Identification and Distribution of New Insertion Sequences in the Genome of the Extremely Halotolerant and Alkaliphilic *Oceanobacillus iheyensis* HTE831

Yoshihiro Takaki, Akiko Matsuki, Gab-Joo Chee, and Hideto Takami

Microbial Genome Research Group, Japan Agency for Marine-Earth Science and Technology, 2-15 Natsushima, Yokosuka 237-0061, Japan

(Received 18 March 2004; revised 4 June 2004)

Abstract

Six kinds of new insertion sequences (ISs), IS667 to IS672, a group II intron (Oi.Int), and an incomplete transposon (Tn8521oi) were identified in the 3,630,528-bp genome of the extremely halotolerant and alkaliphilic *Oceanobacillus iheyensis* HTE831. Of 19 ISs identified in the HTE831 genome, 7 were truncated, indicating the occurrence of internal rearrangement of the genome. All ISs except IS669 generated a 4- to 8-bp duplication of the target site sequence, and these ISs carried 23- to 28-bp inverted repeats (IRs). Sequence analysis revealed that four ISs (IS669, IS670, IS671, and IS672) were newly identified as belonging to separate IS families (IS200/IS605, IS30, IS5, and IS3, respectively). IS667 and IS668 were also characterized as new members of the ISL3 family. Tn8521oi, which belongs to the Tn3 family as a new member, generated a 5-bp duplication of the target site sequence and carried complete 38-bp IRs. Of the eight protein-coding sequences (CDSs) identified in Tn8521oi, three CDSs (OB481, OB482, and OB483) formed a ger gene cluster, and two other paralogous gene clusters were found in the HTE831 genome. Most of the ISs and the group II intron widely distributed throughout the genome were inserted in noncoding regions, while two ISs (IS667-08 and IS668-02) and Oi.Int-04 were inserted in the coding regions.

Key words: *Oceanobacillus iheyensis*; deep-sea; isolate; IS element; group II intron; halotolerant; alkaliphile

1. Introduction

*Bacillus* species are ubiquitous in nature. These organisms have often been isolated from various terrestrial soils and deep-sea sediments. Some *Bacillus* species have adapted to extreme environments, including high and low temperature, high and low pH, and high salinity.1,2 The genome sequences of two terrestrial *Bacillus* species, *Bacillus halodurans*3 and *B. subtilis*,4 have previously been reported. Through a series of genome analysis studies, it has become clear that the *B. halodurans* C-125 genome contains 15 kinds of new insertion sequences (ISs), IS641–IS672, IS650–658, IS660, IS662 and IS663, and a group II intron, in contrast to the genome of *B. subtilis* 168, in which IS elements are absent. *Oceanobacillus iheyensis* HTE831, which was isolated from deep-sea sediment collected at a depth of 1050 m on the Iheya Ridge and recently reclassified from the genus *Bacillus* (formerly *Bacillus* sp. HTE831), has extremely halotolerant and facultatively alkaliphilic properties.5 The entire genome sequence of *O. iheyensis* HTE831 has been determined.6 The *O. iheyensis* genome possesses 21 genes encoding putative transposases (Tpases) and reverse transcriptases/maturases/endonucleases (RT), similar to sequences present in the genomes of *B. halodurans*, *Marinococcus halophilus*, and *Enterococcus faecium*. Fourteen of those genes showed similarities to the Tpases of the IS elements categorized into various IS families such as ISL3,7 IS200/IS605,8,9 and IS30,10 and five were similar to the RT of the group II intron.

In this investigation, we identified and characterized six kinds of new ISs, a group II intron, and a transposon-like element in the 3,630,528-bp genome of *O. iheyensis* HTE831. We report here the distribution and orientation of the members of each IS and of the group II intron in the genome of strain HTE831, the structure and the target site sequence of each, and the alteration of protein-coding regions mediated by IS elements and the group II intron.
Table 1. IS elements, a transposon, and a group II intron in the *O. iheyensis* genome.

| IS element | Length of TSD (bp) | Direction | Position in genome | Length (bp) |
|------------|--------------------|-----------|--------------------|-------------|
| Oi.Int-01  | None               | +         | 81656-83544        | 1889        |
| IS670-01   | -                  |           | 171187-171044      | 149 (partial)|
| IS669-01   | +                  |           | 255507-255716      | 210 (partial)|
|            |                    |           | 255716-256231      | 516 (partial)|
| IS668-01   | -                  |           | 327007-326859      | 148 (partial)|
|            |                    |           | 327181-326999      | 183 (partial)|
| IS672-01   | -                  |           | 516488-507968      | 8521        |
| IS667-01   | +                  |           | 637607-638972      | 1366        |
| IS667-02   | 8                  |           | 666813-666774      | 40 (partial) |
|            |                    |           | 666932-666850      | 85 (partial) |
|            |                    | +         | 666935-666850      | 85 (partial) |
| IS667-03   | +                  |           | 964106-964752      | 647 (partial)|
|            |                    |           | 964755-965030      | 277 (partial)|
| IS670-04   | +                  |           | 137099-1397435     | 333 (partial)|
| IS668-02   | 8                  |           | 1460085-1461433    | 1349        |
| IS667-02   | 8                  |           | 1473203-1473557    | 355 (partial)|
|            |                    |           | 1473558-1474438    | 881 (partial)|
| IS667-05   | 8                  |           | 1484399-1483034    | 1366        |
| IS672-02   | 4                  | -         | 1677083-1678367    | 1285        |
| IS667-06   | 8                  | +         | 1687021-1688386    | 1366        |
| IS670-02   | 6                  | -         | 1765979-1764956    | 1024        |
| IS671-01   | 6                  | -         | 1871903-1870323    | 1581        |
| IS671-02   | 8                  |           | 2284759-2282871    | 1889        |
| IS670-07   | -                  |           | 2349951-2349749    | 204 (partial)|
| IS667-08   | 8                  |           | 2399023-2397658    | 1366        |
| IS668-03   | 8                  |           | 2438190-2439538    | 1349        |
| IS669-02   | None               |           | 2630632-2629894    | 739         |
| IS667-09   | 8                  |           | 2708472-2707107    | 1366        |
| Oi.Int-05  | None               |           | 2787411-2785523    | 1889        |

The direct repeat sequences flanking each intact element member are defined as TSD. The partial IS element without terminal sequence shorter than 100 bp is not basically defined as an IS element in this table.

