMLLKLGVPKDRVKLASRSRKGYWRMSCNSLVNL
ALNDRYLVKQGVPSMRNLWVTFKYGDNVKC*

>IEP-UHB.I2-c(5223..6725)

MIPDKGSALRNMPRNWRSLDWDAAERHVKRLQVRIAKAVEEKKWGKVKALQWTLTHSFYAK
ALAVRRVTRNKGARTPGIDKARWRTDGRKLAAVLQLKRHGYRAKALRRIYILKKNGKKRPLSIP
TMNDRAMQALYALALIPVAEALADPNSYGFREGRCCQDALEQCFVILARRVSPGWILEADIKGC
FDNISHEWLMNHIPLDKSILRQWLEVGYIEEGEWFRSEAGTPQGGIVSPILANLTLNGLEKAIKAS
VPSTETGVNVVRYADDFIVTARSPERLTETIRPVIERFLAERGLSLSEEKTKITSIDEGFDFLGQNA
RKYEGKLLIKPSKTSTQGLLDKVRLIIDAHKGKSAERLIKVLNPVIRGWANYHRHSVCAQTFYYI
DYVISGALFRWIRKRNQNKSKSWIVWKHFRSPLDKSGTFCAKSKNKKGQTVYYHLQKALNIPRA
LHRKVIGKAHPYQPEKAEYFAKRQLKRYRTKGRMSQPMQWIQAHLGFQP*

> IEP-H.ha.F1-1187659.. 1188685

VPEGNTKHPQWRGCGGLAGSSGRGMQGEIRRRTREAPKGSCGGEGRQGPTAIETRRGNLETKR
YRTRRVRRCYIPKEDGGERPLGIPAVEDRLLQAACARILTAIYEADFLDGSYGYRPGKSAKDAVA
DLGST/LHYALDLWFEQVVKPRCRGQALLVRYADDYVCAFQFQEDAQRFYRAVPRRLGRFGLQ
VAPEKTRLMRFSRFHPGLRRRFGFLGFELNWSRDRRGELRVMKRTARKKLQAAKRRLKGWIRA
NRHLPGRVFIQELNRRLVGHYNYFGLRSNEQGLGSYHIFAIRCAFK*LNRRGGKRSSFNWAQYIE
ALRKLGVAQPRITERQRAHGVFA*

> IEP-H.ha.F2 -1196335..1196964

YLHYALDLWFERVVKPRCRGQALLVRYADDYVCAFQFQEDAQRFYRAVPRRLGRFGLQVAPE
KTRLMRFSRFHPGLRRRFGFLGFELYWSRDRRGELRVMKRTVRKKLQAAKRRLKGWIRANRHL
PGRVFIQELNRRLVGHYNYFGLRSNEQGLESYYIFATRCAFKWLNRRGGKRSSFNWAQYIAALR
KLGVEQPRITERQRAHAVFA*

Fig.S6.2 Amino acid sequences of identified IEPs showing stop codons as "*", frameshifts as "/" and insertions underlined. Positions within contigs or genome are indicated as well.

