LBNL-61525
Preprint

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Title

Extending the *cereus* group genomics to putative food-borne pathogens of different toxicity

Author(s) Alla Lapidus, Eugene Goltsman, et al

Division Genomics

Extending the *cereus* group genomics to putative food-borne pathogens of different toxicity.

Alla Lapidus[1], Eugene Goltsman[1], Sandrine Auger[2], Nathalie Galleron[2], Béatrice Ségurens[3], Carole Dossat[3], Miriam L. Land[1], Veronique Broussole[4], Julien Brillard[4], Marie-Helene Guinebretiere[4], Vincent Sanchis[5], Christophe Nguen-the[4], Didier Lereclus[5], Paul Richardson[1], Patrick Winker[3], Jean Weissenbach[3], S. Dusko Ehrlich[2], Alexei Sorokin[2*].

[1] *DOE Joint Genome Institute, Walnut Creek, CA, USA*
[2] *Génétique Microbienne, INRA, Domaine de Vilvert, 78352 Jouy en Josas cedex, France*
[3] *Génoscope, Centre National de Séquençage, 2 rue Gaston Crémieux CP 5706, 91057 Evry cedex, France*
[4] *Sécurité et Qualité des Produits d'Origine Végétale, UMR INRA – Université d'Avignon, Avignon, France*
[5] *Génétique Microbienne et Environement, INRA, La Minière, 78285 Guyancourt cedex France*

Running title: *cereus* genomics

[*]**For correspondence** :

Alexei Sorokin

Génétique Microbienne, INRA, Domaine de Vilvert,

78352 Jouy en Josas cedex, France

Tel. (33) 1.34.65.27.24

Fax. (33) 1.34.65.25.21

E-mail: sorokine@jouy.inra.fr

ABSTRACT

The *cereus* group represents sporulating soil bacteria containing pathogenic strains which may cause diarrheic or emetic food poisoning outbreaks. Multiple locus sequence typing revealed a presence in natural samples of these bacteria of about thirty clonal complexes. Application of genomic methods to this group was however biased due to the major interest for representatives closely related to *B. anthracis*. Albeit the most important food-borne pathogens were not yet defined, existing data indicate that they are scattered all over the phylogenetic tree. The preliminary analysis of the sequences of three genomes discussed in this paper narrows down the gaps in our knowledge of the *cereus* group. The strain NVH391-98 is a rare but particularly severe food-borne pathogen. Sequencing revealed that the strain must be a representative of a novel bacterial species, for which the name *Bacillus cytotoxis* is proposed. This strain has a reduced genome size compared to other *cereus* group strains. Genome analysis revealed absence of sigma B factor and the presence of genes encoding diarrheic Nhe toxin, not detected earlier. The strain *B. cereus* F837/76 represents a clonal complex close to that of *B. anthracis*. Including F837/76, three such *B. cereus* strains had been sequenced. Alignment of genomes suggests that *B. anthracis* is their common ancestor. Since such strains often emerge from clinical cases, they merit a special attention. The third strain, KBAB4, is a typical psychrotrophe characteristic to unbiased soil communities. Phylogenic studies show that in nature it is the most active group in terms of gene exchange. Genomic sequence revealed high presence of extra-chromosomal genetic material (about 530 kb) that may account for this phenomenon. Genes coding Nhe-like toxin were found on a big plasmid in this strain. This may indicate a

potential mechanism of toxicity spread from the psychrotrophic strain community. The results of this genomic work and ecological compartments of different strains incite to consider a necessity of creating prophylactic vaccines against bacteria closely related to NVH391-98 and F837/76. Presumably developing of such vaccines can be based on the properties of non-pathogenic strains such as KBAB4 or ATCC14579 reported here or earlier. By comparing the protein coding genes of strains being sequenced in this project to others we estimate the shared proteome in the *cereus* group to be 3,000±200 genes and the total proteome to be 20-25,000 genes.

INTRODUCTION

Bacillus cereus (Bce) is one of the most ubiquitous bacteria on Earth, its representatives are commonly found in soil, water, plants, dead or live insects and animals. The *cereus* group includes *B. anthracis (Ban)*, *B. thuringiensis (Bth)*, *B. mycoides*, *B. pseudomycoides*, *B. weihenstephanensis (Bwe)* and *B. cereus sensu stricto* (Bce), distinguished by such features as the presence of specific animal or insect toxins in *Ban* and *Bth*, mycoidal growth of *B. mycoides* and *B. pseudomycoides*, and psychrotrophy of *Bwe*. *Bce* is an opportunistic pathogen that causes gastrointestinal diseases manifested by diarrhoeic or emetic syndromes. The majority of emetic strains, similarly to other specific pathogens of this group, represent single evolutionary lineage of closely related strains [1]. Such a specificity cannot however be attributed to the food spoilers causing diarrhoea. Two enterotoxins (haemolysin BL and non-haemolytic enterotoxin Nhe) are usually considered as the major factors responsible of diarrhoeic food poisoning [2, 3, 4]. Other factors such as phospholipase activity and stress adaptation are also considered as important [5, 6, 7]. A rare strain NVH391-98, was recently isolated from a particularly severe food poisoning outbreak, which was reported to cause three fatal cases and shown to synthesise only one toxin important for its pathogenic potential, the cytotoxin K [8]. The corresponding *cytK* allele encodes a particularly effective protein synthesised in elevated amounts [9, 10]. No other genes encoding diarrhoeic toxins seem to be present in this strain [9, 11, 12]. Therefore NVH391-98 appeared to be an interesting model to study the problem of food poisoning. Surprisingly, the nucleotide sequences of conserved genes of NVH391-98 are very distinct from those of other strains of the *cereus* group [13, 14].

Another strain of our interest, F837/76, that also can be related to the problem of food poisoning by *cereus* group bacteria, was known for about 30 years. This strain was isolated from a postoperative prostate wound [15]. Actually the strain toxicity may be beyond of a simple noxious food-borne pathogen and it must be carefully characterised with a relevant illness model, but the strain does not seem to produce toxins other than those known to cause food intoxication [16, 17]. Another interesting characteristic of this strain was that it may carry a smaller chromosome compared to other *cereus* group representatives [18].

