A novel approach for identifying the laterally transferred genetic elements in bacteria

Dongru Qiu, Mitsuhiro Itaya, Masaru Tomita

[1] Institute for Advanced Biosciences and Bioinformatics Program, Graduate School of Media and Governance Keio University, Tsuruoka, 997-0017, Japan

Abstract: In *Bacillus subtilis* Marburg 168 trpC2, there are 11 chromosomal *rap* genes and several *rap* genes reside within or adjacent to the presumptive prophage regions. on the chromosome and multiple *rap* genes are borne by the extra-chromosomal mobile elements including the conjugative plasmid and bacteriophage. Phylogeny analysis is conducted on the *rap* genes from *B. subtilis* and other 4 closely related species, *B. anthracis*, *B. cereus*, *B. thuringiensis*, and *B. halorandus*, to explore the relationship between gene duplication and lateral gene transfer (LGT). Based on the phylogeny inference, it is suggested that at least some of the *B. subtilis* chromosomal *rap* genes may have recently arisen by the recurrent LGT mediated by the temperate phages and/or conjugative plasmids and these multiple chromosomal *rap* genes may derive from a single common origin. In this study, the correlation between LGT and Chi sequence frequency is also explored based on the *B. subtilis* genome sequence. It is demonstrated that the prophage elements and many other identified LGT elements normally exhibit extremely low Chi sequence frequency. Several novel elements, including the poly-▣-glutamic acid (PGA) capsule synthesis genes, the cell wall associated protein *wap*A gene, and penicillin-binding protein 2A gene (*pbp*A) flanking regions, exhibit significantly low Chi frequency and may also be LGT elements, though they are speices-specific characteristics to some extent. Most of the open reading frames of these newly identified regions do not have orthologs in the above-mentioned 4 species and they may represent ancient LGT. These regions exhibit normal base composition of the *B. subtilis* chromosome and therefore are difficult to detect by using base composition and codon usage analyses. It is suggested that Chi sequence frequency bias could indicate genome heterogeneities and could serve as an index for identifying LGT elements and inferring the integration history. It is obvious that LGT plays a crucial role in shaping bacterial genome.

Key words: Chi sequence frequency, Lateral gene transfer, Gene duplication, codon usage, *Bacillus subtilis*

Introduction

The whole genome sequencing of the gram-positive endospore-former, *Bacillus subtilis* Marburg 168 trpC2, had been completed in 1997 and strikingly it was found that the genome contains at least 10 prophages or remnants of prophages, accounting for 9 % of the whole genome. The *B. subtils* genome is 43.5% G+C rich on the whole and the prophage (-like) heterogeneous regions exist as the A+T rich islands on the chromosome and therefore were readily detected. This fact is pointing to a significant role for bacteriophage infection in the transfer of genes during bacterial evolution (Kunst *et al.*, 1997; Moszer *et al.*, 1998). These possible cryptic prophages contain a very high proportion of genes of unknown functions (Y-genes) and three prophage elements, the SP□, PBSX and skin element, had been identified before the whole genome sequencing. Automatic detection of LGT genes is frequently performed by using measures such as codon usage and/or G+C content transfer (Lawrence & Ochman, 1997; Rocha et al, 1998; Ochman *et al.*, 2001). However, these methods had recently been criticized for high rates of both false positives and false negatives (Koski *et al.*, 2001; Wang, 2001). In *B. subtilis*, several efforts have been made to reveal the heterogeneities of *B. subtilis* 168 genome sequence and a series of interesting features of genomic organization have been found (Moszer *et al.*, 1999; Rocha *et al.*, 1999; Nicolas *et al.*, 2002). The *B. subtilis* genome has been partitioned into three well defined classes by the codon usage bias of genes. The third class, with AT-rich codon, corresponds to laterally transferred elements, mainly prophages elements, indicative of the existence of systematic lateral gene transfer in this organism (Moszer *et al.*, 1999). By using hidden Markov models (HMM), 14 new regions were demonstrated to be horizontally transferred elements (Nicolas *et al.*, 2002).

B. subtilis 168 strain is naturally competent and in *B. subtilis*, the AddAB gene is the functional counterpart of the RecBCD of *E. coli*, and may play the similar role in the homologous DNA recombination and the specific Chi sequence that it recognizes is a five-nucleotide sequence (5'-AGCGG-3' or its complement 5'-CCGCT-3') (Chedin *et al.*, 2000). This Chi (Bs) could attenuate the nuclease activity of the AddAB enzyme and therefore a protruding 3'-terminated single-stranded tail is produced, which can facilitate the recombination process. The Chi sequence is overrepresented and is not evenly distributed along the chromosome. The Chi orientation bias is mostly due to the uneven distribution of G content (GC skew), instead of the replication-related function of Chi sequences (Uno *et al.*, 2000). It has been demonstrated that certain prophage inserted regions exhibit lower Chi frequency (El Karoui *et al.*, 1999). In the present study, the correlation between Chi sequence frequency and LGT are systematically explored in the *B. subtilis* genome.

Nevertheless, 11 *rap* genes, encoding the response regulator aspartate phosphatases (Rap), are present on the chromosome (Reizer *et al.*, 1997) and several genes are located within or very close to the prophage regions. Nevertheless, other 8 *rap* genes have been found on the rolling-circle and theta-type plasmids of *B. subtilis*, as well as in one of the sequenced temperate phage, □ -105. More importantly, at least some *rap* genes are functionally redundant (Jiang *et al.*, 2000). It is very unusual for a bacterium to have so many functionally redundant genes because the bacterial genomes are usually streamlined. These facts indicate that some *rap* genes may arise by gene duplication and are LGT elements. The phylogeny clustering is conducted to explore LGT of *rap* genes since other *Bacillus* genome sequences, including *B. halorandus* C-125 (Takami *et al.*, 2000), *B. thuringiensis* serovar *morrisoni*a (Lee *et al.*, 2002), *B. anthracis* A2012 and Ames (Read *et al.*, 2002; 2003), and *B. cereus* ATCC 14579 (Ivanova *et al.*, 2003), are available in the public database. Here we show that several chromosomal *rap* genes may have arisen by prophage and/or conjugative mediated LGT and the Chi sequence frequency bias of specific chromosome regions may serve as a good indicator for identifying LGT elements.

Sequence data and methods

Almost all the sequence data are extracted from the NCBI database and the *rap* gene homologs are retrieved by using BLAST, except the sequence of theta-type large plasmid pLS32 determined recently (Itaya & Tanaka, unpublished). Phylogeny clustering is conducted by using ClustalW multiple-alignment software package. Codon adaptation index (CAI) (Sharp & Li, 1987), frequency of optimal codons (Fop) (Ikemura, 1981) and codon bias index (CBI) (Bennetzen & Hall,1982) are calculated by using John Peden's codonW 1.3 program (http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html). The O/E (observed/expected) values for Chi sequence are calculated based on the base composition of specific genome regions and whole genome sequence (Arakawa *et al.*, 2003; http://www.g-language.org/). The NCBI COG database is also used for orthologous group retrieval (http://www.ncbi.nlm.nih.gov/COG/).

