Exploring the evolution of the *Bacillus cereus* group repeat element *bcr1* by comparative genome analysis of closely related strains

Are Klevan, Nicolas J. Tourasse, Fredrik B. Stabell, Anne-Brit Kolstø and Ole Andreas Økstad

Correspondence

Ole Andreas Økstad
o.a.okstad@biotek.uio.no

Department of Pharmaceutical Biosciences (Microbiology), University of Oslo, PO Box 1068 Blindern, 0316 Oslo, Norway

bcr1 is a chromosomal ~155 bp repeated element found uniquely and ubiquitously in the *Bacillus cereus* group of Gram-positive bacteria; it exhibits several features characteristic of mobile elements, including a variable distribution pattern between strains. Here, highly similar *bcr1* elements in non-conserved genomic loci are identified in a set of closely related *B. cereus* and *Bacillus thuringiensis* strains near the *Bacillus anthracis* phylogenetic cluster. It is also shown that *bcr1* may be present on small RNA transcripts in the 100–400 bp size range. *In silico* folding of *bcr1* at the RNA level indicated that transcripts may form a double-hairpin-like structure predicted to have high structural stability. A functional role of *bcr1* at the RNA level is supported by multiple cases of G–U base-pairing, and compensatory mutations maintaining structural stability of the RNA fold. *In silico* folding at the DNA level produced similar predicted structures, with the potential to form a cruciform structure at open DNA complexes. The predicted structural stability was greater for *bcr1* elements showing high sequence identities to *bcr1* elements in non-conserved chromosomal loci in other strains, relative to other *bcr1* copies. *bcr1* mobility could thus be dependent on the formation of a stable DNA or RNA intermediate. Furthermore, *bcr1* elements potentially encoding structurally stable and less stable transcripts were phylogenetically intermixed, indicating that loss of *bcr1* mobility may have occurred multiple times during evolution. Repeated elements with similar features in other bacteria have been shown to provide functions such as mRNA stabilization, transcription termination and/or promoter function. Similarly, *bcr1* may constitute a mobile element which occasionally gains a function when it enters an appropriate chromosomal locus.

Received 22 December 2006

Revised 27 June 2007

Accepted 13 July 2007

INTRODUCTION

The *Bacillus cereus* group of bacteria encompasses six recognized species which are genetically very closely related (Rasko *et al.*, 2005), including *B. cereus*, an opportunistic pathogen involved in two types of food poisoning syndromes and a range of opportunistic infections (Drobniewski, 1993), *Bacillus thuringiensis*, an insect pathogen widely used as a commercial biopesticide (Schnepf *et al.*, 1998), and *Bacillus*

anthracis, the anthrax pathogen, employed as a bioterror agent in 2001 (Jernigan *et al.*, 2001). By comprehensive analysis of complete genome sequences, the species of the *B. cereus* group have been shown to carry a range of repeated elements in their genomes (Tourasse *et al.*, 2006), including the ~155 bp repeated element *bcr1*, originally discovered during piecemeal sequence analysis of *B. cereus* ATCC 10987 and *B. cereus* ATCC 14579 (Økstad *et al.*, 1999), which appears to be ubiquitous and unique to the *B. cereus* group (Økstad *et al.*, 2004). *bcr1* displays characteristics of a mobile element in showing multiple unique chromosomal insertion loci for each particular strain, being flanked by a directly repeated TTTAT motif at both ends, and occasionally interrupting genes. Furthermore, when a chromosomal locus contains a *bcr1* element in one particular strain and the corresponding locus in a different strain lacks *bcr1*, the locus devoid of *bcr1* frequently still contains one copy of the TTTAT direct repeat (Økstad *et al.*, 2004). Thus it has been suggested that during a mobility event the resident chromosomal TTTAT copy could be duplicated upon *bcr1*

Abbreviations: DEPC, diethyl pyrocarbonate; MLST, multilocus sequence typing; MLEE, multilocus enzyme electrophoresis; TBE, Tris/borate/EDTA.

The multiple-sequence alignment results for the full-length repeats examined in this study have been deposited into the EMBL-ALIGN database with the accession number ALIGN_001090.

Two supplementary figures showing the multiple sequence alignment and the predicted RNA folding of the full-length *bcr1* repeats and two supplementary tables listing the chromosomal coordinates of full-length and partial *bcr1* sequences in the strains studied are available with the online version of this paper.

insertion by a transposon-like mechanism involving staggered cuts and a filling-in reaction, or that the second TTTAT copy may be part of the moving *bcr1* element, with integration into the chromosome occurring through a site-specific-recombination-like mechanism (Økstad *et al.*, 2004).

In a previous study, comparing three relatively distantly related strains, *bcr1* was found to display a random chromosomal distribution, shown by a high variability in the number of repeats in each strain and by the low number of *bcr1* elements found at corresponding chromosomal loci (Økstad *et al.*, 2004). This is in strong contrast to the general gene synteny observed between chromosomes in *B. cereus* group strains (Ivanova *et al.*, 2003; Rasko *et al.*, 2004, 2005; Read *et al.*, 2003). To date, *bcr1* has never been detected in extrachromosomal DNA, and the chromosomal localization of *bcr1* elements exhibits a strong bias towards the leading strand of DNA replication (Økstad *et al.*, 2004).

In this paper we describe a four-way computational analysis of newly available complete genome sequence data from closely related strains near the *B. anthracis* phylogenetic cluster, shedding light on *bcr1* evolution. We also present experimental data suggesting that *bcr1* may be part of small RNA transcripts, show that compensatory mutations are frequently introduced into *bcr1*, presumably in order to maintain stability of the folded molecule, and suggest that mobility of *bcr1* is correlated with the predicted stability of the *bcr1* DNA or RNA secondary structure.

METHODS

Genome sequences analysed. Complete genome sequences (including plasmids) from the following set of strains were included in this study: *B. anthracis* Ames (Read *et al.*, 2003), originally isolated from a dead cow in Texas in 1981 (GenBank accession number AE016879); *B. cereus* ATCC 10987 (Rasko *et al.*, 2004), a dairy strain isolated from spoiled cheese in Canada in the 1930s (Herron, 1930; GenBank accession number AE017194); *B. cereus* E33L [formerly known as *B. cereus* Zebra Killer (Han *et al.*, 2006)], originally isolated from the carcass of a zebra in Etosha National Park, Namibia (GenBank accession number CP000001); *B. thuringiensis* 97-27 [subsp. *konkukian* serotype H34 (Han *et al.*, 2006)], a strain isolated from the leg wound of a 28-year-old French soldier following a landmine explosion, and causing severe tissue necrosis (Hernandez *et al.*, 1998; GenBank accession number AE017355). All strains analysed map close to the *B. anthracis* phylogenetic cluster by multilocus sequence typing (MLST) and multilocus enzyme electrophoresis (MLEE) analyses (Helgason *et al.*, 2000b, 2004), and have been sequenced to completion [for a complete overview of currently available *B. cereus* group genome sequence data see Tourasse *et al.* (2006)]. In addition, *B. cereus* strains AH818 and AH820, isolated from cases of human periodontitis (Helgason *et al.*, 2000a), were analysed for *bcr1* loci shared with *B. anthracis* by PCR screening (see below).

Whole-genome alignments. Multiple Genome Aligner (MGA; Hohl *et al.*, 2002) was employed for full-length genome comparisons. All chromosome sequences were indexed according to the MGA guidelines prior to computation, using the mkvtree program with options '-dna -lcp -suf -tis'. The mga alignment program was set to

run with length thresholds of 50 bp and 20 bp (-l 50 20), a maximum gap length of 3000 bp (-gl 3000), to always recurse into gaps (-always), and to use CLUSTAL W to close short gaps (-clustalw). Both aligned (-alignedseqs) and unaligned (-gaps) sequences were output. All four fully sequenced strains used in the study were compared in pairs, generating six comparisons. For each pairwise comparison, all aligned sequences were concatenated and the sequence identity between the concatenated sequences was calculated after removal of all positions with gaps (i.e. insertions and deletions). This value was taken to represent the average identity between homologous regions in any two genomes.