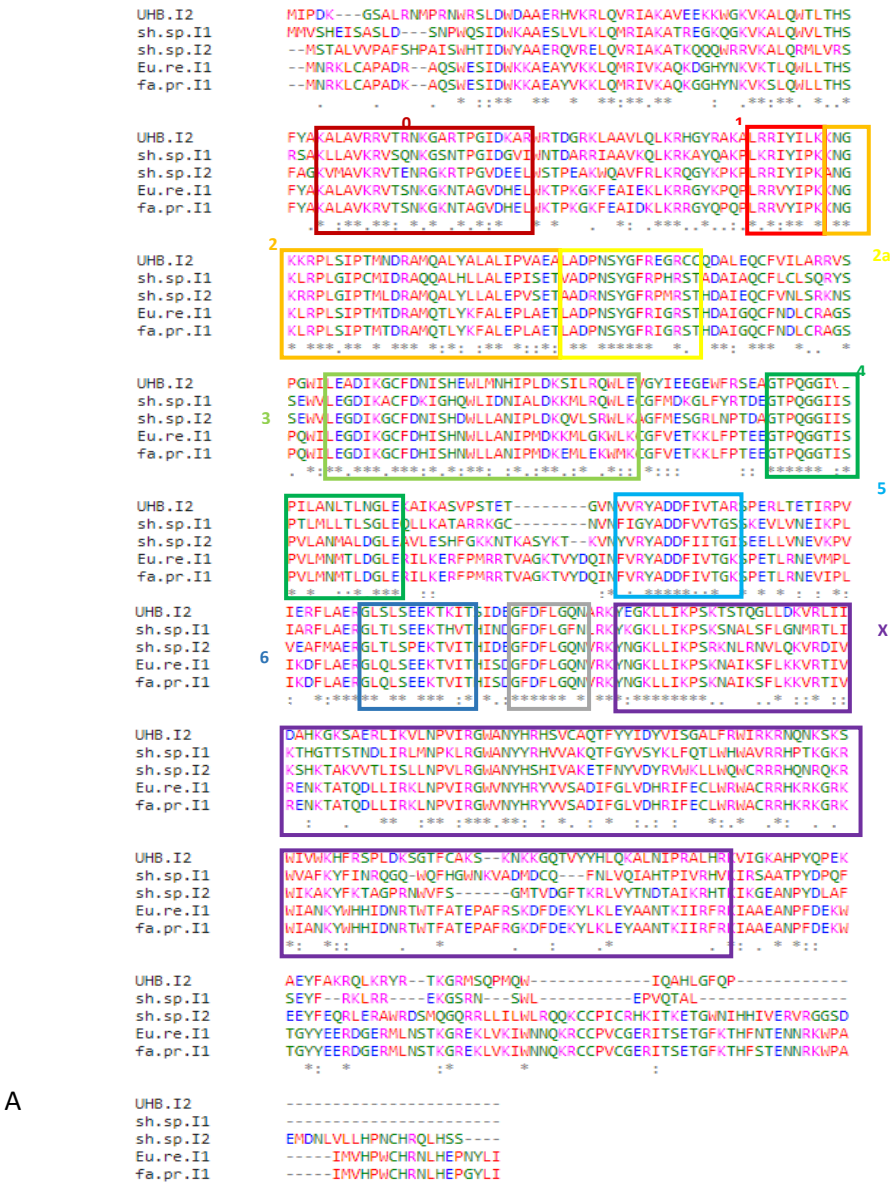


Fig.S6.3 A: Multiple sequence alignment of UHB.I2 with closely related IEPs showing RT domains (RT0-7) and X domain and the highly conserved YADD motif within RT5 domain. B. Schematic representation of UHB.I2 IEP Schematic representation of UHB.I1 IEP showing relative positions of its RT domains (0-7) and X domain