2. Materials and Methods

2.1. Identification of the ISs, the group II intron, and the transposon in the HTE831 genome

The regions 600 bp upstream and downstream from each of the Tpase genes identified in our previous study\(^6\) were searched for inverted repeat (IR) sequences, using the GENETYX-Mac program, version 11, from Software Development Co., Ltd. (Tokyo, Japan). Whenever two Tpase genes overlapped or were located close to each other, the region 600 bp upstream from the first Tpase gene and the region 600 bp downstream from the second Tpase gene were searched for IR sequences in a similar manner. When an IR was found in the region flanking a Tpase gene, the regions adjacent to both IRs (IR\(_R\) and IR\(_L\)) were searched for direct repeat sequences, in order to identify target-site duplication (TSD). When an IR was not found, the whole genome sequence of strain HTE831 was searched for sequences showing nucleotide sequence similarity to the flanking regions 600 bp upstream and 600 bp downstream from the Tpase or RT gene, using the BLAST program, version 2.0 (BLAST2),\(^1^1\) to confirm the IS or group II intron region. The copy number of each IS was determined through a homology search of the entire genome of *O. iheyensis* HTE831, using BLAST2 in the Genome Gambler system.\(^1^2\) On the other hand, when a CDS showed a similarity to the Tpase gene of the transposon, the regions 10 kilobase pairs upstream and downstream from the CDS were searched for IR sequences in a similar manner.

3. Results

3.1. Identification and characterization of new IS elements in the HTE831 genome

In the previous study,\(^6\) we found many CDSs that showed homology with Tpase genes carried by various IS elements. In the present study, we identified and characterized six kinds of new ISs with or without terminal IRs and with or without TSDs, which are a direct repeat (DR) of the target-site duplication. All appear to belong to known IS families. Members of each IS and group II intron are listed in Table 1 and their locations are shown in Table 1 and Fig. 1.
Figure 1. Distribution of insertion sequence (IS) elements, the group II intron, and the transposon in the *O. iheyensis* HTE831 genome. Arrows indicate the direction of the elements and the copy number of each element is in parentheses.

Table 1. Details of terminal inverted repeat (IR) and target-site duplication (TSD) sequences of *O. iheyensis* HTE831 genome.

| Element | TSD (bp) | IR (bp) | Sequence |
|---------|----------|---------|----------|
| IS667-01 | 8        | 23      | 5’-tagtgggattttttTGCCTCAtaatATTGTGGGG--
 |          |          |         | 3’-agtaaatatttttagCGAGATtcctgAAACCCCA-- |
| IS668-02 | 8        | 28      | 5’tcttggataattcgCTCTCagatATGTTGGTT--
 |          |          |         | 3’tttttttcttctaatCTCAAGACagTTATTTACCCCA-- |
| IS669-02 | --       | --      | 5’-tttttatgtccttagaaaaacctcttgatgccccag--
 |          |          |         | 3’-agttttcttagttttggtacctgacctacacaa-- |
| IS670-01 | 6        | 25      | 5’-ttttttatgtagGAgATGATTTGTGAACTACCA--
 |          |          |         | 3’-taaaaagcctgGtaAAGACgtATTCTACGTTGT-- |
| IS671-01 | 6        | 28      | 5’-gccctccctttttGAGGCTCTTCAGGCTGAGAAACccT--
 |          |          |         | 3’-cggaggggctttttCTCCCGAAtAGTCagAACATCTCTTTctA-- |
| IS672-02 | 4        | 25      | 5’-cagcttttttTGAAGACACTAAAttATAGGACA--
 |          |          |         | 3’tctcttatAAGTggCTGGAATcAtTATCTACGT-- |
| Oi.Int-01 | --       | --      | 5’ttttttagtgctcccgctaatgtgctcagctataggg--
 |          |          |         | 3’-gaagttttatatactctttccctgggtgtggacccaa-- |
| Tn8521oi | 5        | 38      | 5’-tttaaaatataCAGGTAGCTCCAGGAAATGCAGATTACACCTAAAG--
 |          |          |         | 3’-agtagtattttCCCAATCTCGGAGTCTCATTGAGTGCACATTT-- |

Figure 2. Terminal inverted repeats (IRs) and target-site duplications (TSDs) of each element identified in the *O. iheyensis* genome. IRs are shown in blue uppercase letters and TSDs are boxed. Red letters indicate the sequence of the *O. iheyensis* genome.