Intuitively an important property of a strain contaminating refrigerated food products is the ability to grow at low temperatures (below 5-10^{o}C). The strains showing this phenotype are commonly found in soil. Many of them have been characterised and recognised to constitute a new species, different from others of the *cereus* group, called *B. weihenstephanensis* [19]. In a recent study many *Bwe* strains were isolated and phylogenetically characterised, revealing also their elevated genetic exchange activity in soil [14, 20]. This observation was based on the calculation of indexes of association or direct estimation of recombination to mutation frequency ratio. A representative psychrotrophic strain of this collection, KBAB4, closely related to independently isolated psychrotrophic type strains [19] was selected for more detailed studies. We decided to initiate a project for determining the complete genome sequence of the NVH391-98, F837/76 and KBAB4 strains. Here we present the genomic comparison of these strains to other *cereus* group representatives. The presence of the mentioned above diarrhoeic toxin genes relevant to the problem of food poisoning is also discussed.

RESULTS AND DISCUSSION

General genomic features of *NVH391-98*.

A comparison of genome sequence of the strain NVH391-98 with other genomes of the *B. cereus* group is shown on Figures 1 to 3. The representatives of the *B. cereus* group possess the genome of approximately 5.2-5.5 Mb [21, 22, 23, 24, 25] while assembled genome of NVH391-98 contains the chromosome of 4,085 kb and a circular plasmid of 7,136 bp. Alignment of the chromosomal contig over the genomes of related strains *B. cereus* ATCC14579 (*Bce* 14579), *B. cereus* ATCC10987 (*Bce* 10987) and *B. anthracis* Ames (*Ban* Ames, Fig. 1 and 2) showed that four regions corresponding to 750-1100, 1670-2000, 2200-3600 and 4930-5100 kb of *Bce* 14579, were barely represented in the NVH391-98 genome. We designate the former three long diverged regions as LDR900, LDR1850 and LDR3000, respectively, referring to the middle of their locations in the *Bce* 14579 genome. In total the size of these LDRs corresponds to a 2250 kb deletion in the genome of NVH391-98 compared to *Bce* ATCC14579, but since the actual reduction of genome size is about 1300 kb (5,412 of *Bce* ATCC14579 genome minus 4,085 kb of NVH391-98 genome), we estimate conserved gene content in these areas of NVH391-98 genome to be 42% (2250 – 1300 / 2250). The rest of the total of 3915 protein-coding genes of this strain is 85% conserved in other *cereus* group genomes and, as seen on Fig. 2, they are collinear. The genome of the NVH391-98 strain is therefore collinear to those of other representatives of the group *cereus*, with exception of four areas, where only 42% of genes are conserved. The locally collinear blocs (LCB) in the mentioned above LDRs have greatly reduced sizes and can be regarded as a result of multiple recombination events, which left only the genes

important for NVH391-98 survival intact (Fig. 2). The presence of phage remnants or inducible prophage (like phBC6A51 near 1,850 kb of Bce14579 genome [26]) near the center of LDR900 and LDR1850 and close to the ends of LDR3000 in the strain *Bce* 14579 (Fig. 1, circles 7 and 8) is notable. This correlates with G+C content and similarity between *Bce* 14579 and *Ban* Ames distributions over the chromosome of this strain (Fig. 1 circles 3 and 4, respectively). Presumably the integrated temperate phages are between the most important factors of genome evolution in the *cereus* group. Only one restructuring event was fixed in the genome of NVH391-98, it is the inversion of a 50 kb area around the replication terminus (Fig. 2, circled in red). The parameters defining sizes of LCB used by the software MAUVE [27] can be chosen to minimize to zero the shuffling between genomes other than NVH391-98 (Fig. 2). At these conditions 55 restructuring events needed to align NVH391-98 genome over others are predicted by GRIMM [28]. It is remarkable, that even given such a big number of restructuring events the genome of NVH391-98 still remains globally collinear to others, indicating a great evolutionary pressure on the collinearity of genomes in the *cereus* group.

Three detected features of the genome of strain NVH391-98 are of interest. One of them is the absence of tryptophan biosynthesis genes. All other known representatives of the *cereus* group, as well as other Bacillus species, do possess such genes and are able to grow without tryptophan in the medium. If confirmed for other NHV391-98-like strains, this feature may lead to an interesting phenotypic property and can be related to a particular ecological niche. This could also be an easily testable distinguishing phenotype of such strains. Another genetic property, which appeared from the completed sequence, is the presence of genes for biosynthesis of a food-poisoning related toxin Nhe. Up to now only

one important toxin, CytK, was identified in this strain and had been considered as the principal factor of the severe food intoxication [8]. Several studies, using PCR or monoclonal antibodies based approaches, failed to identify other hemolytic or non-hemolytic toxin genes [9, 11, 12]. This is not surprising since the protein-coding sequences of this strain are divergent enough to hamper PCR-based gene detection under the stringent conditions used for analysis. The genomic sequence of NHV391-98 confirmed the absence of hemolytic Hbl and HlyII toxins, but unambiguously revealed the three genes encoding NheA (77% identity to the *Bce* 14579 protein), NheB (87%) and NheC (73%) organized in operon. Interestingly that the synthesis of the protein NheB was in fact detected by the antibodies 3F1 in the work of Dietrich et al [12, see Fig. 1D of this paper]. Since the strain was used as the negative control, this result was interpreted as a non-specific reactivity of used antibodies. These genes are therefore expressed and probably the synthesized proteins can contribute to toxicity, in addition to the CytK very efficient in this strain. The third important feature of this strain revealed by genomic sequence, is the apparent absence of genes encoding σ^B transcription factor and related regulatory proteins. Until now this sigma-factor was found in all sporulating Bacilli and was intensively studied in *B. subtilis* [29]. Recent studies using the *Bce* 14579 strain revealed the importance of σ^B in heat-shock and other stresses. At the same time it appeared obvious that the molecular details of regulation of this sigma-factor and its involvement in stresses are significantly different than those known for *B. subtilis* [30]. The absence of this regulatory system in NVH391-98 may indicate that this stress regulation system is a recent acquisition in the *cereus* group.

The strain NVH391-98 may represent a unique natural model for studying stress responses independent of the σ^B function in Bacilli.