Results:

1. The phylogeny inference of *rap* family genes

Though base composition and codon usage bias are normally applied to identify LGT, the most convincing way is based on inter-/intra-specific phylogenetic inference (Ragan, 2001). Here we take advantage of the *rap* gene family phylogeny analysis to explore LGT in *B. subtilis*. There are 11 *rap* genes, *rap*A-*rap*K, residing on the chromosome of the sequenced *B subtilis*168. Three *rap* genes, *rap*40, *rap*50, and *rap*60 are found on the rolling circle plasmid pTA1040, pTA1050 and pTA1060, respectively (Meijer *et al.*, 1998). Theta-type large plasmids pLS20 (Meijer *et al.*, 1995) and pLS32 have also one *rap* gene, *rap*LS20 and *rap*LS32, respectively. In addition, one *rap* gene, *rap*□ -105 was also found in the sequenced *B. subtilis* phage □ -105. There are also several *rap* gene homologs found in *B. anthracis*, *B. cereus*, *B. thuringiensis*, and *B. halorandus*, respectively. Totally 41 gene sequences, encoding Rap proteins or hypothetical phosphatase proteins in the 5 *Bacillus* bacteria, can be retrieved from NCBI database by using BLASTP. The phylogenetic clustering of both nucleotide sequence and amino acid sequence of these *rap* genes is conducted by using ClustalW multiple alignment software package (Figure 1).

It is obvious that the *rap* genes of *B. subtilis* and *B. halorandus* largely constitute distinct within-species clusters, respectively, while the relationship among the *rap* genes of other 3 *Bacillus* species is much more complicated (Figure 1). It is believed that *B. subtilis* is closer to *B. halorandus*, and their common progenitor diverged from the common progenitor of other 3 species, which diverged more recently.

In *B. subtilis*, most of the *rap* family genes, including those from its cognate plasmids and the bacteriophage □ -105, probably derived from a single origin since there are no obvious orthologous genes for each *rap* gene found in other *Bacillus* species. The *rap*E gene, which is currently located within the prophage skin element, is much closer to the *rap* genes found on the rolling circle plasmids pTA10040, pTA10050, and pTA10060 (Meijer *et al.*, 1998), which is substantiated by the bootstrap analysis. The *rap*I gene, which is currently located within the prophage 2, is closely related to the *rap* gene on the theta-type plasmid pLS32 (*rap*LS32), while the *rap*K, located within the prophage 6, and *rap*G, located within a presumptive LGT region, are very similar to and the phage □ -105 *rap* gene (*rap*□ 105). The cluster of *rap*G, *rap*K and *rap*□ 105 seems less related to other *B. subtilis rap* genes, indicating their recent integration of these 2 chromosomal genes. The *rap*LS20 is closer to the *B. halorandus rap* gene cluster than to the major *B. subtilis* cluster. The *rap* A, B, and H genes form a closer cluster based on both nucleotide and amino acid sequence alignment, but their relationship is not significantly supported by bootstrap analysis, indicating they have well diverged. Though *rap*A gene is now situated just outside the presumptive PBSX prophage region, it may be previously part of degrading defective PBSX. It seems that several chromosomal *rap* genes, for example, *rap*E, *rap*K, *rap*I, and *rap*G, recently arose via the phage and/or plasmid mediated LGT in *B. subtilis*. It is suggested that most of *rap* genes of *B. subtilis*, except *rap*LS20, are actually paralogous genes deriving from a single progenitor gene and duplicated via LGT after the speciation of this species, though they have already highly diverged. *Rap*C and *rap*F may represent recently arisen paralogs. This may be also the case for the *rap* genes of *B. halorandus*, and their common ancestor *rap* gene may be the orthologue to that of *B. subtilis*.

However, unlike those of *B. subtilis* and *B. halorandus*, the *rap* genes from other 3 *Bacillus* species constitute several cross-species close clusters, consistent with the close phylogenetic relatedness of these 3 species. The *rap* gene BA1180 from *B. anthracis* is more closely related to *rap* genes from *B. subtilis* and *B. halorandus* than to its within-species homologs, in particular, it is more similar to *B. subtilis rap*D gene in amino acid sequence. Therefore, it may represent an ortholog to the *rap* genes of *B. subtilis* and *B. halorandus*. Most *rap* genes in *B. anthracis*, *B. cereus*, and *B. thuringiensis*, may have arisen before the divergence of these 3 bacteria. Obviously, the clusters of BT-rapB, BA1582, and BC1026 and of BT-*rap*E, BC3581, and BA3790 may represent the orthologues in different species while most of other *rap* genes of same species, for example BT-*rap*C and BT-*rap*F, are paralogues arisen by gene duplication after the species divergence. The lateral gene transfer among these 3 pathogenic species may readily occur because of their recent divergence, though their host is different. The LGT mediated gene duplication may also have occurred in these species. For example, the BA3760 gene, which is located in the prophage □Ba01 region on the chromosome of *B. anthracis* str Ames, is much closer to the *rap*C and *rap*F of *B. thuringiensis*.

The multiple chromosomal *rap* genes of *B. subtilis* may have derived from one single original copy, and most of them, previously borne by the same or different extrachromosomal elements, progressively integrated into chromosome by recurrent prophage and/or conjugative plasmid mediated within-species LGT events. This gene duplication is unique, in contrast to the normal gene duplication occurred within bacterial chromosome.

2. Chi sequence frequency bias along the chromosome of *B. subtilis*

The Chi sequence (including its complement) frequency along the chromosome of *B. subtilis* is shown in Figure 2, with a 10 kb sliding window. The average Chi sequence (including its complement) number over 10kb is 26.91±9.31 and lower than 18/10kb is defined as the low Chi frequency. It is obvious that the frequency of Chi sequence dramatically changes along the chromosome and numerous low Chi sequence frequency islands exist. Interestingly, most of these low Chi sequence frequency regions are particularly corresponding to the previously identified LGT elements, in addition to some well-conserved elements like ribosomal RNA (*rrn*) operons and heat shock protein genes. These LGT elements include the prophages (except PBSX) and most of those LGT elements detected by using HMM (Nicolas *et al.*, 2002) or repeat analysis (Rocha *et al.*, 1999). Only 2 of previously identified LGT elements, including the arsenic resistance regulon and the region of around 1385~1424kb, do not exhibit extremely low Chi sequence frequency (Table2). There seems to be a real correlation between LGT elements and their low Chi sequence frequency, rather than only an accidental coincidence, with such an overall consistency.