Five kinds of IS elements were found to have IRs and to be flanked by TSDs. One of them, at position 637,607–638,972 in the genome (Fig. 1 and Table 1), has imperfect IRs 23 bp long (Fig. 2). The IS element designated IS667 was found to be flanked by an 8-bp TSD (Fig. 2 and Table 2). IS667, which is 1366 bp in length, is 62% identical to the nucleotide sequence of IS652 belonging to the ISL3 family, and the amino acid sequence of the Tpase of IS667 is 56% identical to that of IS652. The DDE motif, which is conserved in most Tpases and other enzymes capable of catalyzing the cleavage of DNA strands, was found in the open reading frame (ORF) of the IS667 element: D (154th amino acid [a.a.]), D (227th a.a.), E (365th a.a.), and K (372nd a.a.). This is a good match to the typical DDE motif pattern identified in ISs belonging to the ISL3 family (IS finder: Downloads from https://academic.oup.com/dnares/article-abstract/11/4/233/336271 by guest on 25 July 2018
Table 2. New IS elements, and a group II intron identified in the *O. iheyensis* genome.

| Insertion sequence | Size (bp) | TSD (bp) | IR (bp) | Number of elements | Family |
|--------------------|-----------|----------|---------|--------------------|--------|
|                    |           |          |         | Total | Intact | Truncated |
| IS667              | 1,366     | 8        | 23      | 9     | 6      | 3         |
| IS668              | 1,349     | 8        | 28      | 3     | 2      | 1         |
| IS669              | 739       | ---      | ---     | 2     | 1      | 1         |
| IS670              | 1,024     | 6        | 25      | 2     | 1      | 1         |
| IS671              | 1,581     | 6        | 28      | 1     | 1      | 0         |
| IS672              | 1,285     | 4        | 25      | 2     | 1      | 1         |
| Oi.Int             | 1,889     | ---      | ---     | 5     | 4      | 1         |

http://www-is.biotoul.fr/is.html. Moreover, IS667 terminates with 5′-GG–CC-3′, like the majority of the IS elements categorized into the ISL3 family. These findings support the view that IS667 should be categorized as a new member of the ISL3 family (Table 2). The genome of strain HTE831 has five other copies of intact IS667 (IS667-02, IS667-05, IS667-06, IS667-08, and IS667-09) and three other copies of truncated or partially deleted IS667 (Fig. 3 and Table 2). The IS element at position 1,460,085–1,461,433 (Fig. 1 and Table 1) has imperfect IRs 25 bp long, similar to IS667 (Fig. 2). This IS element (1349 bp in length), designated IS668, was found to be flanked by an 8-bp TSD (Fig. 2 and Table 2). IS668 is 52% identical to the nucleotide sequence of IS651, which duplicates an 8-bp sequence at the target site, and IS652 shows significant homology to IS667 as well. The amino acid sequence of the Tpase of IS668 is 40% identical to that of IS651, which is a member of the ISL3 family, and is also 38% identical to that of IS667. Similarly to the case of IS667, the typical DDE motif of the ISL3 family was identified in the ORF of IS668: D (154th a.a.), D (228th a.a.), E (300th a.a.), and K (367th a.a.). These results indicate that IS668 is another new member of the ISL3 family (Table 2).

The IS element at position 2,641,523–2,642,666 (Fig. 1 and Table 1) has imperfect IRs 25 bp long, similar to that of IS658 identified in the *B. halodurans* genome (Fig. 2). This IS element (1024 bp in length), designated IS670, was found to be flanked by a 6-bp sequence (Fig. 2 and Table 2). IS670 is 63% identical to the nucleotide sequence of IS658 belonging to the IS30 family. There are two ORFs overlapping at bp position 255–313 in IS670 (Fig. 3). It is evident that this occurred due to a frameshift mutation, because the first and second ORFs are both similar to the Tpase of IS658, with 53.1% and 71.4% identity at the amino acid sequence level, respectively. The typical DDE motif identified in the ISs categorized into the IS30 family was found to be flanked by the first ORF in IS670: D (145th a.a.), D (204th a.a.), E (240th a.a.), and K (247th a.a.). These findings support the view that IS670 should be categorized as a new member of the IS30 family (Table 2).

In addition, there are two other new ISs, IS671 and IS672, which carry terminal IRs and generate a TSD. IS671-01 at bp position 1,870,323–1,871,903 (Fig. 1 and Table 1) was found to be flanked by a 6-bp sequence (Fig. 2 and Table 2). IS671 is 51.1% identical to the nucleotide sequence of IS1562 belonging to the IS5 family, identified in the *Streptococcus pyrogenes* genome. This IS (1581 bp in length) has imperfect IRs 28 bp long and terminates with 5′-GG–CC-3′ (Fig. 2), similarly to other members of the IS5 family. There are two ORFs overlapping at bp position 672–697 in IS671, analogously to bp position 266–313 in IS670 (Fig. 3). It is evident that this occurred due to a frameshift mutation, because the first and second ORFs are both similar to the amino acid sequence of the Tpase of IS1562 (44.2% and 36.9% identity, respectively). The DDE motif was found in the second ORF of IS671 as well as other ISs, although the pattern did not have an exact match in any other IS element categorized into the IS5 family. Actually, the pattern of the DDE motif across the IS5 family varies depending on the IS involved. Based on these results, the inclusion of IS671 into the IS5 family seems to be indicated (Table 2). IS672-02 at bp position 1,870,323–1,871,903 (Fig. 1 and Table 1) has imperfect IRs 25 bp long (Fig. 2). This IS element (1285 bp in length) was found to be flanked by a 4-bp sequence (Fig. 2 and Table 2). IS672 is 49.5% identical to the nucleotide sequence of IS150 belonging to the IS3 family, identified in the *Escherichia coli* K12 genome. There are two ORFs in IS672 (Fig. 3). Mem-
Figure 3. Structure of each IS element and the group II intron identified in the *O. ibeyensis* genome. The box shows the Tpase or RT of each element and the numbers beside each box indicate the position of the Tpase in the element. The black horizontal bar indicates the elements identified in the genome. The black dashed lines indicate deleted parts of the element. The small vertical bar at the end of the element denotes IRs. The partial IS element without terminal sequence and shorter than 100 bp is not drawn in this figure.