Significant difference of the genome of NVH391-98 from the others of the *cereus* group inspires to consider this strain as a representative of a new species, for which we would propose a name *Bacillus cytotoxis*. The whole genome comparison of protein coding regions of *Bce* 14579 and *Ban* Ames strains revealed the levels of identity of 80 to 100%, with an average of 93% (Fig. 3 on the top). The species status difference of these two bacteria is still ambiguous [31]. A similar comparison of genomes of *Bce* 14579 and NVH391-98 gave the distribution of identity values between 70 and 95% with an average of 82%. This is a much higher difference than that found between *Bwe* and *Bce* strains (mean value 90%, data not shown) for which the question of species difference has already been resolved [19]. The comparisons of the 16S rRNA sequences shows that NVH391-98 is as far from *Bce* 14579 and other *cereus* group representatives as *B. subtilis* (*Bsu*) is far from *B. licheniformis* (*Bli*), the two bacteria have always been considered as different species (Fig. 3 bottom). However, at present, only one strain that we would assign to the *B. cytotoxis* species has been reported. Even given the potential importance of this species as poisonous food contaminants, the characterization of other similar strains is needed to follow the formal rules of new species creation [32]. We therefore propose to temporally distinguish the species status of the strain NVH 391-98 by calling it *B. cereus* ssp *cytotoxis*.

General genomic features of *B. cereus* F837/76.

Another strain being sequenced during this work represents an important group of strains which are genetically extremely close to *Ban*, but do not synthesize the anthrax toxin or the protective capsule. These strains are rare in nature, most of them were isolated from clinical cases, usually not as severe as anthrax [33, 34]. The strain that we chose for genomic sequencing, *B. cereus* F837/76 (*Bce* F837/76), isolated from a contaminated prostate wound [15], was characterized to be an efficient producer of hemolytic and non-hemolytic enterotoxins [16, 17], and was reported to possess a small circular chromosome of 2.4 Mb with the rest of genetic material distributed on plasmids [18]. The strain is one of the closest to *Ban* sequenced strains of *Bce*. The analysis similar to that presented on Fig. 3 indicated the mean identity of 98% with *Ban* Ames in the protein coding regions, compared to 93% with *Bce* 14579 (not shown). Using the formalism of Miltiple Locus Sequence Typing (MLST) and the concept of clonal complexes [35] the strain *Bce* F837/76 can be phylogenetically localized between other strains extremely closely related to *Ban* from the public MLST database [34, 36]. Our sequencing data revealed that the strain *Bce* 837/76 is identical, in terms of this MLST schema, to the independently isolated strain R3039/03 from this database [37]. The phylogeny is presented on Fig. 4. The clonal complex to which the strain *Bce* F837/76 can be assigned, is the closest to the one containing *Ban* and in which the latter was identified as the strain ancestor (not shown). The number of strains with available MLST data was not sufficient to identify the ancestor sequence type in this clonal complex. However, the genomic alignment shown on Fig. 5, indicates that *Ban* can be considered as the common ancestor of the four sequenced strains : *Ban* Ames, *Bce* F837/76, *Bce* E33L (“Zebra Killer”) and *Bth* 97-27. Considering that the toxicity plasmids pXO1 and pXO2 can be easily lost [38] a possible scenario of recent evolution of these

strains is that they could have originated from a completely pathogenic *Ban* ancestor which lost these plasmids and succeeded to find a different ecological niche, becoming in this manner less pathogenic for animals and gaining a closer association with plant material. The recent genomic analysis of the *Bce* E33L strain and especially the existence of a plant material degrading palsmid in this strain confirm this model [25].

Our data do not confirm the presence of multiple plasmids in the strain *Bce* F837/76 as it was suggested by pulse-field electrophoresis mapping studies [18]. The size of the chromosome of this strain is 5.2 Mb, as appeared after the assembly of shot-gun sequencing data and subsequent combinatorial PCR based finishing. We detected also two plasmids, one of 10,288 bp and the second of approximately 52 kb. The second plasmid is especially interesting, since it represents a hybrid extra-chromosomal element with a fraction of genes more characteristic for temperate phages and others more characteristic for plasmids.

General genomic features of *B. weihenstephanensis* KBAB4.

The third genome being sequenced in the current project is that of a psychrotrophic strain *B. weihenstephanensis* KBAB4 (*Bwe* KBAB4). This strain was selected for sequencing due to recent phylogenetic studies of a collection of naturally isolated strains of *B. cereus* and *B. thuringiensis* [14, 20]. Most of the studied strains originated from the same geographic location in the forest soil near Versailles (region of Paris, France). The collection can be regarded as representative of the Central Europe forest soil. Probably due to the fact that the Versailles Collection was gathered during a mild winter, many strains appeare to be psychrotrophic. Their sufficient representation in the collection permitted to conclude that an active recombinational exchange in the nature between these

psychrotrophic strains is higher than that seen for the mesophilic *B. thuringiensis* isolates [14]. A typical psychrotrophic strain of this collection, *Bwe* KBAB4, which appeared to be very close to other independently isolated representative psychrotrophic strains [19] was chosen for genomic sequencing. One of the goals of this genomic sequencing work is to try to elucidate the genetic basis of the above mentioned intensive genetic exchange. Currently this project is further away from completion compared to the two mentioned above, due to technical difficulties related to the presence of several plasmids and integrated or non-integrated temperate phages in this genome. The chromosomal sequence, albeit not yet completed, can be represented by a single scaffold, those alignment to other *cereus* group genomes is shown on Fig. 6. It appears from this alignment that the chromosomal 5.2 Mb sequence of *Bwe* KBAB4 is collinear to all other *cereus* group genomes. However, this strain also contains about 650 kb of additional genomic sequence. About 530 kb are distributed on two or several plasmids, the biggest one having the size of more than 400 kb. Three contigs of 27, 31 and 52 kb correspond to integrated or non-integrated temperate phages. At present we cannot definitely assign all these contigs to the chromosomal or separated location, however, the integration of so much genetic material into the chromosome would highly bias the symmetry properties usually seen in the bacterial and especially in the Bacilli genomes. It is therefore more probable that most of these contigs, and especially the one of 400 kb are extra-chromosomal. If these elements are able to transfer DNA by conjugation or transduction, this can provide a basis for explanation of high genetic exchange in the natural *Bwe* population.