Iterative BLAST searches. To identify *bcr1* elements in the whole-genome sequences, a BLASTN (Altschul *et al.*, 1997) search procedure was executed essentially as described previously (Økstad *et al.*, 2004) with the following modifications. To increase the sensitivity of the search, a dual iterative BLASTN routine was employed, in which two parallel runs of BLASTN were conducted at each step, one using lowered gap penalties (opening cost $G=1$ and extension cost $E=1$), the other using increased reward for nucleotide match (match reward, $r=2$). The two BLASTN output files were combined and all identified full-length *bcr1* sequences [repeats of length 120 bp or more, as defined in Økstad *et al.* (2004)] from the strains were used as seeds in a subsequent BLASTN search against all strains in the analysis. This process was repeated until no further full-length repeats could be identified. The complete genome of *B. cereus* ATCC 14579 (type strain, GenBank accession number AE016877) was included in the iterative BLAST search, in order to provide additional seed sequences. This strain, however, belongs to a different phylogenetic subgroup (Helgason *et al.*, 2004), and was thus not included in the comparative analyses. The identity of all full-length repeats was verified by multiple-sequence alignment using CLUSTAL W (Thompson *et al.*, 1994). The alignment was manually checked and corrected using SEAVIEW (Galtier *et al.*, 1996), and deposited into the EMBL-ALIGN database (Lombard *et al.*, 2002; accession number ALIGN_001090). Partial repeats in the genomes (defined as those ranging from 30 to 119 bp; Økstad *et al.*, 2004) were identified by a dual non-iterative BLASTN routine, one using opening cost $G=1$ and extension cost $E=1$, the other using match reward $r=2$, as above. All 218 full-length *bcr1* identified in the iterative BLASTN routine described above were used as seeds (*B. cereus* ATCC 14579 included).

Comparative analysis of chromosome regions flanking *bcr1* repeats. In order to investigate locus conservation of *bcr1* elements, 2 kb of DNA sequence was extracted from both sides of each repeat, for both full-length and partial elements. The sequences were subsequently used as input in an all-against-all BLASTN search with default parameters, using the BLAST-enhancement tool MSPcrunch (Sonnhammer & Durbin, 1994) and CLUSTAL W for sorting and visualization of hits. Repeats (full-length or partial) for which both flanking regions were conserved in different strains were considered as being at a conserved genomic locus.

Comparative analysis of *bcr1* sequence conservation. To identify closely related *bcr1* sequences in the sequenced *B. cereus* group genomes, an all-against-all BLASTN comparison was performed using the identified full-length *bcr1* sequences and default parameters, with the exception of no filtering of low-complexity regions. The output was converted to a format suitable for GenomePixelizer, using a GenomePixelizer parser (Kozik *et al.*, 2002; <http://www.atgc.org/GenomePixelizer/>), retaining BLASTN hits with an expectation value (E-value) lower than 1×10^{-30} , a normalized nucleotide sequence identity >90%, and an alignment length of 120 bp or greater.

Phylogenetic analysis of *bcr1* sequences. A phylogenetic analysis was performed for all 159 full-length *bcr1* repeats identified. The sequences were aligned using CLUSTAL W (Thompson *et al.*, 1994,

1997) followed by manual editing in SEAVIEW (Galtier *et al.*, 1996). The final alignment (189 bp including gaps) was converted into MEGA format. Using MEGA (Kumar *et al.*, 2001), a tree was constructed by the neighbour-joining method (Saitou & Nei, 1987), with the K80 substitution model allowing for transition/transversion substitution rate bias (Kimura, 1980) and gaps treated by pairwise deletion. Similar methods were used to construct trees for two different subsets of *bcr1* elements.

A set of three *bcr1* repeats located in conserved loci in *B. cereus* strains ATCC 10987, E33L, AH818 and AH820, *B. thuringiensis* 97-27 and *B. anthracis* Ames were used for phylogenetic analysis; *bcr1* sequences from the same strain were concatenated. After block extraction using the Gblocks program (http://molevol.ibmb.csic.es/Gblocks_server/) (Castresana, 2000) with default parameters, reducing the alignment from 470 bp to 432 bp, the final alignment was converted to the NEXUS file format and input into MrBayes (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003) for phylogenetic analysis using maximum-likelihood-based Bayesian inference. A total of 1 000 000 generations were executed using a burn-in value of 100 000, and a sampling frequency of 1000. The likelihood was computed using a two-parameter substitution model allowing for transition/transversion substitution rate bias (nst=2, similar to HKY85), with invariant+gamma distribution (rates=invgamma) allowing variable substitution rates among sites, and with base frequencies estimated by the program.

MrBayes consensus trees were visualized using TreeView (Page, 1996). Trees based on all datasets were also constructed in MEGA (Kumar *et al.*, 2001) using the K80 substitution model (Kimura, 1980) and the neighbour-joining method (Saitou & Nei, 1987) and showed branchings nearly identical to those constructed with MrBayes.

DNA and RNA secondary structure predictions. DNA and RNA secondary structures and thermodynamics were predicted using the MFOLD package, version 3.1 (Mathews *et al.*, 1999; Zuker, 2003), with default parameters. All 159 full-length *bcr1* sequences from the complete *B. cereus* group genomes were folded as circular or linear RNA, and circular or linear DNA, with both terminal TTTAT direct repeats included in the structure. For a given *bcr1* sequence the folding with the minimum free energy (ΔG value) was selected. As

folding results for the circular and linear forms showed only very slight variations, results of circular folding were used in the analysis. Furthermore, an alignment of the four largest inverted sequences (two pairs) within *bcr1* was constructed, after removing repeats harbouring significant deletions (55 out of 159 repeats). Compensatory mutations located within these inverted repeats were detected by visual inspection.

PCR amplification and nucleotide sequencing of conserved *bcr1* copies from *B. cereus* strains AH818 and AH820.

PCR primers for the detection of conserved *bcr1* repeats in the unsequenced *B. cereus* strains AH818 and AH820 were designed using the genome sequence of *B. anthracis* Ames and Primer3 software (Rozen & Skaletsky, 2000). Primers were positioned in regions flanking each full-length *B. anthracis* Ames *bcr1* copy (sequences given in Table 1). NetPrimer (Premier Biosoft International) was employed for additional control of primer sequences. PCRs were conducted in a total reaction volume of 50 μ l, with 0.2 mM of each deoxynucleoside triphosphate (Promega), 0.6 μ M of each primer (Invitrogen), 50 ng genomic template DNA and 1 U DyNAzyme (Finnzymes). PCR was run with an initial denaturation step of 5 min at 94 °C, followed by 40 cycles of 1 min denaturation at 94 °C, 1 min annealing at 57 °C and 1 min elongation at 72 °C. A final elongation step at 72 °C for 7 min was included. All PCR products were sequenced using standard methods.

RNA isolation. Total RNA from *B. cereus* ATCC 14579, *B. cereus* AH820, *B. cereus* ATCC 10987 and *B. thuringiensis* 97-27 was isolated using a hot acid phenol-based procedure, from cultures in the mid-exponential growth phase, grown at 30 °C, 240 r.p.m., in Luria-Bertani (LB) medium. Each culture sample was added to an equal amount of preheated acidic phenol (Ambion, pH 4.5), and lysed for 10 min in a 100 °C water-bath with occasional mixing. After cooling, the suspension was centrifuged (12 000 g, 20 min), and the water phase was removed and mixed with an equal volume of phenol/chloroform/isoamyl alcohol (25:24:1, Ambion). The suspension was centrifuged (12 000 g, 20 min), and RNA was ethanol precipitated from the water phase following the addition of 0.3 M sodium acetate (Merck), and dissolved in DEPC-treated H₂O (Ambion). Total RNA was treated with DNase I (FPLC pure, Amersham) and repurified by

Table 1. Primers employed in screening for *bcr1* loci from *B. anthracis* Ames (1F–15R) that are conserved in *B. cereus* AH818 and AH820

All primers were designed using the genome sequence of *B. anthracis* Ames.