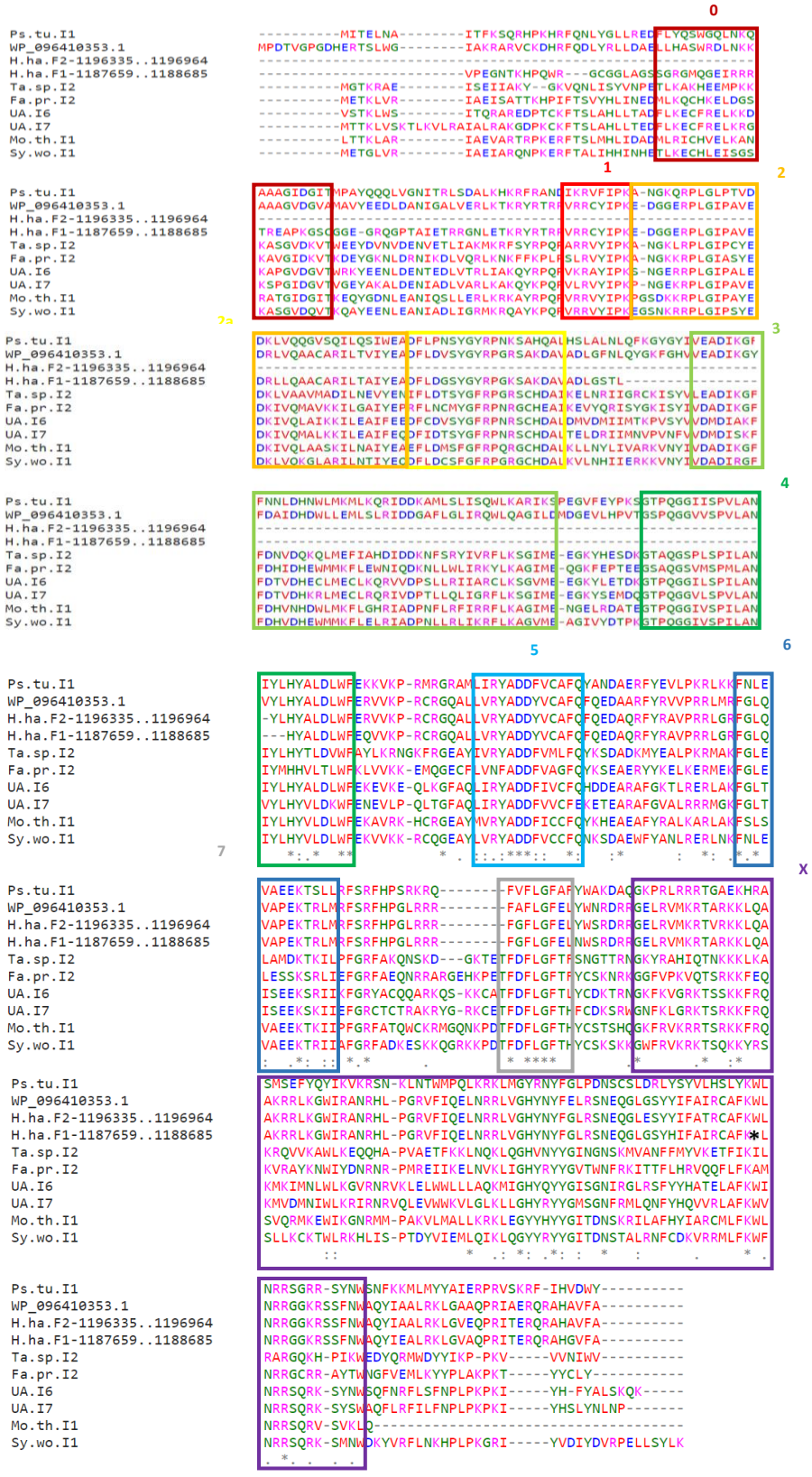


Fig.S6.4 Multiple sequence alignment of H.ha.F1 and H.ha.F2 IEP with closely related IEPs from bacterial class E showing missing RT1, 2,3 and part of RT4 in both ORFs and missed RT0 in H.ha.F2 as well. A internal stop codon in H.ha.F1 is shown as an asterisk. The highly conserved YADD motif is within RT5 domain.

UHB.F1-fragmented_intron-3872..5204

AACAAAGGAGCACCGGGGATCGAGAACATGTCGGTCGCCGGGTTTCCGGCGTTTGCATGGACGCACCTGCCAAG
AATATTGGGGCAGATTCGGGAGGGGCGCTATGCCCCTGCTCCGGTTAAAAGAGCTTGGATCACCAAACCGGACG
GAAGCGAACGCCCCCTGGGCATACCGACCGTTCTGGACCGGGTGATCCAGCAAGCCATGGCTCAAATCCTCAATC
CCATCTTTGATGTGGATTTCAGCGACAGCAGCTATGGTTTCAGATACGGACGCCAGGCTCACGCCGCCGTCGAGC
GGTTGAGTCAGGCGAGTCAGGACGGTTACCGTTGGGGAGTGGACTGCGACCTGAAGTCCTACTTCGACATGGTA
AACCACGACCTGTTGATGCGTCAGCTGGGTAAGCGAGTCCGGGACAAACGCGTTCTCGCCCTGGTCGGCAAATAC
CTGCGTGCAGGTGTCAGGCATGAGAACGGTTGCACGGAGAAAACGATCAAAGGCGTTCCTCAGGGAGGCCCGTT
GTCTCCGCTGCTTGCCAACATCATGCTCGACCCGCTCGACCGGGAAATCGAGGCGATGCACCTGCCGTTTGCCCGC
TACGCGGATGATTTCCTCATCCTCACCCGCACCAAAGCCGAGGCGCTGAGCGCCATGGCCGAAGTCCGGGAGTAT
GTGGAGGGAAAACTGAAGTTGCGGGTTAATAACGACAAAAGTCAGGTTGCTCCCTTGAGGGAATGCAGCTTCCTT
GGGTTCTGTATCCACGGGAAGAAAATCCGGCGGACCGATAAGGCAGCCCGGAGATTCAAACGCCGGATACATGA
GATTACCGCCCGCAGTCGGGGCGTCTCGATGAGGCAACGCCTCAACGAACTCCGGCGTTATTGCGTGGGGTGGTT
TCATTACTTCAAGCCGGGCCTTTCCTATAAGGAAGTCCGACAGTGGGCCTGTGGATACGCAGGCGCGTGCGCCTG
TGCTAGCCGTCTTCGCCCTGCGGGCTACGCCGAGCCACAGCTGGAAACACTGGAAGCGGCCGCGAACGCGGAGA
CGAATGCTCCTGAAACTCGGCGTCCCTAAAGACCGGGTGAAGCTGGCATCCCGCTCCCGCAAGGGCTATTGGCGA
ATGTCGTGCAACAGTCTGGTCAACCTGGCCCTCAATGATCGTTATCTGGTAAAACAAGGGGTACCGTCGATGCGG
AACCTCTGGGTGACCTTCAAATATGGAGATAACGTCAAGTGCTAGTCTCCGGTCACTGATTCTTCGGAACCGCCGT
GATACGGACCCGTATGTCCGGTGGTGTGGGGGCCGGGGAGTTAACGCTCCCGGCTACCCGAT

UHB.I2- intron-c(5096..7296)