The Chi frequency and codon usage indices for the 10 prophages are calculated for comparison. 9 out of the 10 prophage elements, except PBSX, exhibit low Chi frequency (Table 3) as well as high codon usage bias (Table 4). The 3 codon usage indices, CAI, CBI and Fop, are largely consistent for these prophages. There also exists a well correlation between Chi sequence frequency and codon usage indices for these prophages (Table 3 and 4). Moreover, the O/E values for Chi sequence frequency are significantly different among these prophage elements, which is highly implicative of the different selection pressure to each prophage element. On the whole, the Chi sequence is overrepresented in the *B. subtilis* genome because the O/E values for both Chi sequence and its complement are about 2.1, while those are very close to 1 for several prophages, including the prophage 2, 6, 7 and SP□ and the sequenced phage □ -105. The prophages that exhibit higher Chi sequence frequency, also exhibit higher O/E values, indicating that the high Chi frequency is not only dependent on the higher GC content, but also positively selected, as a result of codon usage adaptation or other causes (Table 3).

Normally the larger prophages exhibit lower Chi sequence frequency and the O/E values are close to 1 because they may be more intact in sequence structure and subject to less selection pressure with a relatively recent integration. The largest and most intact prophage SP□ exhibits the lowest Chi sequence frequency among the 10 prophages

(Table3) and constitutes a wide low Chi frequency island on the chromosome (Figure 2). Similarly, it also exhibits high codon usage bias. Both codon usage indices and Chi frequency strongly indicate the recent integration of prophage SP□, which has basically intact operons. Nevertheless, the O/E values for Chi sequence are very close to 1 in SP□ (Table 3), indicating that its low Chi frequency mainly results from base composition, instead of the selection pressure from host. This is also consistent with the experimental evidence that SP□ is absent in the genome of many natto strains, *B. subtilis* (natto), the starter of fermented Japanese food natto.

The skin element is also absent in many natto strains (Sato & Kobayashi, 1998), but some *B. subtilis* strains without SP□, still have skin element (unpublished data). This fact indicates that it may also have recently integrated into the genome of *B. subtilis* 168 before SP□ integration. However, the Chi frequency, as well as codon usage indices (except CBI), is not very high among the prophage elements (Table 3 and 4). The relatively high GC content (40.1%) of skin element may partly account for its high chi sequence frequency. However, the O/E values for skin element are relatively high and close to those of the host genome, indicating the Chi sequence has also been positively selected. Nevertheless, the Chi sequence distribution is also significantly biased within this prophage and the middle region (the late operon) exhibits a higher level of Chi sequences than the rest region, indicating that multiple recombination events may have occurred in this operon region. The operon disruption resulted from recombination may be responsible for the lysogenic conversion of this prophage, which will be further discussed below.

PBSX is the only prophage that could not be detected by automatic computer survey because of its normal base composition and codon usage. Nevertheless, it exhibits the highest Chi sequence frequency and the O/E values for Chi frequency are also the highest among the prophages, indicating frequent recombination has occurred within this region. The frequent recombination events may have greatly facilitated whole genome sequence amelioration, and also made the pseudogenes or short ORFs concentrated in the decaying defective prophages in *B. subtilis*. The local Chi sequence frequency somehow represents the history of LGT integration in the host genome and therefore may provide a tool to identify LGT and infer the integration history. Generally, the sequence fragments with a lower Chi frequency and highly biased codon usage are more recently integrated elements. However, the reverse may not be necessarily true: the element with a higher Chi frequency and/or less biased codon usage, for example, PBSX, may not necessarily have a long integration history.

In addition, some low Chi sequence frequency regions represent the evolutionarily conserved structural metabolic and regulatory gene such as flagellar genes and heat shock proteins, while several other low Chi frequency region may be LGT elements (Table 2).These listed regions normally have multiple unknown genes (Y-gene) and short ORFs (Pseudogene?), which have no orthologs in the closely related *Bacillus* species. However, most of these newly identified regions exhibit normal base composition of *B. subtilis*, in contrast to those LGT elements that have been previously identified by using analyses of codon usage, Hidden Markov Models or repeats (Moszer *et al.*, 1998; Rocha *et al.*, 1999; Nicolas *et al.*, 2002). The polyglutamic acid (PGA) capsule synthesis genes, penicillin binding protein (PBP) 2A gene (*pbp*A), and the wall associated protein encoding *wap*A gene are the most prominent examples for these newly identified elements because to some extent they are the species-specific identity characteristics for this bacterium. The genes upstream of *wap*A are functionally unique and have short ORFs, and particularly they have no othologs in other *Bacillus* species. Therefore, they are obviously LGT elements. The *wap*A itself may also be a LGT element, though In *B. anthracis*, PGA capsule formation are important to its virulence, but the PGA synthesis genes *cap*A, B, and C are the plasmid *p*OX2 encoded genes, instead of chromosomal genes (Makino *et al.*, 1989), strongly indicating its LGT origin. The horizontal transfer of capsular synthesis genes and PBP synthesis genes is well documented in the clinical strains of pathogenic bacteria (Dowson *et al.*, 1989; Coffey *et al.*, 1991). The penicillin binding protein 2A gene is functionally redundant with *pbp*H gene product and is crucial for rod-shape determination (Murray *et al.*, 1997; Wei *et al.*, 2003). Since most of other *pbp* genes exhibit normal Chi sequence frequency, the under-representation of Chi sequence in *pbp*A gene may indicate its LGT origin. On the whole, these regions may represent relatively ancient LGT elements that may arose before the divergence of Bacillus species and have been undergoing differential degradation in different species or strains because a few genes within these areas obviously have orthologs in one or more related species.

Therefore, it is highly indicative of LGT if local low Chi sequence frequency appears on the bacterial chromosome. The low Chi frequency may act as a complementary index for identifying LGT.

Discussion:

1. LGT and *Rap* gene duplication

At least some of the *B. subtilis rap* genes are functionally redundant, for example, RapA, RapB, and RapE proteins specifically dephosphorylate the spo0F~P intermediate of the sporulation phosphorelay (Jiang *et al.*, 2000). Recently it has demonstrated that the plasmid borne *rap*60 gene is also fully functional in *B. subtilis,* dephosphorylating a component of sporulaiton phosphorelay (Koetje *et al.*, 2003). This component may also be spo0F~P because Rap60 is highly homologous to RapE in amino acid sequence (Figure 1). It is suggested that at least several chromosomal *rap* genes, including *rap*E, G, K, C, F, and I, may have recently arisen by LGT mediated gene duplication (Figure 1). Though *rap*D and BA1180 may represent the orthologous gene in the two different species, they may also be highly homologous LGT genes from the same origin. Moreover, *rap*E, *rap*I, and *rap*K are still residing in the skin element, prohage 2 and prophage 6 regions while *rap*A gene lies just outside the presumptive PBSX region on the chromosome in *B. subtilis* 168. Actually, *rap*A may also previously be a part of original PBSX prophage because its borders, as well as most prophage's borders, is not well defined as those of SP□ and skin element, which definitely marked their repetitive attachment site sequences. In addition, *rap*A, E, I, and K genes still have their cognate regulator gene *phr*A, E, I, and G. These cognate polypeptide regulator genes are partly overlapped with their upstream *rap* gene, and overlapping genes are normally found in virus whose genome is highly compact. All these facts strongly indicate the LGT origin of these *rap* genes. Interestingly, *rap*A, E, I, and K genes are free of Chi sequence, while other chromosomal *rap* genes have one or several copies of Chi sequences (data not shown). Furthermore, it seems that more recently arisen *rap* genes tend to be intact in gene structure with the cognate *phr* gene remained, while some *rap* genes may have lost their cognate *phr* gene as a result of mutations or recombination occurred on either chromosome or the phage and plasmid DNA. These chromosomal *rap* genes should be tightly controlled or some of them have already functionally diverged, otherwise the sporulation and competence development, crucial to the survival of this bacterium in the field, will be blocked with so many *rap* genes in its genome. Some Rap proteins without cognate *phr*, such as RapB and RapH, may be cross-regulated by the *phr* peptide from other *rap* genes, or their expression is properly controlled by the upstream regulators. The functional divergence of *rap* genes, which arose from similar LGT origin(s) and are functionally redundant, may still be underway in *B. subtilis*. Nevertheless, these *rap* genes may represent a novel type of selfish genes because their original function is to inhibit the host attempt of sporulation and therefore favor the replication of the phage or plasmid, other than to act as the quorum sensing system of *B. subtilis*.