Members of the IS3 family generally have two consecutive and partially overlapping reading frames, ORFA and ORFB, in relative translational reading phases 0 and −1, respectively. It has been demonstrated in at least three cases (IS150, IS3, and IS911) that, in addition to the product of the upstream frame, ORFA, a fusion protein, ORFAB, is generated by programmed translational frameshifting. The second ORF of IS672 is 34.4% identical to the amino acid sequence of the Tpase of IS150, although the first short ORF did not show significant similarity to it. The typical DDE motif identified in the Tpase of the IS3 family was found in the second ORF of IS672 (D, 145th a.a.; D, 204th a.a.; E, 240th a.a.; K, 247th a.a.). These findings support the view that IS672 is another new member of the IS3 family (Table 2).

Finally, the IS element designated IS669 with no IR or TSD was found to be present at bp position 2,629,894–2,630,632 in the HTE831 genome (Figs. 1–3 and Table 2). IS669 is 83% identical to the nucleotide sequence of IS657 belonging to the IS200/IS605 family, and the amino acid sequence of the Tpase identified in IS657 is 89% identical to that of IS669. Most of the IS elements categorized into the IS200/IS605 family do not carry IRs and are not flanked by TSD, whereas IS1535 has imperfect IRs 16 bp long, and IS657 and ISEnfa are flanked by 2 or 4 TSD (IS finder: http://www-
has been identified. Thus, we concluded that the inclusion of Oi.Int also showed significant

The amino acid sequence deduced from the CDS of Oi.Int is 62% identical to the
designated Bh.Int. The amino acid sequence

cal to the nucleotide sequence of the group II intron of

22

into target sites in double-stranded DNA.

3.2. Identification and characterization of the group II intron in the I. thegensis genome

Group II introns are catalytic RNAs that function as mobile genetic elements by inserting themselves directly into target sites in double-stranded DNA. The element, designated Oi.Int, has no IRs or TSDs (Fig. 2 and Table 2). This element (1889 bp in length) is 66% identical to the nucleotide sequence of the group II intron of B. halodurans designated Bh.Int. The amino acid sequence deduced from the CDS of Oi.Int is 62% identical to the putative RT of the Bh.Int. The amino acid sequence of the putative RT of Oi.Int also showed significant

identity to those of group II introns from Clostridium acetobutylicum (58%) and Nitrosomonas europaea (49%). On the other hand, the conserved motifs of group II intron ORFs consist of subdomains 0–7 of an RT domain, and is best revealed in the bacterial group II intron ORFs (pX01-23). Domains conserved among intron-encoded ORFs are denoted by the line above the alignment. RT0 through RT7 represent RT-like domains. The maturase-specific domain is designated X, and the finger-like zinc domain is designated Zn.

Figure 4. Protein alignment of the putative RT protein of Oi.Int with group II intron-encoded ORF in Bacillus anthracis (pX01-23). Domains conserved among intron-encoded ORFs are denoted by the line above the alignment. RT0 through RT7 represent RT-like domains. The maturase-specific domain is designated X, and the finger-like zinc domain is designated Zn.

is.biotoul.fr/is.html). On the other hand, the distinct DDE motif present across the IS200/IS605 family has not been identified. Thus, we concluded that the inclusion of IS669 into the IS200/IS605 family as a new member is valid, as well as that of IS657, because IS669 shares a very high degree of identity with IS657 at both the nucleotide level and the amino acid sequence level. Another copy with a deletion in an IS

29

and Lactococcus lactis (litrB). Domains conserved among intron-encoded ORFs are denoted by the line above the alignment. RT0 through RT7 represent RT-like domains. The maturase-specific domain is designated X, and the finger-like zinc domain is designated Zn.
Figure 5. Secondary-structure model of Oi.Int. The model was constructed by comparative analysis with a consensus group II intron structure model combined with RNA folding analysis using the MFOLD program (http://bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html). Roman numbers correspond to the six major structural domains of the group II introns. Potential tertiary pairings are designated $\alpha-\alpha'$, $\delta-\delta'$, $\varepsilon-\varepsilon'$, $\gamma-\gamma'$, $\zeta-\zeta'$, $\kappa-\kappa'$, and EBS1-IBS1. EBS3 represents a potential tertiary pairing with IBS3. ORF is encoded completely within domain IV. An asterisk indicates the lariat branch point A. Group II intron splicing occurs through the following two-step transesterification: the 2$'$$'OH$ group of a conserved intron adenosine residue in domain VI attacks the phosphodiester bond at the 5$'$$'splice site, forming a lariat form of the intron that contains the 2$'$$'–5$'$$'linkage and releasing the 5$'$$'exon; then, the 3$'$$'OH$ liberated at the end of the upstream exon attacks the 3$'$$'splice site, resulting in exon ligation and intron release.

Domain I extensively to form a catalytic core with long-range tertiary interaction. In domain V of Oi.Int, potential tertiary pairings ($\zeta-\zeta'$, $\lambda-\lambda'$, and $\kappa-\kappa'$) were identified in a consensus model. Domain VI forms a relatively variable structure. Most of the group II introns have a bulging A on the 3$'$$'side of the basal helix of domain VI, at either seven or eight nucleotides from the 3$'$$'splice site. A branch point of Oi.Int was found seven nucleotides from the 3$'$$'splice site (AU). These findings support the view that Oi.Int should be categorized as a new member of the group II intron class, although the splicing of Oi.Int has not been confirmed experimentally.

The HTE831 genome contains four other copies of the element (Oi.Int-01, and -03 to -05) and a truncated copy of Oi.Int (Oi.Int-02) (Table 1, Figs. 1 and 3).