TTGCGACATGATGTTACGCAAGATACTGATTAACAACAACTATGATATAATGGAAAACTGATTAATCCTCGCATTCC
ACTGCGAGTTCGCCAAGGGGCGTGCGCCCTGTCGGGCGGTTTCGGTTCGCCGGAGCTACTGGCAGGGGGTGCGA
AGCCTTCCTGGAATGTAATGATCCCTGGAGCCTCTGCCGGGGGGGATTGCTAGCGACGGGCGGTATGGTGAGATTA
GACGAAAGGTCGAAAGCGCCGTAAAGGAAACGAGCACAAAAGAGGCTAATGAGCTCAGCCAAAAGGCAAGTAG
TCAGGTTTCGGGTGTAGAACGTGCCCGGACGTGGATGTTGAACTGCCGGCGTAACAGACCTCACCTAACCCGCTC
TCGTATCTTGTGTGGAACATGGGAACCCGGATCTTCTTCCCTCCGCGGGAAGGCGTCACCGCAAGGCGCGTATCG
GAGATCTGGATTGAGAGGTTCGAAAAAGCCAATGCCCGCCTGCAATGGGTGCGGATATGCCCACGTCGAACTGGT
GTTTCTTTGTGAGAGAACCTGATAAACCGATGAATCGCAAAAAGCAGATGATCCCCGACAAAGGGAGTGCATTGC
GAAACATGCCACGAAACTGGCGCTCCCTTGACTGGGACGCCGCCGAACGGCACGTTAAACGGCTCCAGGTGCGTA
TCGCAAAGGCAGTTGAAGAAAAGAAATGGGGCAAGGTGAAAGCCTTGCAATGGACGCTGACCCACTCCTTTTACG
CCAAAGCTTTGGCCGTAAGGAGAGTCACGCGCAACAAGGGAGCTCGCACGCCCGGCATCGACAAAGCCCGCTGG
AGAACCGACGGACGAAAACTCGCTGCCGTGCTCCAGCTCAAACGCCACGGCTACCGAGCCAAGGCGTTGCGTAGA
ATCTATATCCTAAAGAAGAATGGCAAGAAACGTCCCCTGAGTATCCCGACAATGAACGACCGGGCAATGCAGGCG
CTTTACGCGCTTGCGCTGATACCGGTAGCCGAAGCACTGGCCGACCCGAACTCCTACGGATTTCGCGAAGGACGC
TGCTGTCAGGACGCTCTCGAACAATGCTTCGTCATCCTGGCCAGACGGGTCTCCCCCGGATGGATACTGGAGGCG
GACATCAAAGGCTGTTTCGACAACATCAGCCACGAATGGCTGATGAACCATATTCCGCTGGACAAAAGCATTCTGC
GTCAATGGCTGGAAGTTGGTTACATAGAGGAAGGAGAATGGTTCCGGTCGGAAGCGGGAACTCCGCAAGGCGG

Fig.S6.5 Identified introns' DNA sequences with their positions within their contigs (TSL1 and TSL2) or genome (*H. halochloris*). Domains are shown in different colors: DI, DII, DIII, DIV, DV, DVI, ORF underlined. Putative promoters are either underlined with a zigzagged line (same orientation) or with a dotted line (opposite orientation). Intron boundaries are colored in cyan.

AATCGTCTCGCCAATCCTCGCCAATCTCACACTCAACGGACTCGAAAAAGCCATCAAGGCATCGGTCCCGAGCACA
GAGACTGGTGTTAACGTAGTTCGGTATGCCGACGACTTCATTGTCACGGCAAGGTCGCCGGAAAGACTGACGGAG
ACGATTCGACCCGTAATCGAGCGATTTCTCGCCGAACGCGGGTTGAGTCTTTCCGAGGAAAAGACGAAGATCACG
TCCATTGACGAAGGCTTCGATTTCCTCGGTCAAAACGCCCGGAAGTACGAAGGAAAGCTGTTGATCAAACCATCG
AAAACCTCGACTCAGGGACTCCTGGACAAGGTTCGGTTGATCATCGACGCCCACAAAGGCAAATCAGCCGAAAGA
CTGATCAAGGTACTAAACCCGGTCATCCGTGGCTGGGCCAACTACCACCGCCACAGCGTGTGCGCGCAGACCTTCT
ATTACATCGACTATGTGATCAGCGGAGCCTTGTTCCGGTGGATACGCAAAAGGAACCAGAATAAATCGAAAAGTT
GGATTGTATGGAAACACTTCCGCAGTCCCCTCGACAAATCCGGAACCTTCTGCGCGAAATCGAAAAACAAGAAAG
GCCAGACCGTCTACTATCACCTGCAAAAGGCGCTCAACATACCGAGAGCCCTGCATCGAAAGGTAATCGGGAAAG
CCCACCCCTACCAACCCGAAAAGGCCGAGTATTTCGCCAAGCGCCAGCTCAAACGTTACCGCACCAAGGGAAGAA
TGAGCCAGCCGATGCAGTGGATACAAGCCCACCTCGGATTCCAACCATGAAAAGAACAACTGCCGGATCCGCCTC
TACTGAGCGGATTCTAGAAATGCTTGAGCCGTGTGAAGGGAAACTTTCACGCACGGTTCTTAGGGAGAACGGGG
GCCGCAAGGCCCCCTGACCACCCGGTA