2. Chi sequence frequency and LGT

Prophages normally exhibit low Chi sequence frequency, which may result from two causes. First, the occurrence probability of the GC-rich Chi sequence, AGCGG, and its complement CCGCT, is low in the AT-rich prophage sequence. In phages, the preferential usage of A-containing codons is consistent with their rapid replication because cytoplasmic ATP concentration is higher than other three nucleotides. Secondly, the directional mutation pressure to the temporarily integrated phages may not be that high as to the chromosome sequence because of its short residence time in host genome. For example, the O/E values for Chi sequence are very close to 1 in the phage □ -105 and the recently integrated prophage SP□, which may result from less selection pressure. In this case Chi sequence frequency may principally depend on base composition. Some nucleotide sequence will be degraded by the nuclease activity of the AddAB enzyme when recombination happens as a result of DNA damage or other causes. However, the appearance of Chi sequence resulting from mutations will protect prophage sequence from decaying caused by AddAB in recombination. The Chi sequence frequency of LGT will gradually increase because part of sequence may be degraded and at the same time mutations leading to Chi sequence are positively selected by the AddAB exerted selective pressure, which is obvious from the O/E values for Chi sequence (Table 3). It has also been demonstrated that overrepresentation of Chi sequence could only partly explained by their adaptation to codon usage (El Karoui *et al.*, 1999). Chi sequence could accumulate as result of frequent recombination because the DNA damage and mutations of LGT region are not lethal to the host bacteria. On the other hand, LGT elements may rapidly decay as a result of the restriction-modification system or other mechanisms unless they bring particular evolution advantage to the host genome. For instance, prophage 3 is the smallest among the 10 prophages, and has the smallest number of ORFs and the lowest gene density. Therefore, the prophage 3 region was considered the oldest prophage region, with only restriction-modification genes remaining undestroyed because this R-M system is evolutionarily advantageous to the host genome (Ohshima *et al.*, 2002). Nevertheless, relocation of chromosomal genes to prophage regions, as a result of recombination, may also increase the local Chi sequence frequency. The longer the history of prophage element integration, the more the Chi sequence occurrence. This may be also the case for other LGT mediated by transformation and conjugation. From this hypothesis, it is proposed that Chi frequency could serve as the indicator for identifying LGT.

Frequent recombination dramatically disrupted the phage operons and therefore a lot of unique and functionally unknown gene, or pseudogenes, are concentrated in these decaying LGT elements. Though it still could be induced by the SOS response and result in cell lysis with the release of phage-like particles, PBSX has become a defective phage because PBSX cannot properly assemble and it packages the randomly selected bacterial chromosomal DNA instead of its own DNA (Anderson & Bott, 1985). The phage particles kill sensitive bacteria without injecting DNA. This prohage exhibits the highest Chi frequency, even higher than the average level of the whole genome, indicating that multiple recombination events may have disrupted its original sequence with only its late operon remaining largely undisrupted. Skin element, PBSX, and SP□ share a high homology in some genes, which is related to phage functions. Though two non-phage-like operons in the skin element are expressed and have distinct expression profiles that are dependent on the growth and developmental status of the cell, the expression of the late operon was not detected during exponential growth, during sporulation or after induction of the SOS response (Krogh *et al.*, 1996; Krogh *et al.*, 1998). Since skin element also has relatively high Chi frequency, especially in the middle region (late operon), it is speculated that multiple recombination events may have occurred, resulting in the gene loss and operon disruption that help to inactivate this prophage. SP□ is the largest and most intact among the prophages. As a consequence, SP□ exhibits the lowest Chi sequence frequency and high codon usage bias, which is consistent with its more recent integration (Lazarevic *et al.*, 1999).

As compared to phages, the sequenced small cognate plasmids of *B. subtilis* exhibit normal levels of Chi sequence frequency (data not shown). This is because phages could survive in the dormant form and is much more independent of host genome while plasmids replicate and descend along with host chromosome generation after generation and exposed to the same selection pressure, if any, as the chromosomal genes.

The shorter Chi sequence could lead to more frequent recombination in *B. subtilis*, which may have greatly facilitated LGT because frequent recombination events could rapidly disrupt prophage operon and inactivate the prophages as described above, perpetuating the existence of these prophages in the host genome. Low Chi frequency also reflects high conservation of the gene sequence, especially the essential structural, metabolic and regulatory genes, for example, the class III heat shock protein genes (including *clp*X) and cell shape determining genes (including *rod*A) exhibit very low Chi frequency. However, these genes normally have orthologues in closely related species and exhibit normal codon usage, making it easy to distinguish these well conserved genes from real LGT elements.

In summary, we have explored the relationships between the lateral gene transfer and chi sequence frequency and between gene duplication and LGT in *B. subtilis*, raising the possibility of using Chi frequency as an LGT indicator. The more recently integrated elements exhibit lower Chi frequency, as a result of discrepancy in both base composition and selection pressure, but the Chi sequence is normally overrepresented in the *B. subtilis* genome. Therefore, low chi frequency could serve as an indicator for LGT. However, like the compositional/codon signatures, the Chi sequence frequency of LGTs decay over time due to the mode of host genome evolution. It needs to be confirmed whether this method is applicable in other bacterial species. Even though we already have many bacterial genome sequences in hand, it seems far from enough to discriminate LGT because one single strain only represents a "snapshot" of the genome of each species. Moreover, the LGT elements encode some specific traits that are sometimes used to identify bacteria, but this can lead to an "identity crisis" (Doolittle, 2002). This is also the case for *B. subtilis*, for example, the Rap phosphatase-regulated phosphorelay cascade and two-component signaling systems play a central role in the metabolism and development of this bacterium and somehow species-specific, though they may have arisen by LGT. In this work, we proposed that some novel elements are LGT based on the low Chi frequency of these genes and their flanking regions, though it is also possible that the low Chi frequency result from high conservation. To identify LGT, it will be more reliable to use the different ways to generate more information and then carefully scrutinize all evidence before making a conclusion.

