DISCLAIMER

The Pathogenomic Sequence Analysis of *B. cereus* and *B. thuringiensis* isolates closely related to *Bacillus anthracis*.

Cliff S. Han*, Gary Xie*, Jean F. Challacombe*, Michael R. Altherr, Smriti S. Bhotika, David Bruce, Connie S. Campbell, Mary L. Campbell, Jin Chen, Olga Chertkov, Cathy Cleland, M. Dimitrijevic-Bussod, Norman A. Doggett, John J. Fawcett, Tijana Glavina, Lynne A. Goodwin, Karen K. Hill, Penny Hitchcock,, Paul J. Jackson, Paul Keim, Avinash Ramesh Kewalramani, Jon Longmire, Susan Lucas, Stephanie Malfatti, Kim McMurry, Linda J. Meincke, Monica Misra, Bernice L. Moseman, Mark Mundt, A. Christine Munk, Richard T. Okinaka, B. Parson-Quintana, Lee P. Reilly, Paul Richardson, Donna L. Robinson, Eddy Rubin, Elizabeth Saunders, Roxanne Tapia, Judith G. Tesmer, Nina Thayer, Linda S. Thompson, Hope Tice, Lawrence O. Ticknor, Patti L. Wills, Paul Gilna, Thomas S. Brettin.

*These authors contributed equally to this study.

ABSTRACT

The sequencing and analysis of two close relatives of *Bacillus anthracis* are reported. AFLP analysis of over 300 isolates of *B. cereus, B. thuringiensis* and *B. anthracis* identified two isolates as being very closely related to *B. anthracis*. One, a *B. cereus*, BcE33L, was isolated from a zebra carcass in Nambia; the second, a *B. thuringiensis*, 97-27, was isolated from a necrotic human wound. The *B. cereus* appears to be the closest *anthracis* relative sequenced to date. A core genome of over 3,900 genes was compiled for the *Bacillus cereus* group, including *B anthracis*. Comparative analysis of these two genomes with other members of the *B. cereus* group provides insight into the evolutionary relationships among these organisms. Evidence is presented that differential regulation modulates virulence, rather than simple acquisition of virulence factors. These genome sequences provide insight into the molecular mechanisms contributing to the host range and virulence of this group of organisms.

Bacillus anthracis, *Bacillus cereus*, and *Bacillus thuringiensis* are closely related gram-positive, spore forming bacteria of the *B. cereus sensu lato* group(*1*). *B. anthracis* is the causal agent of anthrax, a zoonotic disease that can be lethal to humans. *B. cereus* is a ubiquitous soil organism and an opportunistic human pathogen most commonly associated with food poisoning (*2*). *B. thuringiensis* is an insect pathogen and its spores are widely used as a biopesticide (*3*). While independently derived strains of *B. anthracis* reveal conspicuous sequence homogeneity (*4*, *5*), environmental isolates of *B. cereus* and *B. thuringiensis* exhibit extensive genetic diversity (*6*, *7*). Here we report the sequencing and comparative analysis of the genomes of two members of the *B. cereus* group, *B. thuringiensis* 97-27 subsp. Konkukian, serotype H34 (Bt9727) (*8*), and *B. cereus* E33L (BcE33L). The Bt9727 strain was originally isolated from a case of severe human tissue necrosis (*8*), while the BcE33L strain was originally isolated in December 2000 from a swab taken in April 1996 of the carcass of a zebra suspected of having died of anthrax (PCB Turnbull, personal communication). These two strains, when analyzed by amplified fragment length polymorphism within a collection of over 300 of *B. cereus*, *B. thuringiensis* and *B. anthracis* isolates, appear closely related to *B. anthracis* (*9*). The BcE33L isolate appears to be the nearest relative to *B. anthracis* identified thus far (Figure 1). Whole genome sequencing of these possible pathogenic isolates Bt9727 and BcE33L was undertaken to identify shared and discrete genes among these isolates in comparison to the genomes of pathogenic strains, B. anthracis Ames, Bc G9241 and nonpathogenic strains Bc10987 and Bc14579. Comparison of these genomes revealed differences in terms of virulence, metabolic competence, structural components and regulatory mechanisms (Table 1).