Analysis of the presence of known toxin genes provides a basis for the estimation of pathogenic potential of *Bwe* strains. In the chromosome of *Bwe* KBAB4 we have detected

two operons encoding hemolytic Hbl (66-86% identity in different protein components) and non-hemolytic Nhe (92-100% identity) enterotoxins. No CytK or HlyII homologs were detected. A very important finding was the presence of a second operon encoding all three components of Nhe toxin (42-56% of amino-acid sequence identity for the three components) on the 400 kb plasmid. To our knowledge, this is the first case of this toxin being detected on a plasmid. Since the identity level is not sufficiently high and the chromosomal and plasmid encoded toxin genes have very similar sizes, the plasmid paralogue was not identified during PCR-based screens of *cereus* group isolates [11]. Whether this plasmid encoded toxin causes eukaryotic cells lysis is yet unknown. In fact these issues must be rapidly clarified since the psychrotrophic strains, especially those containing potential diarrheic toxin genes on a plasmid, are the most probable potential risk of refrigerated food poisoning.

Comparison of *cereus* and *subtilis* groups and estimation of the *cereus* group complete proteome.

Several attempts have been done to estimate the number of protein-encoding genes shared by all representatives of the *cereus* group and also in comparison with another important Bacillus group - *subtilis*, the representative species of the latter are *B. subtilis* (*Bsu*), *B. licheniformis* (*Bli*) and *B. amyloliquefaciens*. We did a similar analysis including the genomes of strains NVH391-98 and *Bce* F837/76. Since the former is as distant from other *cereus* group strains as *Bsu* from *Bli*, its inclusion would make such analysis more reliable. The results are presented on Fig. 7 and their summary is in Table 1. Application of different analysis methods and of different stringency parameters may give statistically

scattered data. Taking into account these deviations we estimate the statistical error in genome scale prediction of the number of genes to be about 200. The number of shared genes in the strains of *cerus* group (3,000±200), once the NVH391-98 is included in the analysis, appears to be the same, dependless on whether the comparison was done using very closely related *Ban* Ames and *Bce* F837/76 or, instead of the latter, more distinct *Bce* 14579 as the third strain. About the same number of genes is shared by strains of the *subtilis* group, since the common cluster with the *cereus* group have the size of 1,700 genes, plus 1,300 genes common between *Bsu* and *Bli* (Fig. 7 and Table 1). It is also notable that the same figure of 1,700±200 common genes appears if two representatives of either group and one of another are included into the analysis (Fig. 7, bottom).

Inside of the error limit of 200 genes, the data on proteome comparison presented above are in complete concordance with similar estimations made by using different strains and different methods [23, 24, 25, 39, 40]. For example, in the first paper presenting such analysis [23] the number of shared genes between *Bce* 14579 and *Ban* was 4,302 compared to 4,300±200 estimated here for similar couples of *cereus* group (Fig. 7, on the top). It is also very close to the numbers of shared genes varying between 4,300 and 4,500 for the three strains *Ban* Ames, *Bce* 14579 and *Bce* 10987 used by Rasko [24]. The number of common genes between *Bsu* and *Bli* was estimated to be 3,171 [39], which is also in concordance with our estimation of 3,000±200. In the same paper *B. halodurans* was used to estimate the shared genes between different Bacilli groups and the value appeared to be 1,719 genes, very similar to our estimation of 1,700±200 common genes between *cereus* and *subtilis* groups. It is interesting to mention that each new sequenced strain of the *cereus*

group, not belonging to the same clonal complex with any of already sequenced strains, contains 600±200 new genes. We counted seven such genomes (*Ban* Ames, *Bce* 10987, *Bce* 14579, *Bce* E33L, *Bth* 97-27, *Bce* G9241 and *Bth* ATCC35646) available from GenBank and therefore about 3,000 + 7x600 = 7,200 different genes. The three genomes reported here will bring this figure to 9,000. Analysis of *cereus* group MLST data [14, 36, 41] shows that some 30-40 clonal clusters should be expected to be found upon the exhaustive analysis of these bacteria. We should therefore expect the presence of 20-25 thousands of genes in addition to the 3,000 core genes, as the complete proteome of these bacteria.

CONCLUSIONS

A project to sequence the entire genomes of three food-poisoning problem related strains of the *cereus* group have been initiated and is close to the completion. Preliminary analysis of the sequencing data provides the following conclusions:

(i) All *cereus* group genomes sequenced are collinear with exception of a short (50-150 kb) regions around the terminus of replication. The strain NVH391-98, which must be considered as a novel species of the group, has the smallest genome of 4,085 kb. Three characteristic Long Diverged Regions around 900, 1850 and 3000 kb contain 42% of shared genes compared to 85% in other long collinear regions. The evolution of LDR implicates integrative temperate phages.

(ii) Two new important diarrheic toxin operons were detected in both NVH391-98 and *Bwe* KBAB4 strains. An Nhe-like toxin presented on a 400 kb plasmid in *Bwe*

KBAB4 needs further phylogenetic and biochemical characterization since it may represent a mechanism of toxicity spread between the psychrotrophic strains.

(iii) NVH391-98 is the only known sporulating Bacillus strain that lacks σ^B regulator. The strain represents a natural model to study σ^B-independent stress response regulation.

(iv) The strain *Bce* F837/76 represents an important group of *Bce* strains that are extremely close to *Ban*. Genomic data indicate that the latter is the common ancestor of such strains. The group needs a special attention as a source of potentially dangerous wound contaminants and diarrheic food-borne pathogens.

(v) All strains of *cereus* group share 3,000±200 common genes. The total size of the *cereus* group proteome is estimated to be about 20-25,000 protein-coding genes from which about 9,000 have been already sequenced.