Acknowledgement: The authors would like to thank Dr. Hideyuki Ohshima, Dr. Yoshiaki Ohashi, and Mr. Likai Mao for their help. DQ is the recipient of a Mori grant of Keio University.

References

Anderson LM, Bott KF (1985) DNA packaging by the Bacillus subtilis defective bacteriophage PBSX. J Virol 54: 773-780.

Arakawa K, Mori K, Ikeda K, Matsuzaki T, Kobayashi Y, Tomita M (2003) G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining. Bioinformatics 19:305-306.

Bennetzen JL, Hall BD (1982) Codon selection in yeast. Journal of Biological Chemistry 257: 3026-3031.

Chedin F, Ehrlich SD, Kowalczykowski SC, 2000. The *Bacillus subtilis* AddAB helicase/nuclease is regulated by its cognate Chi sequence in vitro. J Mol Biol 298:7-20

Coffey TJ, Dowson CG, Daniels M, Zhou J, Martin C, Spratt BG, Musser JM (1991) Horizontal transfer of multiple penicillin-binding protein genes, and capsular biosynthetic genes, in natural populations of *Streptococcus pneumoniae.* Mol Microbiol 5:2255-2260.

Doolittle RF (2002) Microbial genomes multiply. Nature 416: 697-700.

Dowson CG, Hutchison A, Brannigan JA, George RC, Hansman D, Linares J, Tomasz A, Smith JM, Spratt BG (1989) Horizontal transfer of penicillin-binding protein genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae.* Proc Natl Acad Sci USA 86:8842-8846

El Karoui M, Biaudet V, Schbath S, Gruss A. (1999) Characteristics of Chi distribution on different bacterial genomes. Res Microbiol 150:579-587.

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* system. J Mol Biol 151: 389-409.

Ivanova N, Sorokin A, Anderson I, Galleron N, Candelon B, Kapatral V, Bhattacharyya A, Reznik G, Mikhailova N, Lapidus A, Chu L, Mazur M, Goltsman E, Larsen N, D'Souza M, Walunas T, Grechkin Y, Pusch G, Haselkorn. R, Fonstein M, Dusko Ehrlich S, Overbeek R, Kyrpides N (2003) Genome sequence of Bacillus cereus ATCC 14579 and comparison with *Bacillus anthracis*. Nature 423:87-91

Jiang M, Grau R, Perego M (2000) Differential processing of propeptide inhibitors of Rap phosphatases in *Bacillus subtilis*. J Bacteriol 182: 303-310.

Koetje EJ, Hajdo-Milasinovic A, Kiewiet R, Bron S, Tjalsma H (2003) A plasmid-borne Rap-Phr system of *Bacillus subtilis* can mediate cell-density controlled production of extracellular proteases. Microbiology 149:19-28.

Koski LB, Morton RA, Golding GB (2001) Codon bias and base composition are poor indicators of horizontally transferred. Mol Biol Evol 18: 404-412.

Krogh S, O'Reilly M, Nolan N, Devine KM (1996) The phage-like element PBSX and part of the skin element, which are resident at different locations on the *Bacillus subtilis* chromosome, are highly homologous. Microbiology 142:2031-2040.

Krogh S, Jorgensen ST, Devine KM (1998) Lysis genes of the *Bacillus subtilis* defective prophage PBSX. J Bacteriol 180:2110-2117.

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

Lazarevic V, Dusterhoft A, Soldo B, Hilbert H, Mauel C, Karamata D (1999) Nucleotide sequence of the *Bacillus subtilis* temperate bacteriophage SPbetac2. Microbiology, 145:1055-1067.

Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44: 383–397.

Makino S, Uchida I, Terakado N, Sasakawa C, Yoshikawa M (1989) Molecular characterization and protein analysis of the *cap* region, whichis essential for encapsulation in *Bacillus anthracis*. J Bacteriol 171: 722-730.

Meijer WJ, de Boer AJ, van Tongeren S, VenemaG., Bron S (1995) Characterization of the replication region of the *Bacillus subtilis* plasmid pLS20: a novel type of replicon. Nucleic Acids Res 23: 3214-3223.

Meijer WJ, Wisman GB, Terpstra P, Thorsted PB, Thomas CM, Holsappel S, Venema G, Bron S (1998) Rolling-circle plasmids from *Bacillus subtilis*: complete nucleotide sequences and analyses of genes of pTA1015, pTA1040, pTA1050 and pTA1060, and comparisons with related plasmids from gram-positive bacteria. FEMS Microbiol Rev 21:337-368.

Moszer I, Rocha EP, Danchin A (1999) Codon usage and lateral gene transfer in *Bacillus subtilis*. Curr Opin Microbiol 2:524-528.

Murray T, Popham DL, Setlow P (1997) Identification and characterization of pbpA encoding Bacillus subtilis penicillin-binding protein 2A. J Bacteriol, 179: 3021-3029.

Nicolas P, Bize L, Muri F., Hoebeke M., Rodolphe F., Ehrlich S.D., Prum B, Bessieres P (2002) Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models. Nucleic Acids Res, 30:1418-1426.

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299-304.

Ohshima H, Matsuoka S, Asai K, Sadaie Y (2002) Molecular organization of intrinsic restriction and modification genes BsuM of *Bacillus subtilis* Marburg. J Bacteriol 184:381-389

Ragan MA (2001) On surrogate methods for detecting lateral gene transfer. FEMS Microbiol Lett 201: 187-191.

Read TD, Salzberg SL, Pop M, Shumway M, Umayam L, Jiang L, Holtzapple E, Busch JD, Smith KL, Schupp JM, Solomon D, Keim P, Fraser CM (2002) Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. Science 296:2028-2033.

Read TD, Peterson SN, Tourasse N, *et al* (2003) The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. Nature. 423:81-86.

Reizer J, Reizer A, Pergo M, Saier MHJr (1997) Charaterization of a family of bacterial response regulator aspartyl- phospatase (RAP) phosphatases. Microb Comp Genomics 2: 103-111.

Rocha EP, Viari A, Danchin A (1998) Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons. Nucleic Acids Res 26:2971-2980.

Rocha EP, Danchin A, Viari A (1999) Analysis of long repeats in bacterial genomes reveals alternative evolutionary mechanisms in *Bacillus subtilis* and other competent prokaryotes. Mol Biol Evol 16:1219-1230.

Sato T, Kobayashi Y (1998) The ars operon in the skin element of *Bacillus subtilis* confers resistance to arsenate and arsenite. J Bacteriol 180:1655-1661.

Sharp PM, Li WH (1987) The codon adaptation index a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15: 1281-1295.

Takami H, Nakasone K, Takaki Y, Maeno G., Sasaki R, Masui N, Fuji F, Hirama C, Nakamura Y, Ogasawara N, Kuhara S, Horikoshi K (2000) Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. Nucleic Acids Res 28 (21):4317-4331.