3.3. Identification and characterization of the transposon-like element in the HTE831 genome

The proteins deduced from the three CDSs (OB485, OB486, and OB487) showed significant similarity to the Tpase of Tn1546 classified in the Tn3 family. The amino acid sequence of OB485 (182 a.a.) was 88.9% identical to that of the region from the 546th to the 691st amino acid of the Tpase consisting of 988 amino acids. Similarly, the amino acid sequences of two other proteins were 90% and 86% identical to those of the regions from the 743rd to the 862nd (OB486) and from the 889th to the 988th (OB487) amino acids of the Tpase, respectively. Thus, the Tpase appears to be divided into three parts in the HTE831 genome, although the total length of the amino acid sequence of these three proteins cover only 42% of whole region of the Tpase of Tn1546 (Fig. 6A).
Figure 6. Structure of the transposon-like element containing the spore germination gene cluster (SGGC), and the distribution of SGGCs in the genomes of bacilli. A. Structure of the transposon-like element identified in the HTE831 genome and comparison with that of Tn1546. Gray arrows show spore germination (ger)-related genes. B. Distribution of SGGCs in the HTE831 genome and comparison with those of B. subtilis 168 and B. halodurans C-125. The 12 o'clock position of the oriC region is assigned 0°. Gray arrows indicate the ger-related gene.
Tn1546 identified in the *E. faecium* genome\(^{34}\) constitutes a transposon unit, which carries a vancomycin-resistance gene cluster. To investigate whether there is a transposon-like unit containing the three Tpase parts in the HTE831 genome, the nucleotide sequences down-stream from OB487 and upstream from OB485 were searched for terminal IRs and TSDs. We found that an 8521-bp fragment containing OB485–OB487 had IRs and was flanked by a TSD (Fig. 6A). The fragment at bp position 507,968–516,488 in the HTE831 genome had perfect IRs 38 bp long and was found to be flanked by a 5-bp TSD (Fig. 2 and Table 1). The nucleotide sequences of the IRs were almost identical to those of Tn1546 possessing imperfect IRs (Fig. 6A). In addition, it was found that there are five other CDSs in the element designated Tn8521oi, excluding the three Tpase parts, whereas Tn8521oi lacks the resolvase identified in Tn1546. The amino acid sequences of the proteins deduced from the three CDSs (OB481–OB483) were similar to those of the gerKC (48%), gerKB (54%), and gerKA (55%) gene products encoding a spore germination protein identified in *B. cereus*. OB480, 73.3% identical to the amino acid sequence of the yraG gene product identified in *B. subtilis*, was annotated as a spore coat protein, and the amino acid sequence of OB484 was 52.6% identical to that of aldehyde dehydrogenase (*BHI005*) identified in *B. halodurans*. Thus, the element designated Tn8521oi appears to be a trace of the transposon unit, which has carried a spore germination (ger) gene cluster, a spore coat protein gene, and an aldehyde dehydrogenase gene to the HTE831 genome, because Tn8521oi lacks the complete form of the Tpase and resolvase, which play an important role in autonomous transposition.

3.4. Alteration of protein-coding regions mediated by ISs and the group II intron

In order to investigate how CDSs are affected by ISs and the group II intron of the genome, the CDSs in the regions adjacent to each IS were analyzed. The nucleotide sequence of the 3-kb region upstream and the 3-kb region downstream from all intact IS elements identified in this study were extracted using “ExtremoBase” from the entire genome sequence of the HTE831 genome database (http://www.jamstec.go.jp/jamstec-e/bio/DEEPSTAR/FResearch.html). The 6-kb sequence from which the IS region was excised (Fig. 7) was searched for CDSs to nr-aa protein database using BLAST2. Although most of the IS elements widely distributed throughout the genome were inserted in non-coding regions, at least two CDSs located downstream from the Tpase were likely affected by the insertion of two kinds of IS elements (IS667-08 and IS668-02). As shown in Fig. 7A, OB2341 and OB1417 appear to have occurred due to the truncation of the original CDS by IS insertion. OB2341 has been annotated as a kinase-associated protein B, but the function of OB1417 is still unknown. The gene encoding OB1417 appears to have become 59 bases longer than the original one through the insertion of IS668-02, and in the case of OB2341, the size of the original gene product (133 a.a.) appears to have increased to 140 amino acids with the insertion of IS667-08. Thus, among the 12 intact ISs of six kinds identified in this study, two intact IS elements of two kinds consequently seem to have affected each CDS by their insertion.

On the other hand, as shown in Fig. 7B, the CDS OB2243 located downstream from the RT of Oi.Int-04 was found to be affected by insertion through homing or retrotransposition, although three other intact Oi.Ints were inserted in noncoding regions, as in the case of Bh.Int identified in the *B. halodurans* genome\(^{13}\). The example of a group II intron inserted in the coding sequence region is very unusual even in other bacterial genomes\(^{35,36}\). In fact, Oi.Int-04 was inserted at the position between the 2nd and 3rd letter in the 149th codon of OB2243, creating a stop codon (TAG). Assuming that the self-splicing of Oi.Int-04 occurs in the HTE831 genome, whereas no example is known thus far of a group II intron being inserted in the coding region with opposite transcriptional direction to the coding sequence, another stop codon (TAA) will occur at the same position instead of TAG due to the intrinsic nucleotide sequence of the genome, and, eventually, no change will occur in the amino acid sequence of OB2243.