H.ha.F1- fragmented _intron-1187659..1188795

GTGCCCGAGGGCAACACGAAGCACCCGCAATGGCGAGGGTGTGGAGGTCTGGCGGGGTCATCAGGCCGTGGCA
TGCAGGGAGAGATACGTCGGAGAACTCGGGAAGCCCCGAAGGGCTCCTGTGGTGGTGAAGGCCGGCAAGGGCC
TACGGCTATTGAGACACGAAGGGGAAACCTGGAGACGAAGCGCTACCGGACCCGTCGGGTCCGGCGTTGCTACA
TCCCCAAGGAGGATGGCGGCGAGCGTCCATTGGGGATACCGGCGGTGGAGGACAGGCTGTTGCAAGCGGCCTGT
GCTCGGATACTGACCGCCATCTACGAGGCGGACTTTCTGGACGGGAGCTACGGCTACCGGCCAGGGAAGAGCGC
TAAGGACGCGGTGGCTGATCTGGGTTCAACCTCTGCACTATGCGCTGGACCTCTGGTTCGAGCAGGTGGTGAAGC
CACGTTGTCGAGGACAGGCGCTGCTGGTTCGGTATGCCGATGACTATGTCTGCGCGTTTCAGTTTCAGGAGGATG
CCCAGCGCTTCTATCGTGCAGTGCCGCGCCGGCTGGGTCGGTTTGGGCTGCAGGTGGCGCCGGAGAAGACACGG
CTGATGCGATTCAGCCGGTTCCATCCGGGGTTGCGGCGACGATTTGGCTTTCTCGGCTTCGAGTTGAACTGGAGCC
GGGATCGACGGGGCGAGCTGCGGGTGATGAAGCGCACGGCCCGCAAGAAACTGCAAGCAGCCAAGCGACGGTT
GAAGGGCTGGATACGGGCCAACCGGCACCTGCCGGGGCGCGTGTTTATCCAGGAGCTGAATCGTAGACTGGTAG
GTCATTACAACTACTTCGGGCTCCGCAGCAATGAGCAGGGTCTAGGGAGCTACCACATCTTCGCCATCCGGTGCGC
CTTCAAGTAGCTGAACCGGCGAGGCGGCAAGCGCAGTAGTTTCAACTGGGCGCAATACATTGAGGCCTTGCGGA
AGCTGGGAGTGGCACAGCCGCGGATTACGGAGAGGCAACGAGCGCATGGGGTCTTTGCATAAGGGCACGCCCG
GCGCGAAGGCGAGTACAACCGAGGAACCGGATGCGGGAAAACCGCACGTCCGGGTCTGTGCGGGGGGGGGCG
CCCGGCAACGGGCGTTCCTACCGTGAG

H.ha.F2-fragmented_intron-1196335..1197101

TATCTGCACTACGCGCTGGACCTTTGGTTCGAGCGGGTGGTGAAGCCACGTTGCCGAGGACAGGCGCTGCTGGTT
CGGTATGCCGATGACTATGTCTGCGCGTTTCAGTTTCAGGAGGATGCCCAGCGCTTCTATCGTGCAGTGCCGCGCC
GGCTGGGTCGGTTTGGGCTGCAGGTGGCGCCGGAGAAGACACGGCTGATGCGATTCAGCCGGTTCCATCCGGGG
TTGCGGCGACGATTTGGCTTTCTCGGCTTCGAGTTGTACTGGAGCCGGGATCGGCGGGGCGAGCTGCGGGTGAT
GAAGCGTACGGTCCGCAAGAAACTGCAAGCAGCCAAGCGGCGGTTGAAGGGCTGGATACGGGCCAACCGGCAC
CTGCCGGGGCGCGTGTTTATCCAGGAGCTGAATCGTAGACTGGTAGGTCATTACAACTACTTCGGGCTCCGCAGC
AATGAGCAGGGTCTAGAGAGCTACTACATCTTCGCCACCCGGTGCGCCTTCAAGTGGCTGAACCGCCGAGGTGGC
AAGCGCAGTAGTTTCAACTGGGCGCAATACATTGCGGCGTTGAGGAAGCTGGGAGTGGAGCAGCCGCGGATTAC
GGAGAGGCAACGAGCGCATGCGGTCTTTGCATAAGGGCACGCCCGTCGCGAAGGCGAGTACAACCGAGGAACC
GGATGCGGGAAAACTGCACGTCCGGGTCTGTGCGGGGGCGGCCGGCAACGGGCGTTCCTACCGTGAGATGCAG
GCCGGACACCGAGGAT