Uno R, Nakayama Y, Arakawa K, Tomita M (2000) The orientation bias of Chi sequences is a general tendency of G-rich oligomers. Gene 259:207-215.

Wang B (2001). Limitations of compositional approach to identifying horizontally transferred genes. J Mol Evol 53: 244-250.

Wei Y, Havasy T, McPherson DC, Popham DL (2003) Rod shape determination by the Bacillus class B peninicillin-binding proteins encoded by pbpA and pbpH. J Bacteriol 185: 4717-4726.

Figure legend

Figure 1 Phylogenetic clustering of *rap* genes found in the chromosome as well as the plasmids and phage of *B. subtilis* (BS), *B. halorandus* (BH), *B. cereus* (BC), *B. thuringiensis* (BT), and *B. anthracis* (BA) by using ClustalW. Upper panel: Amino acid sequence alignment of Rap proteins; Lower panel: Nucleotide sequence alignment of *rap* genes. The numbers at branches represent the bootstrap values estimated from 1000 resampling.

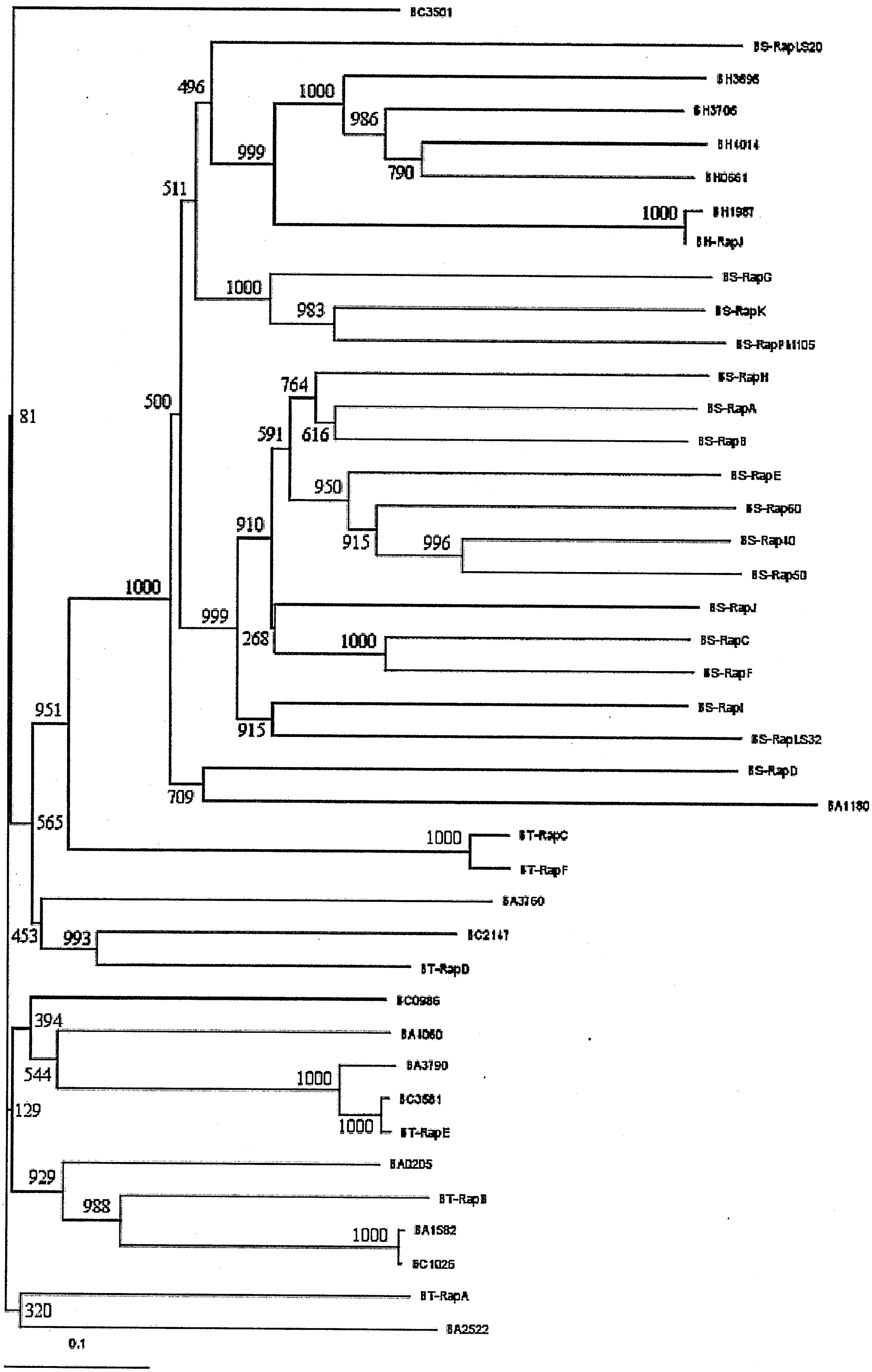
Figure 2 Chi sequence number changes along the chromosome of *B. subtilis* 168 (Sliding window: 10kb)

Table 1 Rap gene or hypothetical gene sequences and their accession numbers.

Gene or locus name	Accession number	Reference
BS-rapA	16078308; NP_389125.1	Kunst *et al.*, 1997
BS-rapB	16080722; NP_391550.1	Kunst *et al.*, 1997
BS-rapC	16077445; NP_388259.1	Kunst *et al.*, 1997
BS-rapD	16080691; NP_391519.1	Kunst *et al.*, 1997
BS-rapE	16079636; NP_390460.1	Kunst *et al.*, 1997
BS-rapF	16080798; NP_391626.1	Kunst *et al.*, 1997
BS-rapG	16081082; NP_391910.1	Kunst *et al.*, 1997
BS-rapH	16077751; NP_388565.1	Kunst *et al.*, 1997
BS-rapI	16077468; NP_388382.1	Kunst *et al.*, 1997
BS-rapJ	16077351; NP_388164.1	Kunst *et al.*, 1997
BS-rapK	16078951; NP_389772.1	Kunst *et al.*, 1997
BS-rap40	2127181; NP_053779.1	Meijer *et al.*, 1995
BS-rap50	1305508; U55043	Meijer *et al.*, 1995
BS-rap60	10956510; NP_053792.1	Meijer *et al.*, 1995
BS-rapLS20	7429804;S58437	Meijer *et al.*, 1995
BS-rapLS32		Itaya *et al.*, unpublished
BS-rapphi-105	22855023; NP_690783.1	Kobayashi et al., 1998
BH1987	25494545; NP_242853.1	Takami *et al.*, 2000
BH3706	25305088; NP_244573.1	Takami *et al.*, 2000
BH3696	25305091; NP_244563.1	Takami *et al.*, 2000
BH0661	11278763; NP_24152.1	Takami *et al.*, 2000
BH4014	25305090; NP_244882.1	Takami *et al.*, 2000
BH-rapJ	5822759; BAA83915.1	Takami *et al.*, 2000
BT-rapA	30265885; AAM51160.1	Lee *et al.*, 2002
BT-rapB	30265888; AAM41162.1	Lee *et al.*, 2002
BT-rapC	30265891; AAM51164.1	Lee *et al.*, 2002
BT-rapD	30265894; AAM51166.1	Lee *et al.*, 2002
BT-rapE	30265897; AAM51168.1	Lee *et al.*, 2002
BT-rapF	30265900; AAM51170.1	Lee *et al.*, 2002
BA2522	30262514; NP_844891	Read *et al.*, 2002
BA4060	21401435; NP_657420.1	Read *et al.*, 2002
BA3790	21401165; NP_657150	Read *et al.*, 2002
BA3760	30263642; NP_846019	Read *et al.*, 2002