Repeat	Forward primer (5'–3')	Reverse primer (5'–3')	T _m (°C)
1F	TGTTATTCCTGGTGGGAAAAG	TCAGTTGAAATATGATACACCTGTTG	59.00/59.57
2R	GAACAACATTACTTTCAACAAATCAGA	AAAGAAAATGAAGAGTTGATCGAAG	59.88/59.37
3F	GGCAAAAAGGCACACTAACC	TGTCAGTGAATCACCTAACC	59.61/58.96
4F	TCACGAGAAGAGTATAAGGGTGAA	TGGCAAAAGAACAAAACGAA	59.33/59.32
5F	GAAAATCAAATTGCAAAAGAAGAAG	ATTTTGCCATCTCTTCATTGAG	59.70/58.33
6R	CCAAACTCGTCTCTTTCTCAA	TGGTGGTATCGGACAATCAA	58.63/59.77
7F	CGCATATTGCTTCTTTCCA	CCTCAAGTTCCTTTTCATCAGC	59.02/58.92
8F	TGTTATTCCTGGTGGGAAAAG	TCAGTTGAAATATGATACACCTGTTG	58.94/59.39
9R	CATTCTCTTTTGTCTTTTGCAT	CCAGAAAAGAACTTGTGAGGA	59.80/59.37
10F	GAAGGGTAAAGTGTACGAGCA	GCGGTGGTACTTTTGAATTTG	58.45/59.51
11R	GATAGATAAAGCCCGGAAGC	TCCATATGTTTTCGGTTGAGG	58.81/59.81
12R	GCGTAGATAAACAAGTGAAGAAG	TACCTGGCATGATGAGATCC	58.39/58.46
13R	TTCATCGATCGGGTTATTCA	AAAAGCGTTTCGTTTACGTT	58.92/59.79
14R	CCGATGATTGGATACAATTTCTT	TCAGTGCTTGTTGCCCTAAA	59.24/59.46
15R	CCGAGCCACTCTCAGGATA	GATTCAGAGTCCTTCGCTGT	58.92/58.54

hot acid phenol and phenol/chloroform/isoamyl alcohol extraction and ethanol precipitation, as described above.

Northern blotting and riboprobe hybridization. For each sample, total RNA (25 µg) in formaldehyde loading buffer (Ambion) was heated for 10 min at 65 °C and loaded onto a denaturing polyacrylamide gel [7 M urea, 6% polyacrylamide, 1 × Tris/borate/EDTA (1 × TBE: 90 mM Tris/borate (Sigma-Aldrich), 2 mM EDTA (Sigma-Aldrich)), 120 V]. Following electrophoresis, RNA was electroblotted overnight (0.5 × TBE, 18 V, 4 °C) to a nylon membrane (Hybond-N+, Amersham) and fixed by UV-irradiation.

Probes for hybridization were designed from the plus and minus strands of a *bcr1* element located between genes BC3105 and BC3106 of *B. cereus* ATCC 14579 (genomic coordinates 3069711–3069552 in AE016877). The *bcr1* element used as the template had been cloned into pUC19 vector before PCR (Økstad *et al.*, 2004), and corresponded to the element originally used as the seed sequence for the iterative BLAST searches. The DNA template for riboprobe construction was synthesized by PCR using primers from each end of the *bcr1* element. A T7 promoter and a *Bam*HI restriction site were incorporated in the 5' end of each primer (Invitrogen). The primer sequences were as follows: primer 721: 5'-TAATACGACTCACTAT-AGGGAGA CCC GGA TCC GGC AGT AAG ACC TCC ACC TC-3', primer 722: 5'-TAATACGACTCACTATAGGGAGA GCG GGA TCC ATA AAG TGA AAC TTT AAT CAG-3' (*Bam*HI restriction sites underlined, T7 promoter sequence in italics). Primers corresponding to 721 and 722, but not containing the T7 promoter sites, were also synthesized (Invitrogen), and two PCR reactions were set up, in which each reaction was run with only one primer carrying the T7 promoter sequence. The PCR products were purified from 3% agarose gels (Nusieve GTG, Cambrex) using the Qiaquick gel extraction kit (Qiagen), and riboprobes representing each of the two *bcr1* strands were synthesized in separate reactions from the single T7 promoter site in each PCR product, using T7 RNA polymerase and employing the Maxiscript kit (Ambion) with 100 µCi (3.7×10^6 Bq) [α -³²P]UTP [800 Ci mmol⁻¹ (29.6 TBq mmol⁻¹), 20 mCi ml⁻¹ (740 MBq ml⁻¹)] and 0.5 mM unlabelled ATP, CTP and GTP, following the suppliers' instructions. Following DNase treatment (DNase I, FPLC pure, Amersham), the full-length riboprobes were purified from a denaturing polyacrylamide gel (7 M urea, 6% polyacrylamide). Hybridization was performed with Perfecthyb Plus (Sigma) as instructed by the supplier, with the highest stringency wash. Membranes were exposed overnight, and signals were visualized using a phosphorimager (STORM 860, Molecular Dynamics).

RESULTS

Patterns of *bcr1* distribution within *B. cereus* group genomes from the same phylogenetic group

The *bcr1* sequence has been detected previously (Økstad *et al.*, 2004) as 79, 12 and 54 chromosomal copies, respectively, in *B. cereus* ATCC 10987, *B. anthracis* Ames and the more distantly related *B. cereus* ATCC 14579 (type strain), which is not part of the same phylogenetic cluster (Helgason *et al.*, 2000b, 2004). In this work *bcr1* distribution was analysed in four *B. cereus* group strains mapping in or close to the *B. anthracis* cluster by MLST (Fig. 1a), and for which complete, closed genome sequences were available, using an improved identification method employing an iterative dual-BLAST procedure. The

total numbers of full-length *bcr1* identified in the four chromosomes by the updated method were as follows: *B. cereus* ATCC 10987, 84; *B. cereus* E33L, 31; *B. thuringiensis* 97-27, 29; and *B. anthracis* Ames, 15. As observed from a multiple-sequence alignment, the newly identified repeats in *B. cereus* ATCC 10987 and *B. anthracis* Ames relative to those analysed by Økstad *et al.* (2004) clearly belonged to the *bcr1* family (see Supplementary Fig. S1, available with the online version of this paper). In addition, 493 partial *bcr1* elements were identified altogether, with the following strain distribution: *B. cereus* ATCC 10987, 212; *B. anthracis* Ames, 91; *B. thuringiensis* 97-27, 93; *B. cereus* E33L, 97 (coordinates of full-length and partial *bcr1* elements are provided in Supplementary Tables S1 and S2, available with the online version of this paper). The partial elements were heterogeneous, representing various regions of the full-length elements.