4. Discussion

We searched the entire genome of strain HTE831 for transposable elements such as ISs, transposons, and group II introns. Nineteen new ISs of six kinds, classified into the ISL3, IS200/IS605, IS3, IS5, and IS30 families, were identified in the genome, although their variety and number are much fewer than the 120 found in the *B. halodurans* genome\(^{13}\) and in *B. cereus* strain C-125 to date, but two other new members (IS667 and IS668) showing significant similarity to those identified in strain C-125 were found in the HTE831 genome in this study. It is intriguing that the IS elements identified in *B. halodurans* were found in species phylogenetically distant from each other. The group II intron and members of families IS3, IS4, IS6, IS21, IS30, IS200/IS605, IS256, IS481, IS630, IS650/IS653, IS656/IS662, IS660/IS1272, IS982, and ISL3 have been reported in other *Bacillus* species.
strains. As described above, four copies of intact Oi.Int were found in the HTE831 genome, and one was inserted in the coding sequence region. To the best of our knowledge, this is the first example among bacilli of a group II intron present in the coding sequence region.

A trace of the transposon unit, designated Tn8521oi and consisting of 8521 bp, is present in the HTE831 genome. A ger gene cluster composed of three CDSs (OB481, OB482, and OB483) was found in the Tn8521oi element (Fig. 6A). The HTE831 genome has two other paralogous gene clusters located at 73° and 114°, respectively, on the circular chromosome, if the 12 o’clock position of the oriC region is assigned 0°. The genome of B. subtilis 168 has five gene clusters (gerA, gerB, gerK, yndDEF, and yfkQRT) orthologous to the ger gene cluster carried to the HTE831 genome by Tn8521oi, although the gerBC gene product did not show significant homology to the amino acid sequence of the putative corresponding protein (OB481) in the fifth gene cluster located at 315° (Fig. 6B). The spores are thought to recognize germinants such as L-alanine, L-valine, L-asparagine, glucose, fructose, and KCl through receptor proteins encoded by the gerA family of operons, which includes gerA, gerB, and gerK.37 To substantiate the receptor function of the gerA family of operons in spore germination, a mutant B. subtilis strain lacking all three gerA-like operons was constructed, as well as two putative gerA homologs, yndDEF and yfkQRT, and it has been confirmed that spores lacking all gerA-like operons germinate at a very low frequency in rich media, compared to single-mutant spores lacking one gerA-like operon. Moreover, it has been concluded that the products of the gerA, gerB, and gerK operons play a major role and the predicted proteins encoded by yndDEF and yfkQRT play a minor role in nutrient-induced spore germination, because the gerA-gerB-gerK triple-mutant spores behaved identically to the quintuple-mutant spores.38 The HTE831 genome has only three paralogous ger gene clusters, even when including the one acquired by the insertion of Tn8521oi, in contrast to five in the case of B. subtilis, and four in

---

**Figure 7.** Alteration of protein-coding regions mediated by ISs and the group II intron. **A.** IS elements. **B.** Group II intron. An alteration occurs in the C-terminal region of the protein-coding sequence (CDS) upon insertion of the IS element and a group II intron. Black and gray boxes indicate the insertion sequences (IS or group II intron) and the HTE831 chromosome, respectively. Dark-gray arrows represent original CDSs before insertion of the IS elements and group II intron.
Table 3. IS family members identified in the genome of Bacillus-related species.

| IS family | O. iheyensis | Other Bacillus-related strains | Reference |
|-----------|--------------|--------------------------------|-----------|
| **IS3**   | 1 (IS672)    | B. thuringiensis sub. aizawai, B. thuringiensis, B. halodurans | 4, 13, 39-41 |
| **IS4**   | 0            | B. thuringiensis sub. thuringiensis, B. halodurans, B. subtilis natto, G. stearothermophilus* | 15, 13, 42-46 |
| **IS5**   | 1 (IS671)    |                                | 0         |
| **IS6**   | 0            | B. thuringiensis sub. israelensis, B. thuringiensis sub. fukuokaensis, B. cereus | 4, 39, 47, 48 |
| **IS21**  | 0            | B. thuringiensis sub. thuringiensis, B. halodurans, G. stearothermophilus* | 5, 13, 44, 46 |
| **IS30**  | 1 (IS670)    | B. halodurans                   | 1, 13     |
| **IS110** | 0            | B. halodurans                   | 2, 13     |
| **IS200/IS605** | 1 (IS669) | B. halodurans                   | 1, 13     |
| **IS256** | 0            | B. halodurans                   | 1, 13     |
| **IS481** | 0            | G. stearothermophilus*          | 1, 49     |
| **IS630** | 0            | B. halodurans, G. stearothermophilus* | 2, 13, 50 |
| **IS650/IS653** | 0    | B. halodurans                   | 1, 13     |
| **IS656/IS662** | 0    | B. halodurans                   | 1, 13     |
| **IS660/IS1272** | 0  | B. halodurans                   | 1, 13     |
| **IS982** | 0            | B. thuringiensis, G. stearothermophilus* | 2, 50, 51 |
| **ISL3**  | 2 (IS667)    | B. halodurans                   | 2, 13     |

*Geobacillus stearothermophilus

the case of B. halodurans (Fig. 6B). Although it is unclear what role these three gerA-like operons identified in the HTE831 genome play in spore germination, and no experimental data for the frequency of germination from O. iheyensis spores has been obtained, these genomic characteristics in the ger gene cluster presumably imply a low frequency of spore germination in O. iheyensis. On the other hand, a low frequency of spore formation in generally rich media has been observed in strain HTE831.5 Thus, a low-frequency spore former, O. iheyensis may have no serious problems in spore germination under enriched conditions, even if the number of paralogous gerA-like operons is smaller than that of high-frequency spore formers.