Fig.S6.5. Continued

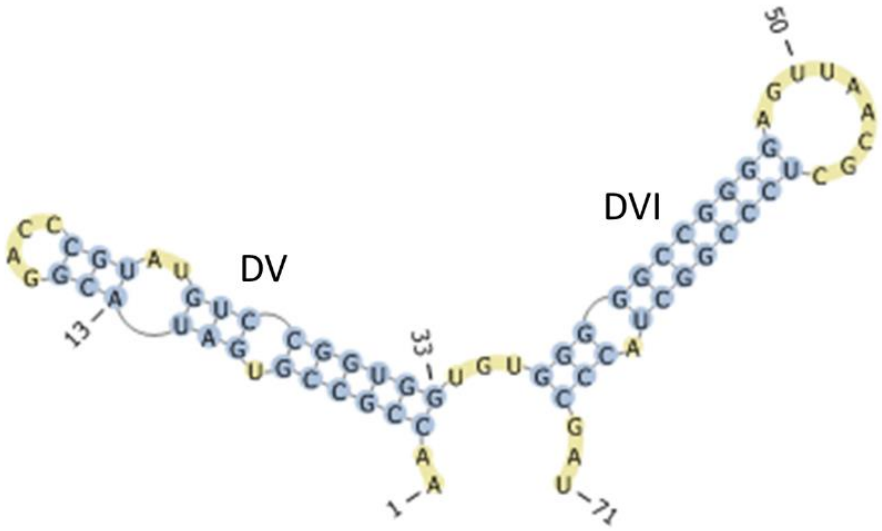


Fig.S6.6 Folding of DV and DVI RNA of truncated UHB.F1 within TSL1 metagenomic contig

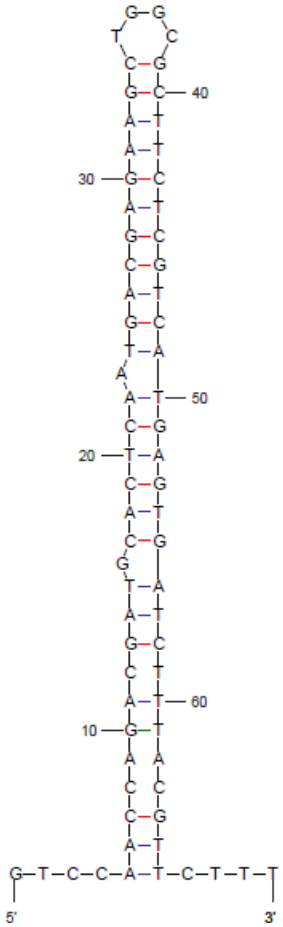
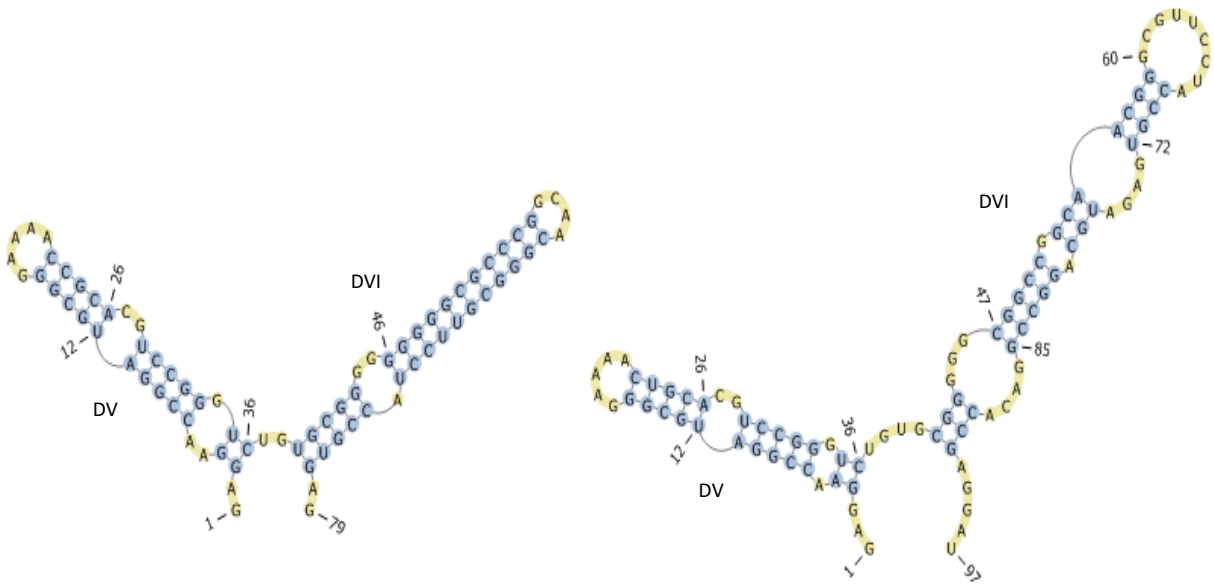
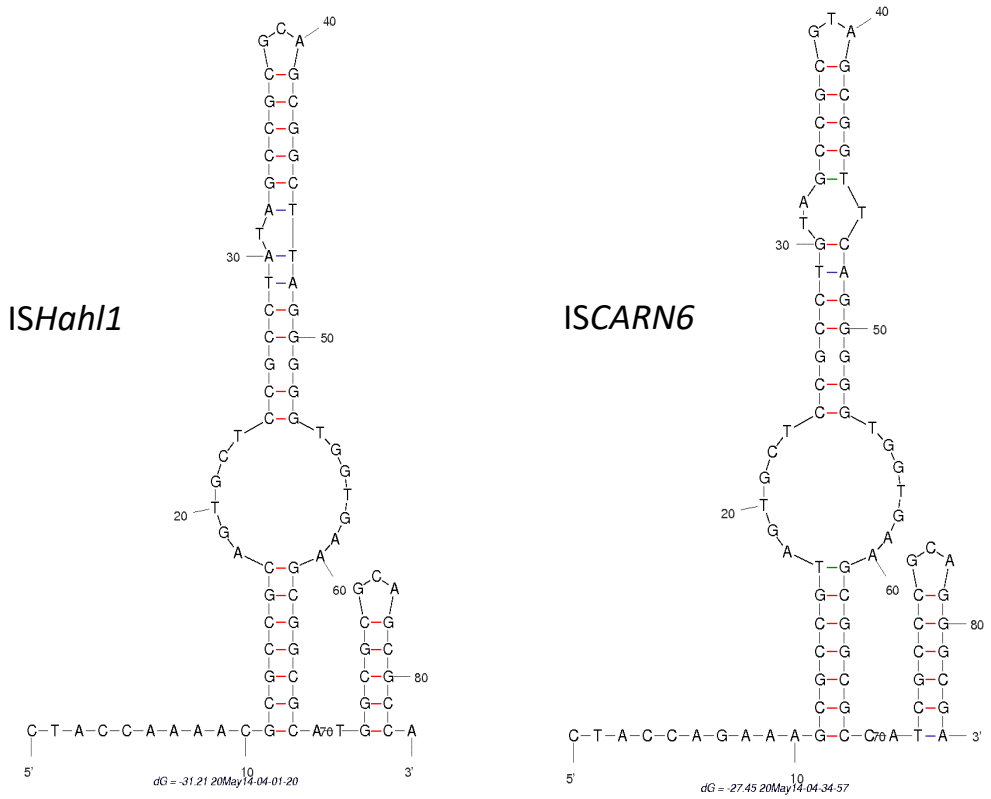