By direct sequence comparison of 2 kb regions flanking each side of the identified *bcr1* elements, we investigated the degree of *bcr1* locus conservation in the closely related strains. The results showed that six *bcr1* loci were conserved in all four strains (locus numbers 1, 2, 6, 7, 23 and 26 in Table 2). Interestingly, the number of *bcr1* loci (full-length and partial) of each strain shared with *B. anthracis* (by pairwise comparison) was largely inversely proportional to the phylogenetic distance of the strain from *B. anthracis* (*B. thuringiensis* 97-27, 14; *B. cereus* E33L, 10; *B. cereus* ATCC 10987, 7) (Table 2, Fig. 1a). This trend was further supported by PCR screening of two clinical *B. cereus* strains isolated from cases of human periodontitis (AH818 and AH820; Helgason *et al.*, 2000a), which are among the most closely related *B. cereus* strains to *B. anthracis* known to date (Helgason *et al.*, 2004; Fig. 1a). Using primers to neighbouring regions flanking each full-length repeat element identified in *B. anthracis*, it was shown that *B. anthracis* Ames shares 12 of its 15 full-length repeats with *B. cereus* AH818 and AH820.

bcr1 RNA transcript mapping

bcr1 elements are highly overrepresented in intergenic regions (Økstad *et al.*, 2004). Northern blotting has previously shown that *bcr1* elements are part of longer transcripts in the size range 1.0–2.5 kb (Økstad *et al.*, 2004), indicating co-transcription of *bcr1* with neighbouring genes. To investigate whether *bcr1* could also be part of small RNA transcripts, distinct from those of neighbouring genes, total RNA isolated from cells during mid-exponential growth (4.5 h) was separated by PAGE, electroblotted and hybridized with a *bcr1* riboprobe, employing both strands in separate hybridization reactions. This revealed that *bcr1* can be part of transcripts in the size range 100–400 bp (Fig. 2; the riboprobe was constructed employing the T7 promoter site in primer 722, which contains a 21 nt region complementary to a conserved end of *bcr1*), which is compatible with the sizes of full-length *bcr1* elements (range: 120–163 bp). Interestingly, the hybridization pattern showed variability between strains (Fig. 2); *B.*

Fig. 1. (a) Phylogenetic relationships among selected *B. cereus* group strains, including those used in this study, based on multilocus sequence typing (MLST). *B. cereus* strains are shown in yellow, *B. thuringiensis* strains in green, and *B. anthracis* in red. The MLST method was based on seven housekeeping loci, following the scheme of Tourasse *et al.* (2006). The scale bar is in mean number of nucleotide substitutions per site. (b) Phylogenetic tree constructed using maximum-likelihood-based Bayesian inference, based on the concatenated sequence of the three conserved *bcr1* loci (loci 1, 2 and 7 in Table 2) found in all six strains close to the *B. anthracis* cluster (*B. cereus* ATCC 10987, *B. cereus* E33L, *B. thuringiensis* 97-27, *B. cereus* AH818, *B. cereus* AH820 and *B. anthracis* Ames). The scale bar indicates a nucleotide sequence variation of 1%.

cerus ATCC 10987, which has the highest *bcr1* copy number among the strains investigated, also had the highest number of differently sized small RNA molecules hybridizing with the probe, and in addition showed significantly stronger bands. This strain also contained a hybridizing RNA of approximately 100 bp, which could originate from transcription of a partial repeat element (Fig. 2).

It is known that *bcr1* elements may be located on the plus or minus DNA strand relative to their neighbouring gene, thus allowing the potential co-transcription of *bcr1* in both orientations within the cell. Hybridization using the riboprobe from the opposite *bcr1* strand (T7 promoter site in primer 721; primer containing a 20 nt region complementary to *bcr1*) produced bands in the same size range as those produced using primer 722 (data not

Table 2. Conserved *bcr1* loci in four complete *B. cereus* group genomes from or phylogenetically close to the *B. anthracis* cluster, based on the comparison of 2 kb of DNA sequence flanking each side of the full-length *bcr1* repeats identified in each respective strain.

Conserved locus no.	<i>B. anthracis</i> Ames*	<i>B. thuringiensis</i> 97-27*	<i>B. cereus</i> E33L*	<i>B. cereus</i> ATCC 10987*
1†	1F	1F	1F	1F
2†	2R	2R	2R	2R
3	3F	3F		
4	6R	10R		
5	7F	11F	6F	
6	9R	12R	9R	P84R
7†	10F	13F	10F	38F
8	11R	20R	P50R	
9	12R	25R		
10	15R	29R		84R
11		4F	3F	5F
12		6F	P15F	16F
13			5F	29F
14		15F		45F
15	P45F	16F	12F	
16			15R	51R
17		23F	17F	53F
18		24R	19R	P132R
19			20R	58R
20	P73R	26R	24R	
21		27R	29R	P194R
22		28R		81R
23	P74R	P77R	25R	69R
24			27R	72R
25			28R	76R
26	P90F	P90F	31F	83F

*Repeat names ending with F and R correspond to repeats located on the forward and reverse DNA strand, respectively. Names starting with a P indicate partial *bcr1* copies.

†These repeats were also conserved in *B. cereus* strains AH818 and AH820. Given that they were full-length repeats and conserved in the same genomic loci in all six strains, they were used as phylogenetic markers in Fig. 1(b).

shown), some of which could potentially be common to the two hybridization experiments. The observed hybridization patterns may be due to the inverted repeat character of the *bcr1* elements and/or to *bcr1* being transcribed in both orientations.

In silico RNA secondary structure prediction of *bcr1* elements

To investigate secondary structure stability and predict the RNA folding for each *bcr1* element, all 159 full-length *bcr1* copies in the genomes of *B. cereus* ATCC 10987, *B. cereus* E33L, *B. thuringiensis* 97-27, and *B. anthracis* Ames were analysed using MFOLD (Mathews *et al.*, 1999; Zuker, 2003).

Fig. 2. Northern hybridization of total RNA isolated from vegetative cells of *B. cereus* and *B. thuringiensis*, employing a *bcr1*-specific riboprobe transcribed from the *bcr1* PCR product with the T7 promoter in primer 722. The gel used for RNA separation contained 6% polyacrylamide, to preferentially separate small RNA. Lanes: 1, *B. cereus* AH820; 2, *B. thuringiensis* 97-27; 3, *B. cereus* ATCC 10987.

Computed minimum folding energies (ΔG) were found to vary from -17.5 to -87.0 kcal mol $^{-1}$ (-73.2 to -364.0 kJ mol $^{-1}$), with an average of -57.1 kcal mol $^{-1}$ (-238.9 kJ mol $^{-1}$) [see Supplementary Fig. S2, available with the online version of this paper; for linear folding, the range was -19.3 to -87.0 kcal mol $^{-1}$ (-80.6 to -364.0 kJ mol $^{-1}$), average -57.9 kcal mol $^{-1}$ (-242.3 kJ mol $^{-1}$)]. The most stable secondary structure was predicted to constitute a double-hairpin-like fold, as exemplified by *bcr1* copy 77R from *B. cereus* ATCC 10987, due to the presence of internal inverted repeat motifs [Cereus_10987_77R, $\Delta G = -87.0$ kcal mol $^{-1}$ (-364.0 kJ mol $^{-1}$); Fig. 3a]. Several cases of G–U base pairing were observed in the secondary structures, in line with the observation that *bcr1* is transcribed (Fig. 2) and was predicted to fold as RNA (Fig. 3a, Supplementary Fig. S2). Folding *bcr1* at the DNA level (ssDNA) produced similar structures (data not shown). *bcr1* may thus have the capacity to form a cruciform-like structure at open DNA complexes, for instance during replication or transcription, when DNA strands may be separated into single strands around *bcr1*. The cruciform structure would then be constituted by one double hairpin forming at each single strand of *bcr1* in the open DNA complex, and the double-stranded DNA extending from each side of the pair of double hairpins. Such a cruciform structure would bring the TTTAT-termini into close proximity in three-dimensional space, and one could hypothesize the recognition of such a structure by a putative transposing or recombination enzyme supplied *in trans* from another genomic locus (Økstad *et al.*, 2004).