BA1582	21398957; NP_654942	Read *et al.*, 2002
BC2147	30020282; NP_831913	Ivanova *et al.*, 2003
BC0986	30019141; NP_830772	Ivanova *et al.*, 2003
BC1026	30019181; NP_830812	Ivanova *et al.*, 2003
BC3501	30021603; NP_833234	Ivanova *et al.*, 2003
BC3518	30021620; NP_833251	Ivanova *et al.*, 2003

Fig.1


BC3501
BS-RapLS20
BH3696
BH3706
BH4014
BH0661
BH1987
BH-RapJ
BS-RapG
BS-RapK
BS-RapPhi105
BS-RapH
BS-RapA
BS-RapB
BS-RapE
BS-Rap60
BS-Rap40
BS-Rap50
BS-RapJ
BS-RapC
BS-RapF
BS-RapI
BS-RapLS32
BS-RapD
BA1180
BT-RapC
BT-RapF
BA3760
BC2147
BT-RapD
BC0986
BA4060
BA3790
BC3581
BT-RapE
BA0205
BT-RapB
BA1582
BC1026
BT-RapA
BA2522
496
1000
986
999
790
511
1000
1000
983
764
500
591
616
81
950
910
915
996
1000
999
268
1000
951
915
709
565
1000
453
993
394
544
1000
129
1000
929
988
1000
320
0.1

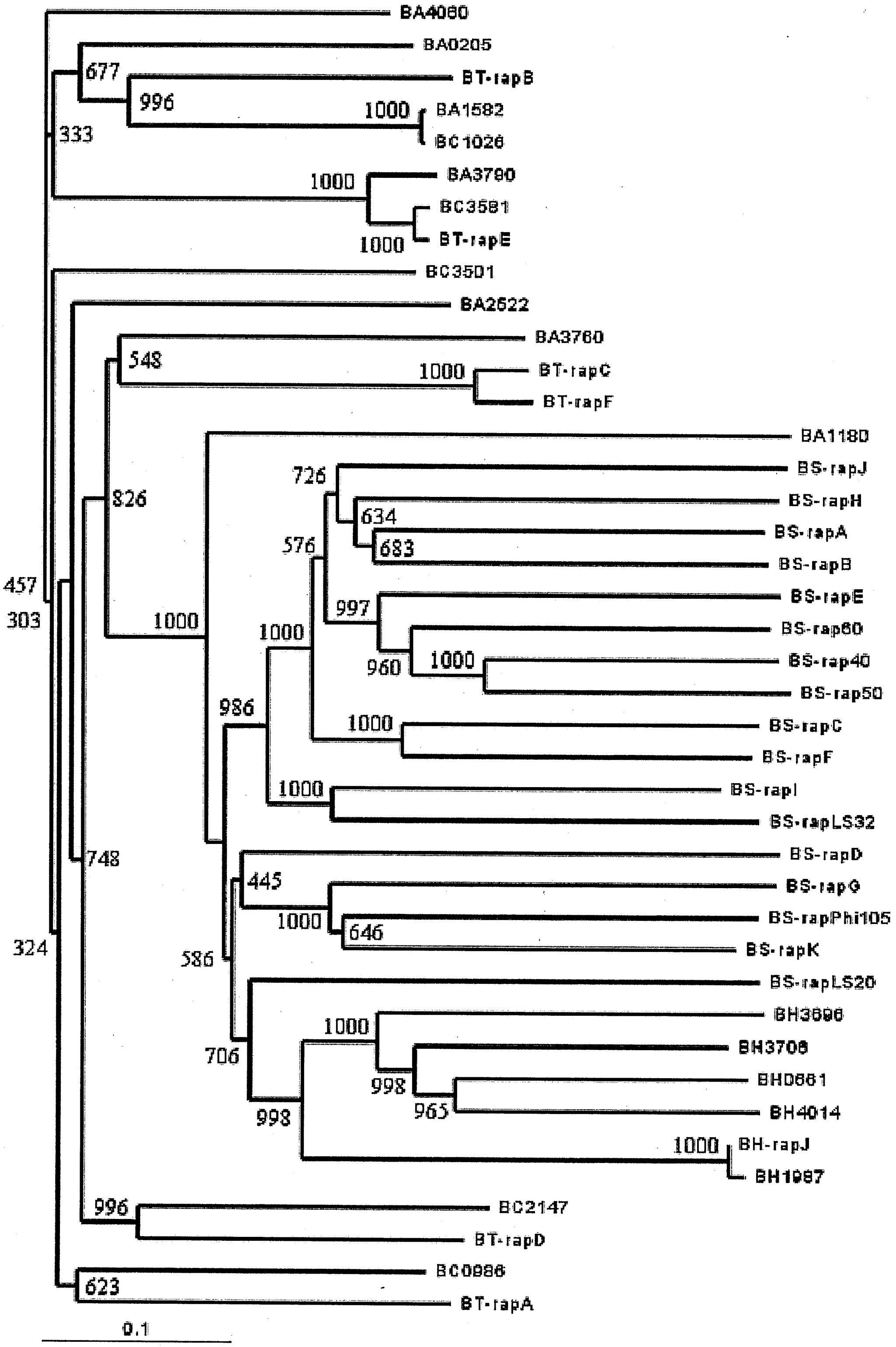
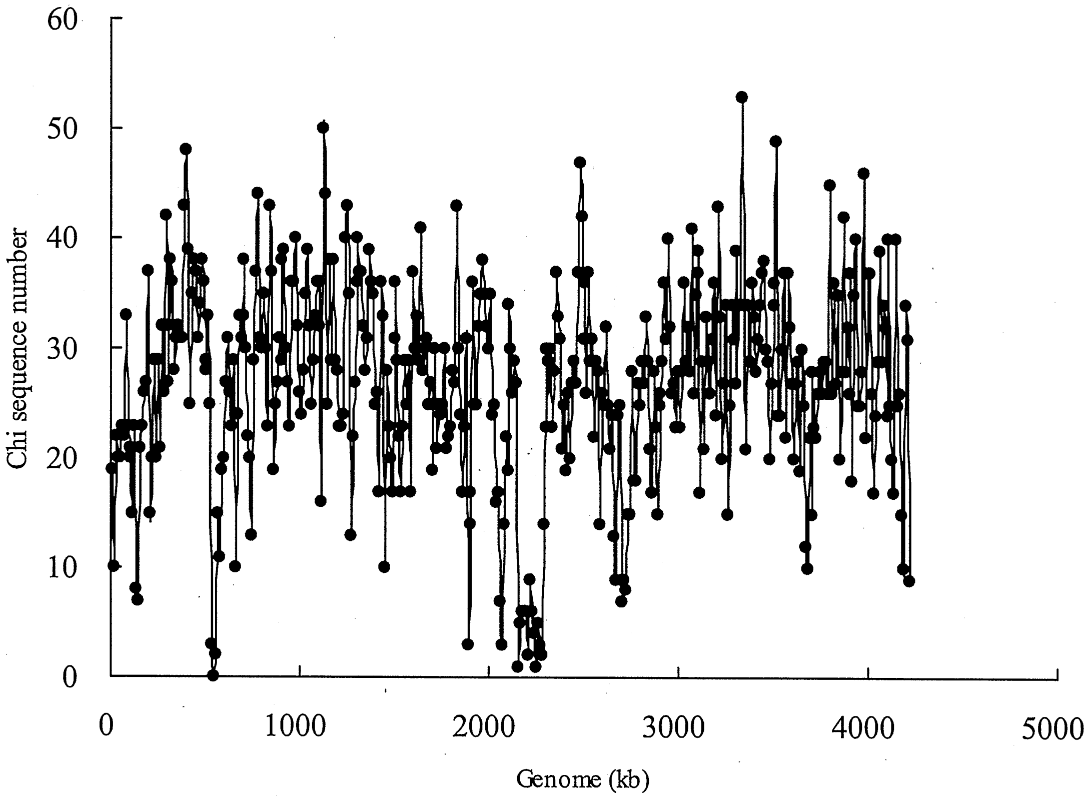