MATERIALS AND METHODS

Genomic sequencing of the *cereus* group strains. The strain NVH391-98 was obtained from Marie-Laure De Buyser (AFSSA, Maisons-Alfort, France), the strain *B. cereus* F837/76 was obtained from Anne-Brit Kolsto (University of Oslo, Norway), the strain *B. weihenstepahnensis* KBAB4 was from the Versailles Collection [14, 20] kept in GM (Jouy-en-Josas). Total DNA was prepared as described [13]. The random shotgun method of cloning, sequencing and assembly was applied as described [25, 42]. Gaps were closed by primer walking over clone inserts and by sequencing of LR PCR products using the finishing strategy based on MLA PCR [43]. The shot-gun genome sequences contained about 80,000 reads, achieving the coverage of 8 fold. Several regions (3-4 in each genome) with sizes of 10-25 kb, including the one of about 20 kb, corresponding to ribosomal protein gene cluster, were not covered by plasmid clones. These regions were entirely amplified by LR PCR and sequenced by primer walking over the LR PCR products. The number of *rrn* operons in NVH391-98 (13) and KBAB4 (14) was determined earlier by Southern hybridization [13]. It was confirmed during this project by LR PCR, using the shot-gun data, and each separately amplified operon is being sequenced by primer walking. The 13 *rrn* operons in F837/76 were only studied by LR PCR and separately sequenced by primer walking. Sequences of *Ban* Ames (accession NC003997), *Bce* 14579 (NC004722), *Bce* 10987 (NC003909), *Bce* E33L (NC006274), *Bth* 97-27 (NC005957), *Bsu* 168 (NC000964) and *Bli* 14580 (NC006270) and their annotations (faa, ffn and ptt files) were obtained from NCBI (Bethesda, Maryland). The draft assemblies of sequences reported

here are available from NCBI by accession numbers AALL01000001-AALL01000114 and AAOY01000001-AAOY01000182.

PCR amplification and sequencing. PCR using the Expand Long Template PCR System (Roche Diagnostics) was applied to obtain templates for sequencing. The cycling program was: 94^{o}C for 5 min; 12 cycles of 94^{o}C for 10 sec, 55^{o}C for 10 sec and 65^{o}C for 12 min; 24 cycles of 94^{o}C for 10 sec, 55^{o}C for 10 sec and 65^{o}C starting from 12 min and increasing this time for 15 sec each cycle. The final extension step was 10 min at 72^{o}C. PCR products were treated for 1h at 37^{o}C with Exonuclease I and Shrimp Alkaline Phosphatase (USB corporation). Sequencing reactions were performed using ABI PRISM sequencing kit (Applied Biosystems). Products of reactions were ethanol precipitated and analysed with ABI 3700 sequencer (Applied Biosystems).

Computer methods. Random shot-gun sequences were assembled using *phred-phrap* software [44]. For the finishing the consensus sequences and novel reads were analyzed using *consed* editor [44] or *gap4* software [45]. Protein-coding gene prediction and genome browsing was done using Oak Ridge National Laboratory (ORNL) Genome Channel (genome.ornl.gov/microbial/) [46] and Integrated Microbial Genomes resource at JGI (img.jgi.doe.gov/cgi-bin/pub/main.cgi) [47]. The gene models provided by EasyGene server (cbs.dtu.dk/services/EasyGene) [48] were used for comparisons of proteomes, that was done using the BLAST Score Ratio (BSR) approach [49]. The software MAUVE [27] was used for the multiple whole genome alignments. The link implemented in MAUVE outputs to the GRIMM web server [28, 50] was used to calculate the numbers of rearrangements of LCBs at a given weight between the genomes aligned by MAUVE.

ACKNOWLEDGEMENTS

We thank Prof. Anne-Brit Kolsto and Dr. Marie-Laure De Buyser for the gifts of strains. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. DE-AC52-06NA25396.Genomic sequence of NVH391-98 and KBAB4 strains was performed by the DOE Joint Genome Institute (JGI),US. The work at INRA was partially supported by the ANR- Agence Nationale de la Recherche - The French National Research Agency, under the « Programme National de Recherche en Alimentation et nutrition humaine » (project ANR-05-PNRA-013).

FIGURE LEGENDS

Figure 1. Comparison of the chromosomal map of *B. cereus* ATCC 14579 and NVH391-98.

The first inner circle after the scale represents CDS present (yellow) or absent (black) in NVH 391-98. The second circle (same as circle 8 from the center) is the quantitative representation, similarity of *Bce* 14579 protein coding sequences to the NVH391-98 sequence, average identity in a 100 kb window minus that over the whole genome (positive, red ; negative, blue), step is 10 kb. Red stretches outside highlight the regions absent in NVH391-98 compared to *Bce* 14579. Other internal circles represent distributions of different properties over the *Bce* 14579 genome. Circle 1 from the center, IS elements; circle 2, rRNA operons; circles 3-6, graphs showing distribution over the genome of different properties in a window of a given size : circle 3, G+C content in the window minus average G+C content of the whole genome (positive, red ; negative, blue), window is 200 kb, step is 10 kb; circle 4, similarity of *Bce* 14579 protein coding sequences to the *Ban* Ames sequence, average identity in a 100 kb window minus that over the whole genome (positive, red ; negative, blue), step is 10 kb ; circle 5, (C-G)/(C+G) distribution, window 10 kb, step 5 kb ; circle 6, similarity of *Bce* 14579 proteins to those of the phage protein database, average Smith-Waterman score in the 50 kb window minus 250 (positive, blue ; negative, red), step over genome is 5 kb ; circle 7, ORFs colored according to the ERGO database assigned function [23], phage-related genes are in black ; circle 8, similar to circle 4 with NVH391-98 genome instead of *Ban* Ames.

Note the correlation of deletions in the NVH391-98 genome with location of pro-phages in the *Bce* 14579 genome (black bars in the 7th and 8th circle from the center).

Figure 2. **Alignment of the chromosomal sequences of *B. anthracis* Ames, *B. cereus* ATCC 10987, *B. cereus* ATCC 14579 and NVH391-98.**

On the top : alignment of the nucleotide sequences of the four genomes by the program MAUVE [27]. The minimum weight of Locally Collinear Blocks (LCB) is set to 236 [for the detailed explanation see 27]. A short (50 kb) region inverted in NVH 391-98 genome in respect to others is circled in red. The reciprocal LCBs in different genomes are indicated by identical colors. Bottom : the numbers of re-structural events between different genomes determined by the GRIMM algorithm [28]. These numbers show the structural closeness of the strains on the whole-genome scale.

Figure 3. Phylogenetic comparison of NVH391-98 to other Bacillus.