Some *bcr1* elements found to have below average ΔG values still deviated from the double-hairpin-like shape, while other elements displayed ΔG values higher than the average and were still predicted to form the double-hairpin structure. The latter group (32 full-length repeats) could in most cases be explained by shorter *bcr1* sequences (120–130 bp) and/or a higher number of bulges and/or a higher AT content in the sequence. With only two exceptions, these repeats displayed less than 95 % BLASTN identity to other *bcr1* repeats in the chromosomes included in this study. The smaller group (8 repeats), comprising *bcr1* elements with low ΔG values and deviating folding structures, could be a result of a limited number of crucial mutations disturbing the double-hairpin structure. Despite their deviating structures (Supplementary Fig. S2), some of these repeats still displayed high sequence identity to other *bcr1* elements. No correlation was found between the location of the deviating *bcr1* elements and the function of their neighbouring genes.

The importance of maintaining a stable *bcr1* secondary structure was corroborated by the frequent observation of compensatory mutations when comparing the two largest pairs of inverted repeats within *bcr1* (Fig. 3b). As a consequence, full-length repeats harbouring large deletions and/or nucleotide substitutions were predicted to form less stable secondary structures, in some cases with different conformations, in particular when a deletion fell within one of the inverted repeat regions forming the stems (Supplementary Figs S1 and S2). When sorting repeats according to their lengths we observed a sharp shift, resulting in a subdivision of the repeats, mainly into either ~155 bp or ~125 bp variants (Fig. 4; Supplementary Fig. S1). This was largely due to a 33 bp deletion near the 3' end which spans internal stem B1–B2 (Fig. 3a) and is present in 27 out of the 159 *bcr1* repeats, but deletions of similar sizes could also be detected in other regions (Supplementary Fig. S1). The 33 bp deletion makes the B1–B2 stem 16 bp shorter (as compared to *Cereus_10987_77R*), and sustains its integrity but not its nucleotide sequence symmetry (e.g. compare structures of *Cereus_10987_77R* in Fig. 3a and *Cereus_10987_49R* in Supplementary Fig. S2; corresponding to *Bce_77R* and *Bce_49R* aligned in Supplementary Fig. S1). The *bcr1* repeats in the ~155 bp size range clearly exhibited a generally higher structural stability ($\Delta G < -50$ kcal mol⁻¹; -209.2 kJ mol⁻¹) than those in the ~125 bp range ($\Delta G > -50$ kcal mol⁻¹) (Fig. 4).

Fig. 4. Relationship between length and predicted folding energy of the 159 full-length *bcr1* elements from *B. anthracis* Ames, *B. cereus* E33L, *B. cereus* ATCC 10987 and *B. thuringiensis* 97-27. The full-length elements mainly centred around two size categories, with approximate sizes of 125 bp and 155 bp. The *bcr1* elements indicated in red and blue were predicted to exhibit a higher and lower than average secondary structure stability, respectively. Although there was a clear correlation between *bcr1* length and structural stability, a considerable number of *bcr1* elements in the 155 bp category were predicted to exhibit structural stability below the average for full-length elements.

***bcr1* copies with high predicted stability frequently share high sequence identity with multiple repeats in non-conserved loci**

Whole-genome sequencing has shown that *B. cereus* group genomes are generally highly syntenic (Rasko *et al.*, 2004, 2005). To calculate the average sequence identity of shared chromosome regions between the four strains in this study for which complete genome sequences exist, a pairwise comparison of the chromosome sequences was performed using Multiple Genome Aligner (MGA; Hohl *et al.*, 2002). Pairwise sequence identity scores varied from 94.2% for *B. cereus* ATCC 10987 and *B. thuringiensis* 97-27, to 97.8 % for *B. anthracis* Ames and *B. thuringiensis* 97-27 (Table 3), and correlated well with the strain phylogeny reconstructed by MLST (Fig. 1a). *bcr1* repeats are known to exhibit a wide

Fig. 3. (a) RNA secondary structure of *bcr1* predicted by *in silico* folding using MFOLD (Mathews *et al.*, 1999; Zuker, 2003), exemplified by repeat 77R from *B. cereus* ATCC 10987 ($\Delta G = -87.0$ kcal mol⁻¹; -364 kJ mmol⁻¹). Brackets mark the inverted repeats aligned in (b). (b) Multiple alignment of the two largest stem-loops (A1–A2 and B1–B2) predicted by *in silico* folding of *bcr1*. The marked regions display compensatory mutations (I, II and III; linked by arrows) observed within the folded stems. These compensatory mutations constitute nucleotide substitutions in the *bcr1* sequence which serve to maintain the secondary structure, e.g. so that when a C→T base change occurs in a site X within a stem region, a simultaneous G→A change is observed in the site which is predicted to pair with site X. Some bases within the fold, which are predicted to be part of an internal loop within the A1–A2 stem, appeared to be less conserved.

Table 3. Multiple genome aligner (MGA) analysis of whole genome sequences, by pairwise comparison of *B. anthracis* Ames (Ba), *B. thuringiensis* 97-27 (Bt), *B. cereus* E33L (BcE33L), and *B. cereus* ATCC 10987 (Bc10987).

Pairwise genome sequence comparison	Coverage (%)	Identity (%)
Ba/Bt	92.7/93.3	97.8
Ba/BcE33L	93.7/93.8	97.5
Ba/Bc10987	87.4/87.8	94.4
Bt/BcE33L	93.2/92.4	97.2
Bt/Bc10987	88.3/88.1	94.2
BcE33L/Bc10987	87.3/88.1	94.5

range of pairwise sequence identity (Økstad *et al.*, 2004), and the full-length *bcr1* copies identified in this study were found to exhibit between 44.0 % and 100 % sequence identity, with an average of 82.8 %. An all-versus-all BLAST

comparison between the 159 full-length *bcr1* sequences from *B. cereus* ATCC 10987, *B. cereus* E33L, *B. thuringiensis* 97-27 and *B. anthracis* Ames revealed that, in addition to a remarkably high sequence conservation of repeats located in corresponding genomic loci (79.0–100 % sequence identity, average 94.4 %; Table 2), there were numerous cases of *bcr1* sequence identity above the genomic DNA average between repeats located in non-corresponding loci in different genomes (Fig. 5). Repeat 14R from *B. anthracis* shows 98 % identity to *B. cereus* ATCC 10987 repeats 33F, 46R, and 65R, above average for shared chromosomal regions (94.4 %, Table 3). These *bcr1* copies are all located in different genomic loci and in one case in the opposite relative orientation (Fig. 5). *B. anthracis* Ames 10F, *B. thuringiensis* konkukian 13F and *B. cereus* ATCC 10987 75R also display 98–99 % identity. Unexpectedly, the high degree of sequence similarity, displaying 98 % sequence identity or more between non-corresponding repeats, is most apparent between *B. cereus* ATCC 10987 and *B.*

Fig. 5. DNA sequence comparison of full-length *bcr1* elements in the genomes of *B. anthracis* Ames (line 1), *B. thuringiensis* 97-27 (line 2), *B. cereus* E33L (line 3) and *B. cereus* ATCC 10987 (line 4), using BLASTN. The figure was drawn using GenomePixelizer. The *bcr1* elements are indicated by rectangles. Lines connecting the *bcr1* elements indicate *bcr1* sequence identities of >95 % (yellow) and >98 % (purple), which are identity values above the genomic average for each pairwise comparison of strains. The *bcr1* repeats with a higher predicted structural stability (lower ΔG value) than the average for all full-length elements are indicated by red rectangles, while those with lower than average structural stability are indicated in blue (see Fig. 4). Elements marked above the horizontal lines representing the chromosome are on the forward strand, while those below the lines are on the reverse strand. *B. anthracis* repeat 14R is labelled with an asterisk (line 1).

anthracis Ames, the least phylogenetically related among the four strains included in the comparison. Furthermore, within each of the four genomes, multiple *bcr1* elements exhibited sequence identities of 94 % or more. This was most pronounced in *B. cereus* ATCC 10987, with 10 *bcr1* copies showing 97 % or more sequence identity to other elements in the genome.