OB480, similar to the gene encoding the spore coat protein, appears to have been acquired by the genome
with the insertion of Tn8521oi as well as a ger gene cluster. Orthologs can easily be found in the genomes of *B. subtilis* (yraG and yraE) and *B. halodurans* (BH0742). However, no Tpase gene has been identified in the *B. ter. Orthologs can easily be found in the genomes of *B. subtilis* (BH0742) downstream from the yraG gene and there is no CDS showing similarity to the Tpase gene in the regions upstream and downstream from the ger gene clusters and BH0742 in the *B. halodurans* C-125 genome. Therefore, the acquisition of spore-related genes due to the insertion of a transposon seems to be very unusual in bacilli.

References

1. Sneath, P. H. A. 1986, In: Sneath, P. H. A., Mair, N. S., Sharpe, M. E., and Holt, J. G. (eds.), Bergey’s Manual of systematic bacteriology, vol. 2, Williams and Wilkins, Baltimore, MD, pp. 1104–1139.

2. Priest, F. G. 1993, In: Sonenshein, A. L., Hoch, J. A. and Losick, R. (eds.), *Bacillus subtilis* and other Gram-positive bacteria, American Society for Microbiology, Washington, D.C., pp. 3–16.

3. Takami, H., Nakasone, K., Takaki, Y., et al. 2000, Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. *Nucleic Acids Res.*, 28, 4317–4331.

4. Kunst, F., Ogasawara, N., Moszer, I., et al. 1997, The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, 390, 249–256.

5. Lu, J., Nogi, Y., and Takami H. 2001, *Oceanobacillus iheyensis* gen. nov., sp. nov., a deep-sea extremely halotolerant and alkaliphilic species isolated from a depth of 1050 m on the Iheya Ridge. *FEMS Microbiol. Lett.*, 205, 291–297.

6. Takami, H., Takaki, Y., and Uchiyama, I. 2002, Genome sequence of *Oceanobacillus iheyensis* isolated from the Ihey Ridge and its unexpected adaptive capabilities to extreme environments. *Nucleic Acids Res.*, 30, 3927–3935.

7. Germond, J. E., Lapierre, L., Delley, M., and Mollet, B. 1995, A new mobile genetic element in *Lactobacillus delbrueckii* subsp. *bulgaricus*. *Mol. Gen. Genet.*, 248, 407–416.

8. Beuzon, C. R. and Casadesus, J. 1997, Conserved structure of IS200 elements in *Salmonella*. *Nucleic Acids Res.*, 25, 1355–1361.

9. Censini, S., Lange, C., Xiang, Z., et al. 1996, cag, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proc. Natl. Acad. Sci. USA*, 93, 14648–14653.

10. Caspers, P., Dalrymple, B., Iida, S., and Arber, W. 1984, IS30, a new insertion sequence of *Escherichia coli* K12. *Mol. Gen. Genet.*, 196, 68–73.

11. Altschul, S. F., Maiden, T. L., Schaffer, A. A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389–3402.

12. Sakiyama, T., Takami, H., Ogasawara, N., et al. 2000, An automated system for genome analysis to support microbial whole-genome shotgun sequencing. *Biosci. Biotech. Biochem.*, 64, 670–673.

13. Takami, H., Han, C. G., Takaki, Y., and Ohtsubo, E. 2001, Identification and distribution of new insertion sequences in the genome of alkaliphilic *Bacillus halodurans* C-125. *J. Bacteriol.*, 183, 4345–4356.

14. Mahillon, J. and Chandler, M. 1998, Insertion sequence. *Microbiol. Mol. Biol. Rev.*, 62, 725–774.

15. Engler, J. A. and van Bree, M. P. 1981, The nucleotide sequence and protein-coding sequence capability of the transposable element IS5. *Gene*, 14, 155–163.

16. Ferretti, J. J., McShan, W. M., Ajdic, D., et al. 2001, Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc. Natl. Acad. Sci. USA*, 98, 4658–4663.

17. Timmerman, K. P. and Tu, C. P. 1985, Complete sequence of IS3. *Nucleic Acids Res.*, 22, 2127–2139.

18. Schwartz, E., Kroger, M., and Rak, B. 1988, IS150 distribution, nucleotide sequence and phylogenetic relationships of a new *E. coli* insertion element. *Nucleic Acids Res.*, 16, 6789–6802.

19. Vogele, K., Schwartz, E., Welz, C., Schiltz, E., and Rak, B. 1991, High-level ribosomal frameshifting directs the synthesis of IS150 gene products. *Nucleic Acids Res.*, 19, 4377–4385.

20. Sokine, Y., Eisaki, N., and Ohtsubo, E. 1994, Translational control in production of transposase and in transposition of insertion sequence IS3. *J. Mol. Biol.*, 235, 1406–1420.

21. Polard, P., Prere, M. F., Chandler, M., and Fayet, O. 1991, Programmed translational frameshifting and initiation at an AUU codon in gene expression of bacterial insertion sequence IS911. *J. Mol. Biol.*, 222, 465–477.

22. Belfort, M., Derbyshire, V., Parker, M. M., Cousineau, B., and Lambowitz, A. M. 2002, In: Craig, N. L., Crigie, R., Gellert, M., Lambowitz, A. M., (eds.), Mobile DNA II, American Society for Microbiology, Washington, D.C., pp. 761–783.

23. Nölling, J., Breton, G., Omelchenko, M. V., et al. 2001, Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. *J. Bacteriol.*, 183, 4823–4838.

24. Chain, P., Lamerdin, J., Larimer, F., et al. 2003, Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph *Nitrosonomas europaea*. *J. Bacteriol.*, 185, 2759–2773.

25. Xiong, Y. and Eickbush, H. 1990, Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.*, 9, 3353–3362.

26. Mohr, G., Perlman, P. S., and Lambowitz, A. M. 1993, Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Res.*, 21, 4991–4997.