BA4060
BA0205
677
BT-rapB
996
1000
BA1582
333
BC1026
1000
BA3790
BC3581
1000
BT-rapE
BC3501
BA2522
BA3760
548
1000
BT-rapC
BT-rapF
BA1180
726
BS-rapJ
826
BS-rapH
634
BS-rapA
576
683
BS-rapB
457
BS-rapE
997
303
1000
BS-rap60
1000
960
1000
BS-rap40
BS-rap50
986
1000
BS-rapC
BS-rapF
1000
BS-rapI
BS-rapLS32
748
BS-rapD
445
BS-rapG
1000
BS-rapPhi105
646
324
586
BS-rapK
BS-rapLS20
BH3696
1000
BH3708
706
998
BH0661
965
998
BH4014
1000
BH-rapJ
BH1987
996
BC2147
BT-rapD
BC0986
623
BT-rapA
0.1

Fig. 2



60
50
40
30
20
10
0
Chi sequence number
0
1000
2000
3000
4000
5000
Genome (kb)

Table 2 Coordinates (kb) of potential laterally transferred elements on the chromosome of *Bacillus. subitlis* 168

Functions	Low Chi frequency	HMM	Repeats
Prophage 1	200-220	202-220	202-213
Prophage 2	530-570	529-570	555-567
-	570-600	570-600	-
Prophage 3	650-660	651-664	-
Site-specific recombinase	730-750	738-747	-
Multidrug-efflux transporter	815-825	818-822	-
-	1120-1130	1124-1130	-
Prophage 4	1260-1280	1262-1270	-
Prophage PBSX	-	-	-
-	-	1397-1399	1385-1424
-	1440-1450	1442-1447	-
-	1480-1490	1478-1482	-
-	1530-1540	-	-
Prophage 5	1880-1910	1879-1891	-
-	2030-2040	2038-2041	-
Prophage 6	2040-2080	2046-2073	2050-2060
Prophage SP□	2150-2290	2151-2286	-
-	2400-2410	-	-
Penicillin-binding protein 2A gene (*pbp*A)	2580-2590	-	-
Prophage Skin element	2650-2710	2652-2701	2654-2701
Prophage 7	2710-2750	2707-2756	2725-2735
Competence	3250-3260	3253-3257	-
Arsenic resistance regulon	-	3463-3467	3462-3469
-	3600-3610	-	3608-3634
Cell wall synthesis	3660-3680	3658-3685	3665-3672
Poly-glutamic acid synthesis genes	3690-3700	-	-
Wall-associated protein	4020-4030	-	-
ABC transporter	4120-4140	4123-4134	-
ABC transporter	4170-4180	4171-4176	4170-4176
Streptothricin, tetracycline, mercury regul.	4180-4190	4184-4190	4189-4190

The HHM column provides the positions of LGT elements identified by Nicolas *et al* (2002); the repeats column provides the positions long repeats described by Rocha *et al* (1999).

Table 3 Frequency and O/E values for chi sequence of the 10 prophages and the whole genome of *Bacillus subtilis* 168 and bacteriophage □ -105

	Coordinates (bp)	Chi sequence number	O/E value	Chi sequence complement number	O/E value	Chi sequence frequency (1/kb)
Prophage 1	202,000-220,000	18	1.93	10	1.71	1/0.64
Prophage 2	529,000-570,000	12	0.72	11	1.06	1/1.24
Prophage 3	652,000-664,500	6	1.14	6	1.71	1/1.04
Prophage 4	1,263,000-1,279,000	18	2.13	9	1.72	1/0.59
PBSX	1,320,000-1,348,000	57	2.19	36	2.60	1/0.30
Prophage 5	1,879,000-1,900,000	13	1.48	7	1.21	1/1.05
Prophage 6	2,046,000-2,078,000	9	0.74	15	1.26	1/1.33
SP□	2,151,274-2,285,689	6	0.97	50	0.95	1/2.40
Skin	2,265,598-2,700,635	27	1.85	42	1.50	1/0.70
Prophage 7	2,707,000-2,750,000	9	0.86	15	1.13	1/1.79
Whole genome	1-4,214,810	5,681	2.14	5,692	2.12	1/0.37
Phage □ -105	1-39,325	32	1.04	11	0.65	1/0.91

Table 4 Codon usage indices and GC content of the prophages in the chromosome of *Bacillus subtilis* 168

	CAI	CBI	Fop	GC content
Prophage 1	0.377	-0.017	0.324	0.396
Prophage 2	0.385	-0.015	0.330	0.369
Prophage 3	0.371	-0.022	0.326	0.281
Prophage 4	0.391	-0.011	0.337	0.394
Prophage PBSX	0.331	-0.028	0.314	0.462
Prophage 5	0.389	-0.013	0.334	0.373
Prophage 6	0.407	0.040	0.370	0.367
Prophage SP□	0.404	0.039	0.362	0.358
Prophage Skin	0.376	0.036	0.357	0.401
Prophage 7	0.388	0.019	0.350	0.383
Phage □ -105	0.358	-0.017	0.351	0.438