When the folding energy was considered, an interesting pattern emerged: the repeats showing identity higher than the chromosomal average to multiple other repeats in non-conserved locations generally had higher than average predicted secondary structure stability ($\Delta G < -57.1$ kcal mol⁻¹; -238.9 kJ mmol⁻¹) indicated in red in Figs 4 and 5), while repeats with less than average folding stability ($\Delta G > -57.1$ kcal mol⁻¹) tended to be located in corresponding loci and/or share high sequence

identity with only a few other repeats (indicated in blue in Figs 4 and 5). Taken together, this could suggest that *bcr1* copies with multiple high-sequence-identity connections are functionally mobile copies and that proper folding, at either the DNA or RNA level, is important for the mobility function. An unrooted phylogenetic tree of all 159 full-length *bcr1* repeats from the strains close to the *B. anthracis* cluster was constructed by the neighbour-joining method (Saitou & Nei, 1987) with the K80 substitution model (Kimura, 1980), demonstrating that repeat elements with a higher than average predicted structural stability are intermixed with those exhibiting less than average stability (Fig. 6), and that 'high-energy' *bcr1* repeats do not exhibit a propensity to cluster phylogenetically.

To further investigate the relationship between folding and putative mobility of *bcr1*, separate phylogenetic trees were

Fig. 6. Unrooted neighbour-joining tree of all 159 full-length *bcr1* repeats. The K80 (Kimura, 1980) nucleotide substitution model was used to compute evolutionary distances. Red circles and blue triangles designate *bcr1* elements with predicted folding stabilities above and below the average value, respectively (see Fig. 4). Notably, some of the *bcr1* repeats from conserved genomic loci form outgroups in the analysis, and are indicated by asterisks.

Fig. 7. Unrooted neighbour-joining tree constructed using the K80 (Kimura, 1980) nucleotide substitution model to compute evolutionary distances. Trees were reconstructed using (a) a subset of 21 *bcr1* repeats displaying multiple sequence-related copies in the different genomes (see Fig. 5), and (b) a subset of 43 *bcr1* repeats located in a conserved genomic locus (in two strains or more) and not displaying sequence identity above the 95% cut-off to any repeats outside this particular locus. In (a) and (b) underlined repeats and their corresponding predicted folding structures are shown (to the right) as examples of typical structures.

built using (1) a subset of 21 *bcr1* repeats displaying multiple sequence-related copies in the different genomes, and (2) a subset of 43 *bcr1* repeats located in a conserved genomic locus (in two strains or more) and not displaying sequence identity above the 95% cut-off to any repeats outside this particular locus. In the first subset 19 out of the 21 repeats (~90%) were predicted to form double-hairpin-shaped structures, and only two displayed a deviating secondary structure, while in the second subset 22 of 43 repeats (~51%) were predicted to exhibit a double-hairpin-like fold, while 21 displayed a deviating structure (Fig. 7). These results strongly indicate that the ability to form a double-hairpin-like structure is correlated with the potential mobility of *bcr1* elements.

Use of *bcr1* as a molecular typing tool

Recently, a highly conserved end of *bcr1* (26 bp), which is also part of many of the partial elements in *B. cereus* group genomes, was used in a repetitive extragenic palindromic (rep)-PCR study of *B. thuringiensis* strains, taking advantage of its variable chromosomal distribution pattern for phylogenetic purposes (Reyes-Ramirez & Ibarra, 2005). In the present study, a phylogenetic tree was constructed based on the concatenated sequence of three *bcr1* repeats (loci 1, 2 and 7 in Table 2) found to be located in corresponding genomic loci in the four complete genomes examined here (*B. cereus* ATCC 10987, *B. cereus* E33L, *B. thuringiensis* 97-27 and *B. anthracis* Ames), as well as in *B. cereus* strains AH818 and AH820 by PCR analysis (Fig. 1b). The phylogeny of the concatenated *bcr1* sequences

corresponded well with that reconstructed by MLST (Fig. 1a, b). Thus, several features of *bcr1* follow the MLST phylogeny for the strains examined in this study (Fig. 1a): with increased phylogenetic distance from *B. anthracis* (1) the total number of full-length repeats in each particular strain increases, (2) the number of shared *bcr1* loci decreases (Table 2), and (3) the sequence-based phylogeny of concatenated repeats from shared chromosomal loci is congruent with the phylogeny obtained by MLST.

DISCUSSION

In this paper we have described an extensive analysis of the DNA repeat element *bcr1* in a set of closely related strains from the *B. cereus* group of bacteria, allowing the investigation of the relationship between *bcr1* folding and mobility. Alignment with previously known *bcr1* repeats (Økstad *et al.*, 2004) showed that the newly identified *bcr1* copies were part of the *bcr1* family and that the iterative dual-BLAST procedure enabled improved *bcr1* identification. A recent study describing the genome comparison of *B. cereus* E33L and *B. thuringiensis* 97-27 with the previously sequenced genomes in the group (Han *et al.*, 2006) lists lower numbers of *bcr1* elements, presumably due to the use of a less comprehensive identification procedure in that study. As the strains compared here are more closely related than those previously compared (Økstad *et al.*, 2004), there was a larger number of repeats sharing the same genomic locus in different strains (Table 2), which allowed us to gain insights into *bcr1* evolution.

In this work we present evidence that *bcr1* is transcribed and may be present on small RNA molecules, and that it may form a secondary structure that is maintained by compensatory mutations. Northern hybridizations using full-length *bcr1* as a probe indicated that *bcr1* elements were part of RNA transcripts, in both the longer [1.0–2.5 kbp; (Økstad *et al.*, 2004)] and shorter size range (120–400 bp; Fig. 2). Hybridization signals were obtained for both strands, which is probably due to the inverted repeat

character of *bcr1* (Fig. 3a) and/or *bcr1* exhibiting transcription in both directions, potentially in a locus-dependent fashion. The presence of *bcr1* on long transcripts suggests that it may be co-transcribed with neighbouring genes. This would be expected from the fact that in many cases *bcr1* lies very close to annotated genes, sometimes overlapping the stop codon (Økstad *et al.*, 2004). Interestingly, the detection of small transcripts whose sizes were about the size of full-length *bcr1* could

indicate an independent or autonomous expression of the repeat element. However, it may also be possible that the smaller RNAs are the result of *bcr1* being cut off from longer transcripts, as is the case for the NEMIS (Correia) repeats in *Neisseria* species (Mazzone *et al.*, 2001). Furthermore, the higher number of *bcr1* bands and stronger banding pattern observed for *B. cereus* ATCC 10987 compared to *B. cereus* AH820 and *B. thuringiensis* 97-27 (Fig. 2) may be explained by the higher number of full-length *bcr1* elements in the chromosome of the ATCC 10987 strain, and the possibility that different *bcr1*-containing transcripts (originating from variable chromosomal loci) may be present in any one band in the gel.