Fig.S6.7 5’ exon secondary structure of UHB.I2. attC top strand (ts) upstream of UHB.I2.

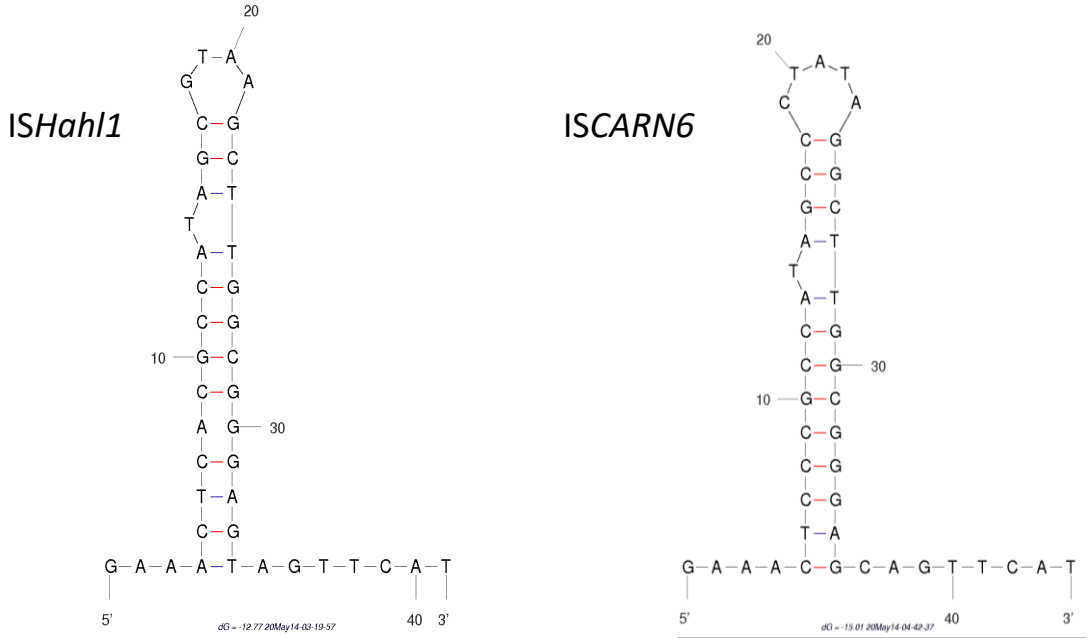


H.ha.F1 **H.ha.F2**

Fig.S6.8 Folding of DV and DVI RNA of fragmented group II introns identified within a CALIN in *H. halochloris.*



Left end hairpin structure



Right end hairpin structure

Fig.S6.9 Left and right end hairpin structures of IS*Hahl1* compared to IS*CARN6*, both belonging to IS*605* group of IS*200/605* superfamily. A conservation in secondary structure and to a lesser extent in primary structure is shown between left and right ends of both IS elements.