The 5.31 Mb genome of Bt9727 comprises two replicons: a circular chromosome, encoding at least 5198 open reading frames (ORFs), and the pBT9727 plasmid (see Supplementary Materials Figure S1a). The 5.84 Mb genome of BcE33L comprises six replicons: a circular chromosome, encoding at least 5682 ORFs, and five plasmids (Figure S1b). Bt9727 and BcE33L have broad similarities to and share a high degree of synteny with *B. anthracis* Ames (*10*), *B. cereus* ATCC 14579 (*11*) and *B. cereus* ATCC 10987 (*12*). Within the *B. cereus* group, *B. anthracis*, BcE33L, and Bt9727 are part of a distinct cluster which contains many pathogenic organisms in the B. cereus group (*9*).

As illustrated in Figure S2, a total of 3917 putative proteins are shared among *B. anthracis* Ames, *B. cereus* ATCC 14579, Bt9727 and BcE33L using as a criterion whether genes were bidirectional best hits in blast searches. Comparison of the genomes of BcE33L and Bt9727 with *B. anthracis* Ames and *B. cereus* 14579 also identified strain specific genes in each organism. Of the 5682 predicted BcE33L proteins, 253 of the chromosomally encoded genes and 416 of the plasmid genes are unique. Of the 5199 predicted Bt9727 proteins, 307 of the chromosomally encoded genes and 66 of the plasmid genes are unique. BcE33L and *B. anthracis* Ames are the closest pair and share the most (221) common proteins. A list of all genes shared in each subgroup can be found in the supplementary tables (Tables S1-S14).

The chromosomally encoded virulence genes in Bt9727 and BcE33L are common to the *B. cereus* group of bacteria(*13*). Neither Bt9727 nor BcE33L have the highly characterized *B. anthracis* toxin genes (*pag, lef and cya*) encoded on pX01or the *cap* genes encoded by pXO2 (*14, 15*). Nonetheless, our results indicate that both Bt9727 and BcE33L share a set of virulence factors common to the members of the *B. cereus* group. These common virulence genes include three non-hemolytic enterotoxins, two channel-forming type III hemolysins, a perfringolysin O (listeriolysin O), phosphatidyl-inositol-specific and a phosphatidyl-choline preferring phospholipase, RNA polymerase sigma-B factor, and a p60 family extracellular protease. These last 5 genes are homologous to virulence genes encoded by the Gram-positive pathogen *Listeria monocytogenes* 12. Bt9727 and BcE33L also have a gene encoding cytotoxin K previously identified in *B. cereus* (ATCC 14579). While BcE33L lacks the *hbl* operon, suspected as a primary factor in diarrhea *B. cereus* food poisoning (*16*), Bt9727 has the *hbl* operon containing the hemolytic enterotoxin genes *hblCDBA* also found in *B. cereus* (ATCC 14579) (Figure 2A). Interestingly, the *hbl* gene cluster consists of the hemolytic enterotoxin (*hblCDBA*), and other genes encoding the spore germination proteins *gerIABC*, as well as other related proteins, that are ordered and oriented in a way that suggests their expression is coordinated by the transcriptional regulator TrrA. This gene cluster is part of a large, approximately 17.7 kb, 11 gene insertion (Figure 2B). A degenerate ISRso11 transposase fragment is found at the presumed insertion boundary region, and direct repeats that overlap with C-terminal *uvrC* like protein were identified. This observation suggests a

mechanism for the acquisition of these virulence factors in a lineage that includes Bt9727, *B.cereus* 14579 and *B. cereus* G9241.

The opportunistic pathogenicity of *B. cereus* and *B. thuringiensis* may depend on the secretion of non-specific extracellular virulence factors in response to transcriptional activation by PlcR (*17*). However in all B. anthracis strains, the *plcR* gene is inactivated by a frame shift mutation which creates an early stop codon (*18*). In other *B. cereus* isolates, the *plcR* gene product up regulates the transcription of genes encoding enterotoxins, proteases, phospholipases, metabolic enzymes, proteins involved in motility and chemotaxis, proteins involved in sporulation, DNA metabolism, transcriptional regulators, and a variety of transporters by binding to a specific upstream motif (*11, 17, 18*). The genes encoding PlcR appear intact in Bt9727 and in BcE33L. Analyzing the upstream sequences of coding regions for PlcR binding motifs identified Genes likely to be activated by PlcR in Bt9727 and BcE33L. We found motifs upstream of most of the genes previously identified as potential members of a *plcR* regulon in *B. cereus* (*11*) (Table S15). Of particular interest are genes encoding probable virulence factors. In this respect, we found that the nonhemolytic enterotoxin genes (*nheA, nheB and nheC*) in both Bt9727 and BcE33L contained upstream PlcR motifs. In Bt9727, BcE33L and *B. cereus* (ATCC 14579), there are PlcR motifs upstream of cytotoxin K and several proteases, including collagenase, bacillolysin, enhancin, aminopeptidase Y, and peptidase T. In addition, we found PlcR motifs upstream of the phospholipase C and phosphatidylinositol-specific phospholipase C genes in all three genomes. Another gene that has an upstream PlcR motif in Bt9727, BcE33L and *B. cereus* (ATCC 14579) is error prone DNA polymerase IV. This gene was previously suggested to induce adaptive point mutations that may affect pathogenicity. These observations (*11*) support the hypothesis that differences in virulence among *B. anthracis*, *B. cereus* and *B. thuringiensis* are predominately due to alterations in gene expression rather than simple gain or loss of gene functions.