The *bcr1* sequence has the potential to fold into a stable double-hairpin-like secondary structure whose folding free energy is comparable to that of repeated elements of similar sizes from other organisms (see examples below) (Fig. 3a and Supplementary Fig. S2). The identification of several compensatory mutations points towards the importance of maintaining this structure, implying that *bcr1* has a functional role and/or an activity which depends on the integrity of the structure. As described previously (Økstad *et al.*, 2004), *bcr1* exhibits a number of characteristics of mobile elements, in particular a heterogeneous chromosomal distribution between strains, the occasional insertion inside chromosomal genes, and a TTTAT target site duplication at its termini. An interesting pattern that emerged from the analysis of *bcr1* genomic distribution and sequence identity was that many *bcr1* repeats displayed sequence identities above the chromosomal average, to repeats in non-corresponding loci in other strains (Fig. 5). In addition, some of the *bcr1* copies that are found at a corresponding genomic locus in different strains (e.g. *bcr1* 2R) also display high similarity to repeats in other locations. Given the strong correlation between high predicted folding stability, ability to form a double-hairpin-like structure, and high sequence identity to multiple *bcr1* copies in non-conserved loci (Figs 5 and 7), it is tempting to suggest that folding is linked to mobility, at either the DNA or RNA level. Maintaining a stable secondary structure may in this respect be important for the mobility mechanism. These observations underline the potential mobile nature of *bcr1* and could be explained by duplication events having occurred in individual genomes. Very striking is the fact that *bcr1* repeats in different genomes can actually share higher sequence identity than copies within each genome, possibly representing recent mobility events. To explain the small number of loci shared between strains, independent duplication and/or excision seems more likely than differential loss, since the latter would imply that the ancestor of the *B. cereus* group carried an unreasonably large number of *bcr1* copies (Økstad *et al.*, 2004). Even though there are a few cases of *bcr1* being missing from a conserved insertion locus in one particular strain (Table 2), in most cases the sequence context appeared to have undergone additional rearrangements. Therefore, there is no conclusive evidence indicating precise *bcr1* excision.

The *bcr1* sequence is probably specific to the *Bacillus cereus* group of bacteria, which also harbours five other specific repeated elements of 110–310 bp, exhibiting variable copy numbers and genomic localizations (Tourasse *et al.*, 2006). Miniature repeats with properties related to *bcr1* have also been found in other prokaryotic and eukaryotic species. ERICs of enterobacteria (Bachelier *et al.*, 1999; Stern *et al.*, 1984), NEMIS in *Neisseria* spp. (Buisine *et al.*, 2002; Correia *et al.*, 1988; Mazzone *et al.*, 2001), BOX and RUP elements in *Streptococcus pneumoniae* (Knutsen *et al.*, 2006; Martin *et al.*, 1992; Oggioni & Claverys, 1999) and the MITEs commonly found in eukaryotic genomes (Bureau & Wessler, 1994; Izsvak *et al.*, 1999; Wessler *et al.*, 1995) have all been predicted *in silico* to have the ability to form potentially stable secondary structures. NEMIS and RUPs exist in high copy numbers and both generate a specific TA target site duplication. The NEMIS elements can be co-transcribed with cellular genes (Mazzone *et al.*, 2001), and at least one of the BOX elements is expressed (Martin *et al.*, 1992). Interestingly, it has been proposed that maintenance of the stable stem-loop structure of the MITEs is involved in their amplification (Izsvak *et al.*, 1999), as seems to be the case for *bcr1*. Furthermore, examination of entries in the Rfam database (Griffiths-Jones *et al.*, 2003, 2005) indicated that the predicted double hairpin-like structure of *bcr1* is highly similar to the structure of a group of small nucleolar RNA (snoRNA) molecules from eukaryotes (reviewed by Kiss, 2002), more specifically those of certain SNORA families. Similar to *bcr1*, snoRNAs constitute non-coding RNA, and are known to exhibit a variety of functions related to RNA or DNA modification or processing (Kiss *et al.*, 2004; Kiss, 2002).

When the maintenance of structural stability is analysed in the context of the *bcr1* phylogeny, it appears that repeats with folding energies above and below the chromosomal average are intermixed in the tree (Fig. 6). This pattern may suggest that *bcr1* has lost structural stability multiple times during its evolution. During mobility events, repeats could insert into new genomic loci where they may or may not provide a novel function to the cell. One may speculate that due to the loss of selection pressure, *bcr1* elements inserting into non-favourable genomic loci could accumulate mutations and deletions and/or be subject to degradation, thereby disfavours the formation of the double-hairpin secondary structure which may be essential for mobility. These repeats would then lose structural stability and thus the ability to move further, and could, through deletion and/or substitution processes, be the origin of the multitude of partial *bcr1* elements observed. Interestingly, a phylogenetic analysis based on the three full-length *bcr1* repeats that are present in the same locus in all strains analysed here produced a tree corresponding to the chromosomal MLST phylogeny (Fig. 1). This suggests that repeats in conserved loci are old and have followed genomic evolution. With a few exceptions, these elements also exhibit weak folding stability. It would thus seem that when the pressure to maintain the secondary structure is

lost, *bcr1* evolves along with the host genome. As a consequence, by identification of the conserved repeats in strains of interest, *bcr1* might have the potential to be used as a high-resolution typing tool. This could be particularly useful for highly similar strains, where sufficient resolution is unattainable by MLST, e.g. the AH818 and AH820 strains studied here. Finally, a puzzling aspect of *bcr1* and host genome evolution is the fact that the number of *bcr1* copies decreases along the phylogenetic tree leading to *B. anthracis*. The number of partial *bcr1* elements in *B. anthracis* is however comparable to those of the *B. cereus* and *B. thuringiensis* strains except for *B. cereus* ATCC 10987 (Supplementary Table S2). Also the total copy number of other repeat families identified in *B. anthracis* is similar to those in *B. cereus* and *B. thuringiensis* strains (Tourasse *et al.*, 2006), suggesting that the difference in copy number may be specific to the *bcr1* repeat.

Repeats found in other prokaryotes have been assigned a multitude of functions (Hofnung & Shapiro, 1999; Versalovic & Lupski, 1998). In this study we see signs of mobility and transcription of the *bcr1* repeat element, and the importance of maintaining a stable DNA or RNA secondary structure in order for mobility to occur. Furthermore, the potential to form secondary structures at the RNA level is apparent, and may suggest that *bcr1* could provide function(s) to the cell, such as modulation of mRNA stability, transcription termination and/or promoter activity, as has been observed for elements with similar features in *Neisseria* and *Streptococcus* species (Buisine *et al.*, 2002; Correia *et al.*, 1988; Knutsen *et al.*, 2006; Martin *et al.*, 1992; Mazzone *et al.*, 2001; Oggioni & Claverys, 1999). Since no apparent general function can currently be assigned to *bcr1*, it could represent a form of 'selfish' mobile DNA, which occasionally gains a function when entering an appropriate chromosomal locus.

ACKNOWLEDGEMENTS

We thank Erlendur Helgason for constructing the MLST tree for Fig. 1(a). This work was supported by grants from The Research Council of Norway through a Strategic University Programme (SUP, project number 146534/420) and through the FUGE Consortium for Advanced Microbial Sciences and Technology (FUGE-CAMST, project number 152020/310).