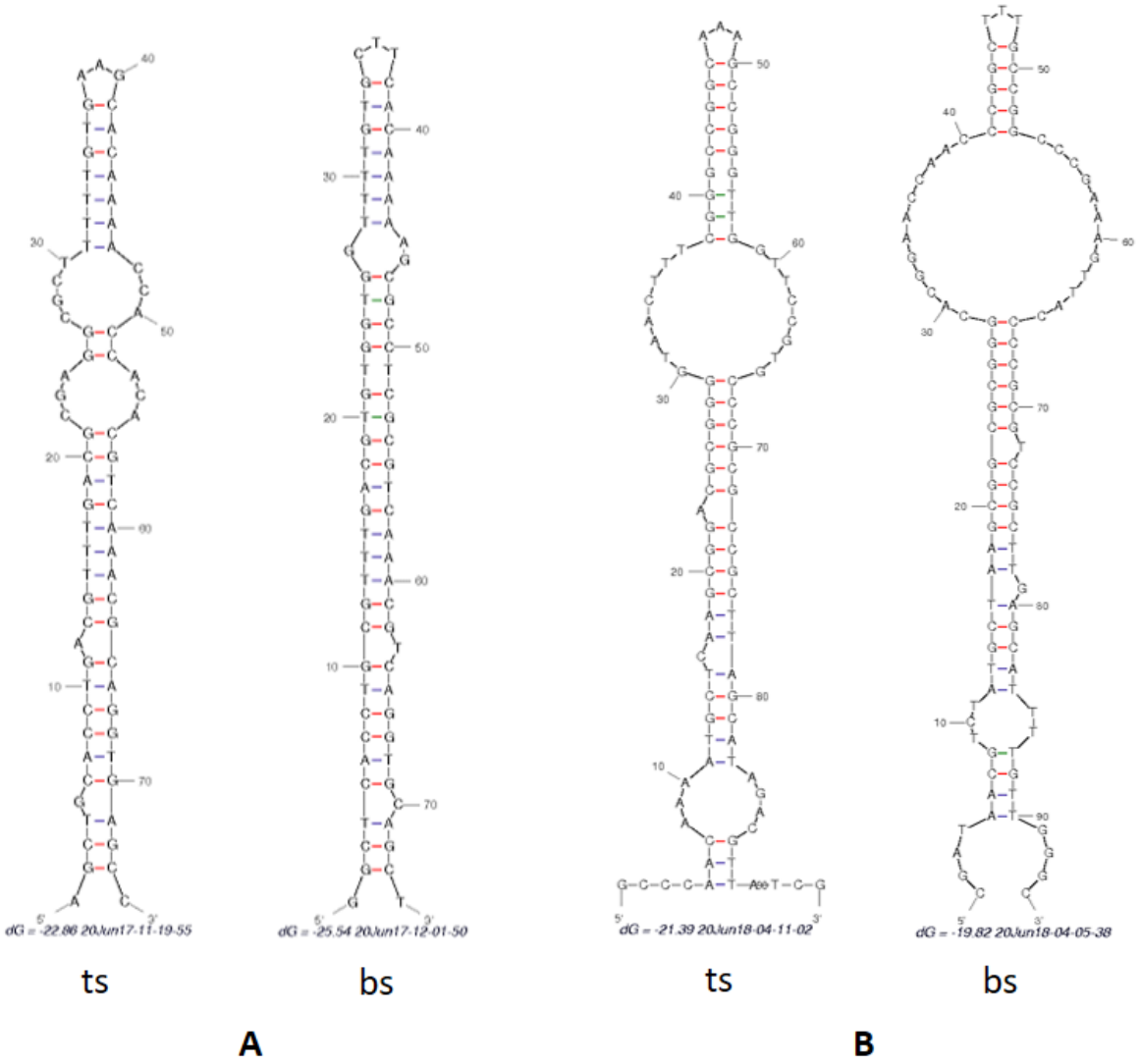


Fig.S6.10 Secondary structure of putative *att*C sites top strands (ts) and bottom strands (bs) undetected by integron Finder upstream H.ha.F1 and H.ha.F2. A: Atypical *att*C upstream H.ha.F1, B: Putative *att*C upstream H.ha.F2