On the top : genome-scale comparison of nucleotide sequences encoding proteins of three strains of the *cereus* group. CDS of *Bce* ATCC14579 were compared using FASTX to the genomes of NVH391-98 and *Ban* Ames. The number of hits in a given identity interval is plotted. The high- and low-identity peaks correspond to orthologous and paralogous genes respectively. The mean identity of orthologs for *Bce* ATCC14579 genome is 93% for *Ban* Ames and 82% for NVH391-98, respectively. Bottom : a phylogenetic tree based on 16S rRNA sequence for different Bacillus, Streptococcus and *S. aureus*. NVH 391-98 is as different from all other *cereus* group strains as *Bsu* is far from *Bli*. It can therefore be considered as a novel species.

Figure 4. Clonal complexes or groups closely related to *B. anthracis*.

The phylogenetic tree was constructed using the START software [51] and the allelic sequence type (ST) profiles for the strains closely related to the *anthracis* (group 7) or emetic (group 2) clonal complexes in the MLST *B. cereus* database [36]. Strain names, ST numbers and profiles (in parentheses) are indicated the same as in the MLST database. *Bce* F837/76 has the same sequences as the strain R_3039/03 (ST-75) from this database. Group 6 represents the set of *B. cereus* strains closest to *B. anthracis*. Bacteria of the group 8 are considered to be the *B. anthracis* strains [34]. *Bth* 97-27 (ST-113) was sequenced earlier [25]. Group numbers are arbitrary assigned by the START.

Figure 5. Alignment of the chromosomal sequences of *Ban* Ames, *Bth* 97-27, *Bce* F837/76 and *Bce* E33L.

On the top : alignment of the nucleotide sequences of the four genomes by the program MAUVE [27]. The minimum weight of LCB is set to 71. The reciprocal LCBs in different genomes are indicated by identical colors. Bottom : the numbers of re-structural events between different genomes determined by the GRIMM algorithm [28]. The closeness of *Ban* to three other strains suggests that it is their common ancestor.

Figure 6. Alignment of the chromosomal sequences of *Bwe* KBAB4, *Bce* 14579, *Ban* Ames and *Bce* 10987.

On the top : alignment of the nucleotide sequences of the four genomes by the program MAUVE [27]. The minimum weight of LCB is set to 82. The reciprocal LCBs in different

genomes are indicated by identical colors. Note the inversion of a 150 kb region near the replication terminus (approximately 2.6 Mb). Bottom : the numbers of re-structural events between different genomes determined by GRIMM [28] indicate structural divergence of the strain *Bwe* KBAB4.

Figure 7. Proteome comparisons of *cereus* and *subtilis* group bacteria.

Venn diagrams illustrating the similarity of shared proteomes in the *cereus* and *subtilis* groups. Each strain is indicated by a unique color. Exact numbers of predicted and shared proteins are indicated (see text for the details of error estimation).

REFERENCES

1. M. Ehling-Schulz, B. Svensson, M.-H. Guinebretiere, T. Lindback, M. Andersson, A. Schulz, M. Fricker, A. Christiansson, P.E. Granum, E. Martlbauer, C. Nguyen-The, M. Salkinoja-Salonen, and S. Scherer. 2005. Emetic toxin formation of *Bacillus cereus* is restricted to a single evolutionary lineage of closely related strains. Microbiology 151 (2005) 183-197.
2. D.J. Beecher, A.C.L. Wong. Improved purification and characterization of hemolysin BL, a hemolytic dermonecrotic vascular permeability factor from *Bacillus cereus*. Infect. Immun. 62 (1994) 980-986.
3. T. Lund, P.E. Granum. Characterization of a non-haemolytic enterotoxin complex from *Bacillus cereus* isolated after a foodborne outbreak. FEMS. Microbiol. Lett. 141 (1996) 151-156.
4. P. Granum. *Bacillus cereus*. In : M. Doyle, L. Beuchat, T. Montville (Eds.), Fundamentals in Food Microbiology, ASM Press, Washington, D.C. 1997, pp. 327-336.
5. D.J. Beecher, T.W. Olsen, E.B. Somers, A.C.L. Wong. Evidence for contribution of tripartite hemolysin BL, phosphatidylcholine-prefferring phospholipase C, and collagenase to virulence of *Bacillus cereus* endophtalmitis. Infect. Immun. 68 (2000) 5269-5276.
6. M.C. Callegan, D.C. Cochran, S.T. Kane, M.S. Gilmore, M. Gominet, D. Lereclus. Contribution of membrane-damaging toxins to *Bacillus endophthalmitis* pathogenesis. Infect. Immun. 70 (2002) 5381-5389.
7. P.M. Periago, W. van Schaik, T. Abee, J.A. Wouters. Identification of proteins involved in the heat stress response of *Bacillus cereus* ATCC 14579. Appl. Environ. Microbiol. 68 (2002) 3486-3495.
8. T. Lund, M. L. De Buyser, and P. E. Granum. A new cytotoxin from *Bacillus cereus* that may cause necrotic enteritis. Mol. Microbiol. 38 (2000) 254-261.
9. A. Fagerlund, O. Ween, T. Lund, S.P. Hardy, P.E. Granum. Genetic and functional analysis of the cytK family of genes in Bacillus cereus. Microbiology. 150 (2004) 2689-2697.