27. Shub, D. A., Goodrich-Blair, H., and Eddy, S. R. 1994, Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns. *Trends Biochem. Sci.*, 19, 402–403.

28. Lehman, K. and Schmidt, U. 2003, Group II introns: Structure and catalytic versatility of large natural ribozymes. *Crt. Rev. Biochem. Mol. Biol.*, 38, 249–303.
fer of pRS01 in Lactococci. J. Bacteriol., 178, 3531–3538.
30. Michel, F. K., Umesono, K., and Ozeki, H. 1989. Comparative and functional anatomy of group II introns—a review. Gene, 82, 3–30.
31. Costa, M. and Michel, F. 1995. Frequent use of the same tertiary motif by self-folding RNAs. EMBO J., 14, 1276–1285.
32. Jacquier, A. and Michel, F. 1990. Base-pairing interactions involving the 5′ and 3′-terminal nucleotides of group III self-splicing introns. J. Mol. Biol., 213, 437–447.
33. Hefferon, F., MacCarthy, B. J., Ohtsubo, H., and Ohtsubo, E. 1979. DNA sequence analysis of the transposon Tn3: three genes and three sites involved in transposition of Tn3. Cell, 18, 1153–1163.
34. Arthur, M., Molinas, C., Depardieu, F., and Courvalin, P. 1993. Characterization of Tn1546, a Tn3-related transposon conferring glycopeptide resistance by synthesis of depsipeptide peptidoglycan precursors in Enterococcus faecium BM4147. J. Bacteriol., 175, 117–127.
35. Bonen, L. and Vogel, J. 2000, The ins and outs of group II introns. Trends Genet., 17, 322–331.
36. Edgell, D. R., Belfort, M., and Shub, D. A. 2000, Barriers to intron promiscuity in bacteria. J. Bacteriol., 182, 5281–5289.
37. Moir, A. and Smith, D. A. 1990. The genetics of bacterial spore germination. Annu. Rev. Microbiol., 44, 531–553.
38. Paidhungat, M. and Setlow, P. 2000, Role of Ger proteins in nutrient and nonnutrient triggering of spore germination in Bacillus subtilis. J. Bacteriol., 182, 2513–2519.
39. Smith, G. P., Ellar, D. J., Keeler, S. J., and Seip, C. E. 1994, Nucleotide sequence and analysis of an insertion sequence from Bacillus thuringiensis related to IS150. Plasmid, 32, 10–18.
40. Dunn, M. G. and Ellar, D. J. 1997, Identification of two sequence elements associated with the gene encoding the 24-kDa crystalline component in Bacillus thuringiensis ssp. fukaoaensis: An example of transposable element archaeology. Plasmid, 37, 205–215.
41. Hodgman, T. C., Ziniu, Y., Shen, J., and Ellar, D. J. 1993, Identification of a cryptic gene associated with an insertion sequence not previously identified in Bacillus thuringiensis. FEMS Microbiol. Lett., 114, 23–30.
42. Nagai, T., Tran, L.-S., P., Inatsu, Y., and Itoh, Y. 2000, A new IS4 family insertion sequence, IS4Bsu1, responsible for genetic instability of poly-γ-glutamic acid production in Bacillus subtilis. J. Bacteriol., 182, 2387–2392.
43. Mahillon, J., Seurinck, J., Van Rompuy, L., Delcour, J., and Zabeau, M. 1985, Nucleotide sequence and structural organization of an insertion sequence element (IS231) from Bacillus thuringiensis strain berliner 1715. EMBO J., 4, 3895–3899.
44. Lereclus, D., Ribier, J., Kiler, A., Menou, G., and M.-M., Lecadet. 1984, A transposon-like structure related to the δ-endotoxin gene of Bacillus thuringiensis. EMBO J., 3, 2561–2567.
45. Kronstand, J. and Whiteley, H. R. 1984, Inverted repeat sequences flank a Bacillus thuringiensis crystal protein gene. J. Bacteriol., 160, 95–102.
46. Xu, K., He, Z.-Q., Mao, Y.-M., Sheng, R.-Q., and Sheng, Z.-J. 1993, On two transposable elements from Bacillus steatorrhophilus. Plasmid, 29, 1–9.
47. Delecuse, A., Bourgouin, C., Kiler, A., and Rapoport, G. 1989, Nucleotide sequence and characterization of a new insertion element, IS240, from Bacillus thuringiensis israelensis. Plasmid, 21, 71–78.
48. Bourgouin, C., Delecuse, A., Ribier, J., Kiler, A., and Rapoport, G. 1988, A Bacillus thuringiensis subsp. israelensis gene encoding a 125-kilodalton larvicidal polypeptide is associated with inverted repeat sequences. J. Bacteriol., 170, 3575–3583.
49. Shulami, S., Gat, O., Sonenshein, A.-L., and Shoham, Y. 1999, The glucuronic acid utilization gene cluster from Bacillus steatorrhophilus T-6. J. Bacteriol., 181, 3695–3704.
50. Kiel, J. A., Bosels, J. M., Berge, A. M., and Venema, G. 1993, Two putative insertion sequences flank a truncated glycogen branching enzyme gene in the thermophile Bacillus steatorrhophilus CU21. DNA Seq., 4, 1–9.
51. Yu, W., Mierau, I., Mars, A., Johnson, E., Dunny, G., and McKay, L. L. 1995, Novel insertion sequence-like element IS982 in Lactococci. Plasmid, 33, 218–225.
52. Robert, A. R., Montgomery, K. N., Smith K. L., and Zimmerly, S. 2004, Principles of 3′ splice site selection and alternative splicing for an unusual group II intron from Bacillus anthracis, RNA, 10, 854–862.