REFERENCES

- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.
- Bachelier, S., Clement, J. M. & Hofnung, M. (1999). Short palindromic repetitive DNA elements in enterobacteria: a survey. *Res Microbiol* **150**, 627–639.
- Buisine, N., Tang, C. M. & Chalmers, R. (2002). Transposon-like Correia elements: structure, distribution and genetic exchange between pathogenic *Neisseria* sp. *FEBS Lett* **522**, 52–58.
- Bureau, T. E. & Wessler, S. R. (1994). Mobile inverted-repeat elements of the Tourist family are associated with the genes of many cereal grasses. *Proc Natl Acad Sci U S A* **91**, 1411–1415.
- Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540–552.
- Correia, F. F., Inouye, S. & Inouye, M. (1988). A family of small repeated elements with some transposon-like properties in the genome of *Neisseria gonorrhoeae*. *J Biol Chem* **263**, 12194–12198.
- Drobniowski, F. A. (1993). *Bacillus cereus* and related species. *Clin Microbiol Rev* **6**, 324–338.
- Galtier, N., Gouy, M. & Gautier, C. (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* **12**, 543–548.
- Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. (2003). Rfam: an RNA family database. *Nucleic Acids Res* **31**, 439–441.
- Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R. & Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**, D121–D124.
- Han, C. S., Xie, G., Challacombe, J. F., Altherr, M. R., Bhotika, S. S., Brown, N., Bruce, D., Campbell, C. S., Campbell, M. L. & other authors (2006). Pathogenomic sequence analysis of *Bacillus cereus* and *Bacillus thuringiensis* isolates closely related to *Bacillus anthracis*. *J Bacteriol* **188**, 3382–3390.
- Helgason, E., Caugant, D. A., Olsen, I. & Kolstø, A. B. (2000a). Genetic structure of population of *Bacillus cereus* and *B. thuringiensis* isolates associated with periodontitis and other human infections. *J Clin Microbiol* **38**, 1615–1622.
- Helgason, E., Økstad, O. A., Caugant, D. A., Johansen, H. A., Fouet, A., Mock, M., Hegna, I. & Kolstø, A. B. (2000b). *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis* – one species on the basis of genetic evidence. *Appl Environ Microbiol* **66**, 2627–2630.
- Helgason, E., Tourasse, N. J., Meisal, R., Caugant, D. A. & Kolstø, A. B. (2004). Multilocus sequence typing scheme for bacteria of the *Bacillus cereus* group. *Appl Environ Microbiol* **70**, 191–201.
- Hernandez, E., Ramisse, F., Ducoureaux, J. P., Cruel, T. & Cavallo, J. D. (1998). *Bacillus thuringiensis* subsp. *konkukian* (serotype H34) superinfection: case report and experimental evidence of pathogenicity in immunosuppressed mice. *J Clin Microbiol* **36**, 2138–2139.
- Herron, W. M. (1930). *Rancidity in Cheddar cheese*. Master's Thesis, Queen's University, Kingston, Ontario, Canada.
- Hofnung, M. & Shapiro, J. A. (1999). Introduction – special issue on repetitive DNA sequences in microbes. *Res Microbiol* **150**, 577–578.
- Hohl, M., Kurtz, S. & Ohlebusch, E. (2002). Efficient multiple genome alignment. *Bioinformatics* **18** (Suppl 1), S312–S320.
- Huelsenbeck, J. P. & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755.
- Ivanova, N., Sorokin, A., Anderson, I., Galleron, N., Candelon, B., Kapatral, V., Bhattacharyya, A., Reznik, G., Mikhailova, N. & other authors (2003). Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. *Nature* **423**, 87–91.
- Izsvak, Z., Ivics, Z., Shimoda, N., Mohn, D., Okamoto, H. & Hackett, P. B. (1999). Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. *J Mol Evol* **48**, 13–21.
- Jernigan, J. A., Stephens, D. S., Ashford, D. A., Omenaca, C., Topiel, M. S., Galbraith, M., Tapper, M., Fisk, T. L., Zaki, S. & other authors (2001). Bioterrorism-related inhalational anthrax: the first 10 cases reported in the United States. *Emerg Infect Dis* **7**, 933–944.

- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**, 111–120.
- Kiss, T. (2002). Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell* **109**, 145–148.
- Kiss, A. M., Jady, B. E., Bertrand, E. & Kiss, T. (2004). Human box H/ACA pseudouridylation guide RNA machinery. *Mol Cell Biol* **24**, 5797–5807.
- Knutsen, E., Johnsborg, O., Quentin, Y., Claverys, J. P. & Håvarstein, L. S. (2006). BOX elements modulate gene expression in *Streptococcus pneumoniae*: impact on the fine-tuning of competence development. *J Bacteriol* **188**, 8307–8312.
- Kozik, A., Kochetkova, E. & Micheltore, R. (2002). GenomePixelizer – a visualization program for comparative genomics within and between species. *Bioinformatics* **18**, 335–336.
- Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**, 1244–1245.
- Lombard, V., Camon, E. B., Parkinson, H. E., Hingamp, P., Stoesser, G. & Redaschi, N. (2002). EMBL-Align: a new public nucleotide and amino acid multiple sequence alignment database. *Bioinformatics* **18**, 763–764.
- Martin, B., Humbert, O., Camara, M., Guenzi, E., Walker, J., Mitchell, T., Andrew, P., Prudhomme, M., Alloing, G. & other authors (1992). A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*. *Nucleic Acids Res* **20**, 3479–3483.
- Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**, 911–940.
- Mazzone, M., De Gregorio, E., Lavitola, A., Pagliarulo, C., Alifano, P. & Di Nocera, P. P. (2001). Whole-genome organization and functional properties of miniature DNA insertion sequences conserved in pathogenic Neisseriae. *Gene* **278**, 211–222.
- Oggioni, M. R. & Claverys, J. P. (1999). Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in *Streptococcus pneumoniae*. *Microbiology* **145**, 2647–2653.
- Økstad, O. A., Hegna, I., Lindbäck, T., Rishovd, A. L. & Kolstø, A. B. (1999). Genome organization is not conserved between *Bacillus cereus* and *Bacillus subtilis*. *Microbiology* **145**, 621–631.
- Økstad, O. A., Tourasse, N. J., Stabell, F. B., Sundfaer, C. K., Egge-Jacobsen, W., Risøen, P. A., Read, T. D. & Kolstø, A. B. (2004). The *bcr1* DNA repeat element is specific to the *Bacillus cereus* group and exhibits mobile element characteristics. *J Bacteriol* **186**, 7714–7725.
- Page, R. D. (1996). TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* **12**, 357–358.
- Rasko, D. A., Ravel, J., Økstad, O. A., Helgason, E., Cer, R. Z., Jiang, L., Shores, K. A., Fouts, D. E., Tourasse, N. J. & other authors (2004). The genome sequence of *Bacillus cereus* ATCC 10987 reveals metabolic adaptations and a large plasmid related to *Bacillus anthracis* pXO1. *Nucleic Acids Res* **32**, 977–988.
- Rasko, D. A., Altherr, M. R., Han, C. S. & Ravel, J. (2005). Genomics of the *Bacillus cereus* group of organisms. *FEMS Microbiol Rev* **29**, 303–329.
- Read, T. D., Peterson, S. N., Tourasse, N., Baillie, L. W., Paulsen, I. T., Nelson, K. E., Tettelin, H., Fouts, D. E., Eisen, J. A. & other authors (2003). The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature* **423**, 81–86.
- Reyes-Ramirez, A. & Ibarra, J. E. (2005). Fingerprinting of *Bacillus thuringiensis* type strains and isolates by using *Bacillus cereus* group-specific repetitive extragenic palindromic sequence-based PCR analysis. *Appl Environ Microbiol* **71**, 1346–1355.
- Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.
- Rozen, S. & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365–386.
- Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425.
- Schnepf, E., Crickmore, N., Van Rie, J., Lereclus, D., Baum, J., Feitelson, J., Zeigler, D. R. & Dean, D. H. (1998). *Bacillus thuringiensis* and its pesticidal crystal proteins. *Microbiol Mol Biol Rev* **62**, 775–806.
- Sonnhammer, E. L. & Durbin, R. (1994). A workbench for large-scale sequence homology analysis. *Comput Appl Biosci* **10**, 301–307.
- Stern, M. J., Ames, G. F., Smith, N. H., Robinson, E. C. & Higgins, C. F. (1984). Repetitive extragenic palindromic sequences: a major component of the bacterial genome. *Cell* **37**, 1015–1026.
- Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680.
- Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876–4882.
- Tourasse, N. J., Helgason, E., Økstad, O. A., Hegna, I. K. & Kolstø, A. B. (2006). The *Bacillus cereus* group: novel aspects of population structure and genome dynamics. *J Appl Microbiol* **101**, 579–593.
- Versalovic, J. & Lupski, J. R. (1998). Interspersed repetitive sequences in bacterial genomes. In *Bacterial Genomes—Physical Structure and Analysis*, pp. 38–48. Edited by F. J. de Bruijn, J. R. Lupski, & G. M. Weinstock. New York: Chapman & Hall.
- Wessler, S. R., Bureau, T. E. & White, S. E. (1995). LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr Opin Genet Dev* **5**, 814–821.
- Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406–3415.

Edited by: D. A. Mills