10. J. Brillard, D. Lereclus. Comparison of cytotoxin *cytK* promoters from *Bacillus cereus* ATCC 14579 and from a *B. cereus* food-poisoning strain. Microbiology. 150 (2004) 2699-2705.
11. M.H. Guinebretiere, V. Broussolle, C. Nguen-the. Enterotoxigenic profiles of food-poisoning and food-borne *Bacillus cereus* strains. Journ. Clin. Microbiol. 40 (2002) 3053-3056.
12. R. Dietrich, M. Moravek, C. Buerk, P.E. Granum, M. Martlbauer. Production and characterization of antibodies against each of the three subunits of the *Bacillus cereus* nonhemolytic enterotoxin complex. Appl. Environ. Microbiol. 71 (2005) 8214–8220.
13. B. Candelon, K. Guilloux, S.D. Ehrlich, A. Sorokin. Two distinct types of rRNA operons in the *Bacillus cereus* group. Microbiology. 150 (2004) 601-611.
14. A. Sorokin, B. Candelon, K. Guilloux, N. Galleron, N. Wackerow-Kouzova, S.D. Ehrlich, D. Bourguet, V. Sanchis. Multiple-locus sequence typing of *Bacillus cereus* and *Bacillus thuringiensis* reveals separate clustering and a distinct population structure of phychrotrophic strains. Appl. Environ. Microbiol. 72 (2006) 1569-1578.
15. P.C.B. Turnbull, K. Jorgensen, J.M. Kramer, R.J. Gilbert, J.M. Parry. Severe clinical conditions associated with *Bacillus cereus* and the apparent involvement of exotoxins. Journ. Clinical. Path. 32 (1979) 289-293.
16. D.J. Beecher, J.D. McMillan. A novel bicomponent hemolysin from *Bacillus cereus*. Infect. Immun. 58 (1990) 2220-2227.
17. T. Lund, P.E. Granum. Comparison of biological effect of the two different enterotoxin complexes isolated from three different strains of *Bacillus cereus*. Microbiology 143 (1997) 3329-3336.
18. C.R. Carlson, A.B. Kolsto. A small (2.4 Mb) *Bacillus cereus* chromosome corresponds to a conserved region of a larger (5.3 Mb) *Bacillus cereus* chromosome. Mol. Microbiol. 13 (1994) 161-169.
19. S. Lechner, R. Mayr, K.P. Francis, B.M. Pruss, T. Kaplan, E. Wiessner-Gunkel, G.S. Stewart, and S. Scherer. *Bacillus weihenstephanensis* sp. nov. is a new psychrotolerant species of the *Bacillus cereus* group. Int. J. Syst. Bacteriol. 48 (1998) 1373-1382.

20. G. Vilas-Boas, V. Sanchis, D. Lereclus, M.V.F. Lemos, D. Bourguet. Genetic differentiation between sympatric populations of *Bacillus cereus* and *Bacillus thuringiensis*. Appl. Environ. Microbiol. 68 (2002) 1414-1424.
21. C.R. Carlson, A. Grønstad, A.B. Kolstø. Physical map of the genomes of three *Bacillus cereus* strains. J. Bacteriol. 174 (1992) 3750-3756.
22. T.D. Read, S.N. Peterson, N. Tourasse, L.W. Baillie, I.T. Paulsen, K.E. Nelson, H. Tettelin, D.E. Fouts, J.A. Eisen, S.R. Gill, E.K. Holtzappel, O.A. Okstad, E. Helgason, J. Rilstone, M. Wu, J.F. Kolonay, M.J. Beanan, R.J. Dodson, L.M. Brinkac, M. Gwinn, R.T. DeBoy, R. Madpu, S. C. Daugherty, A.S. Durkin, D.H. Haft, W.C. Nelson, J.D. Peterson, M. Pop, H.M. Khouri, D. Radune, J.L. Benton, Y. Mahamoud, L. Jiang, I.R. Hance, J.F. Weidman, K.J. Berry, R.D. Plaut, A.M. Wolf, K.L. Watkins, W.C. Nierman, A. Hazen, R. Cline, C. Redmond, J.E. Thwaite, O. White, S.L. Salzberg, B. Thomason, A.M. Friedlander, T.M. Koehler, P.C. Hanna, A.B. Kolsto, and C.M. Fraser. The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. Nature. 423 (2003) 81-86.
23. N. Ivanova, A. Sorokin, I. Anderson, N. Galleron, B. Candelon, V. Kapatral, A. Bhattacharyya, G. Reznik, N. Mikhailova, A. Lapidus, L. Chu, M. Mazur, E. Goltsman, N. Larsen, M. D'Souza, T. Walunas, Y. Grechkin, G. Pusch, R. Haselkorn, M. Fonstein, S. D. Ehrlich, R. Overbeek, N. Kyrpides. Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. Nature. 423 (2003) 87-91.
24. D.A. Rasko, J. Ravel, O.A. Okstad, E. Helgason, R.Z. Cer, L. Jiang, K.A. Shores, D.E. Fouts, N.J. Tourasse, S.V. Angiuoli, J. Kolonay, W.C. Nelson, A.B. Kolsto, C.M. Fraser, T.D. Read. The genome sequence of *Bacillus cereus* ATCC 10987 reveals metabolic adaptations and a large plasmid related to *Bacillus anthracis* pXO1. Nucl. Acids Res. 32 (2004)977-88.
25. C.S. Han, G. Xie, J.F. Challacombe, M.R. Altherr, S.S. Bhotika, D.Bruce, C.S. Campbell, M.L. Campbell, J. Chen, O. Chertkov, C. Cleland, M. Dimitrijevic, N.A. Doggett, J.J. Fawcett, T. Glavina, L.A. Goodwin, K.K. Hill, P. Hitchcock, P.J. Jackson, P. Keim, A.R. Kewalramani, J. Longmire, S. Lucas, S. Malfatti, K. McMurry, L.J. Meincke, M. Misra, B.L. Moseman, M. Mundt, A.C. Munk, R.T. Okinaka, B. Parson-

Quintana, L.P. Reilly, P. Richardson, D.L. Robinson, E. Rubin, E. Saunders, R. Tapia, J.G. Tesmer, N. Thayer, L.S. Thompson, H. Tice, L.O. Ticknor, P.L. Wills, T.S. Brettin, P. Gilna. Pathogenomic sequence analysis of *Bacillus cereus* and *Bacillus thuringiensis* isolates closely related to *Bacillus anthracis.* J. Bacteriol. 188 (2006) 3382-3390.

26. B. Candelon. Caractérisation des opérons ARN ribosomoques et des prophages comme facteurs potentiels de la plasticité génomique chez *Bacillus cereus*. Ph.D. thesis. (2004) Universite Paris XI, Orsay, France.
27. A.C.E. Darling, B. Mau, F.R. Blattner, N.T.Perna. Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Res. 14 (2004) 1394-1403.
28. G. Tesler. GRIMM: genome rearrangements web server. Bioinformatics, 18 (2002) 492-493.
29. C.W. Price. General stress response, in : A.L. Sonenshein, J.A. Hoch, R. Losick (Eds), *Bacillus subtilis* and its closest relatives. From genes to cells. ASM Press, Washington, D.C. 2002, pp. 369-384.
30. W. van Schaik, T. Abee. The role of σ^B in the stress responce of Gram-positive bacteria – targets for food preservation and safety. Curr. Opin. Biotechnol. 16 (2005) 218-224.
31. E. Helgason, O.A. Okstad, D.A. Caugant, H.A. Johansen, A. Fouet, M. Mock, I. Hegna, A.B. Kolsto. *Bacillus anthracis, Bacillus cereus,* and *Bacillus thuringiensis*-one species on the basis of genetic evidence. Appl. Env. Microbiol. 66 (2000) 2627-2630.
32. H. Christensen, M. Bisgaard, W. Frederiksen, R. Mutters, P. Kuhnert, J.E. Olsen. Is characterization of a single isolate sufficient for valid publication of a new genus or species? Proposal to modify Recommendation 30b of the Bacteriological Code (1990 Revision). Int. J. Syst. Evol. Microbiol. 51 (2001) 2221-2225.
33. E. Helgason, D.A. Caugant, I. Olsen, A.B. Kolsto. Genetic structure of population of *Bacillus cereus* and *B. thuringiensis* isolates associated with periodontitis and other human infections. J. Clin. Microbiol. 38 (2000) 1615-1622.

34. C.K. Marston, J.E. Gee, T. Popovic, A.R. Hoffmaster. Molecular approaches to identify and differentiate *Bacillus anthracis* from phenotypically similar species isolates. BMC Microbiol. 6 (2006) 22-28.
35. E.J. Feil, M.C.J. Maiden, M. Achtman, B.G. Spratt. The relative contribution of recombination and mutation to the divergenece of clones of *Neisseria meningitidis*. Mol. Biol. Evol. 16 (1999) 1496-1502.
36. F.G. Priest, M. Barker, L.W. Baillie, E.C. Holmes, M.C. Maiden. Population structure and evolution of the *Bacillus cereus* group. J. Bacteriol. 186 (2004) 7959-7970.
37. M. Barker, B. Thakker, F.G. Priest. Multilocus sequence typing reveals that Bacillus cereus strains isolated from clinical infections have distinct phylogenetic origins. FEMS Microbiol. Lett. 245 (2005) 179-184.
38. C.K. Marston, A.R. Hoffmaster, K.E. Wilson, K.E. Bragg, B. Pikaytis, P. Brachman, S. Johnson, T. Popovic. Effects of long-term storage on plasmid stability in *Bacillus anthracis*. Appl. Environ. Microbiol. 71 (2005) 7778-7780.
39. M.W. Rey, P. Ramaiya, B.A. Nelson, S.D. Brody-Karpin, E.J. Zaretsky, M. Tang, A. Lopez de Leon, H. Xiang, V. Gusti, I.G. Clausen, P.B. Olsen, M.D. Rasmussen, J.T. Andersen, P.L. Jorgensen, T.S. Larsen, A. Sorokin, A. Bolotin, A. Lapidus, N. Galleron, S.D. Ehrlich, R.M. Berka. Complete genome sequence of the industrial bacterium *Bacillus licheniformis* and comparisons with closely related Bacillus species. Genome Biol. 5 (2004) R77.
40. I. Anderson, A. Sorokin, V. Kapatral, G. Reznik, A. Bhattacharya, N. Mikhailova, H. Burd, V. Joukov, D. Kaznadzey, T. Walunas, M. D'Souza, N. Larsen, G. Pusch, K. Liolios, Y. Grechkin, A. Lapidus, E. Goltsman, L. Chu, M. Fonstein, S.D. Ehrlich, R. Overbeek, N. Kyrpides, N. Ivanova. Comparative genome analysis of *Bacillus cereus* group genomes with *Bacillus subtilis*. FEMS Microbiol. Lett. 250 (2005) 175-184.
41. E. Helgason, N.J. Tourasse, R. Meisal, D.A. Caugant, A.B. Kolsto. Multilocus sequence typing scheme for bacteria of the *Bacillus cereus* group. Appl. Env. Microbiol. 70 (2004) 191-201.
42. V. Barbe, D. Vallenet, N. Fonknechten, A. Kreimeyer, S. Oztas, L. Labarre, S. Cruveiller, C. Robert, S. Duprat, P. Wincker, L.N. Ornston, J. Weissenbach, P.

Marliere, G.N. Cohen, C. Medigue. Unique features revealed by the genome sequence of *Acinetobacter sp*. ADP1, a versatile and naturally transformation competent bacterium. Nucl. Acids Res. 32 (2004) 5766-5779.

43. A. Sorokin, A. Lapidus, V. Capuano, N. Galleron, P. Pujic, S.D. Ehrlich. A new approach using multiplex long accurate PCR and yeast artificial chromosomes for bacterial chromosome mapping and sequencing. *Genome Res*. 6, (1996) 448-453.
44. D. Gordon, C. Abajian, P. Green. Consed : a graphical tool for sequence finishing. Genome Res. 8 (1998) 195-202.
45. J.K. Bonfield, K. Smith, R. Staden. A new DNA sequence assembly program. Nucl. Acids Res. 25 (1995) 4992–4999.
46. L. Hauser, F. Larimer, M. Shah, E. Uberbacher. Analysis and Annotation of Microbial Genome Sequences. Genetic Engineering. 26, (2004) 225-238.
47. V.M. Markowitz, F. Korzeniewski, K. Palaniappan, E. Szeto, G. Werner, A. Padki, X. Zhao, I. Dubchak, P. Hugenholtz, I. Anderson, A. Lykidis, K. Mavromatis, N. Ivanova, N.C. Kyrpides. The integrated microbial genomes (IMG) system. Nucl. Acids Res. 34 (2006) D344-D348.
48. T.S. Larsen, A. Krogh. EasyGene – a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics 4 (2003) 1471-2105.
49. D.A. Rasko, G.S.A. Myers, J. Ravel. Visualization of comparative genomic analyses by BLAST score ratio. BMC Bioinformatics, 32 (2005) 977-88.
50. G. Bourque, P. Pevzner. Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Res. 12 (2002) 26-36.
51. K.A. Jolley, E.J. Feil, M.S. Chan, M.C.J. Maiden. Sequence type analysis and recombination tests (START). Bioinformatics 17 (2001) 1230-1231.