Many microbial pathogens produce polymeric capsules that provide protection against host immune systems during the invasion process. *Bacillus* species can produce both polysaccharide capsules that are common to many Gram positive/Gram negative species and the less common polyglutamic acid capsule. A summary of the capsule

biosynthetic content in the sequenced members of the *B. cereus* group is provided in supplementary materials (Figure S3). In *B. anthracis*, three pXO2-encoded genes, *capB*, *capC*, and *capA* are required for synthesis of the polyglutamic acid capsule, and this structure plays a key role in the virulence of this organism (*19*). To date, all *B. cereus* group strains appear to contain a weak homolog of the pXO2-*capA* gene. This is also true in BcE33L, which contains a putitive protein with 32% identity to *capA*. However, BcE33L does not have a polyglutamic acid capsule (P.C.B Turnbull, personal communication, also table 1), nor does it appear to encode any genes involved in polysaccharide capsule synthesis. Interestingly, the Bt9727 genome and *B. cereus* 14579 encode a homolog of a member of a polysaccharide capsule synthesis pathway recently identified on a plasmid of the pathogenic *B. cereus* G9241 (*20*).

In anthrax, spores are the agent of infection. Spore formation occurs in response to nutrient limitation in the environment. In *B. subtilis*, sporulation is initiated by a deficiency in carbon or nitrogen (*21*) and is linked to changes in the expression of genes for degradative enzymes such as alpha amylase, neutral protease, and alkaline protease (*22*). The *B. subtilis* spore coat is composed of at least 15 polypeptides plus an insoluble protein fraction arranged in three morphological layers (*23*). We found differences in the number and composition of genes encoding spore coat proteins among Bt9727, BcE33L, *B. cereus* ATCC 14579 and *B. anthracis* (Table S16). BclA is the major spore surface antigen of *B. cereus* and *B. anthracis*, and is an exosporium depth determinant in *B. anthracis* (*24*). It is possible that BclA proteins affect spore morphology and surface properties. *B. anthracis* Ames and *B.cereus* ATCC 14579 each contain a single chromosomal copy of *bclA*, whereas *B.cereus* ATCC 10987 encodes both a chromosomal copy and a plasmid copy. Curiously, *B.cereus* BcE33L encodes 4 chromosomal copies and 1 plasmid copy, while Bt9727 encodes 3 chromosome copies of this gene.

Both *B. anthracis* and *B. cereus* spores germinate in response to L-alanine and ribosides (*25*-*27*); inosine is the most effective riboside germinant for *B. cereus* (*28*), while *B. anthracis* germination requires adenosine (*26*). The germination response to L-alanine and ribosides requires proteins of the *gerA* family(*25*). Bt9727, BcE33L, *B. cereus* 569, and *B. cereus* ATCC 14579 have a *gerI* operon that is involved in an inosine-induced germination. The *gerI* operon is homologous to the *gerA* family operons of *B.*

subtilis (*25*) and the *gerH* operon in *B. anthracis* (*27*). Similarly, the *gerQ* operon encodes germinant receptors that respond to inosine (*29*). Bt9727 encodes *gerQ*, while BcE33L does not.

Similar to *B. cereus* 14579 and *B. anthracis* (*10, 11*), Bt9727, and BcE33L appear predisposed to an environment rich in protein, having fewer genes for carbohydrate catabolism and more genes for amino acid metabolism (Table 2). Although the *B. cereu*s group species are closely related, variation in the sugar catabolism pathways is observed. Of particular note, BcE33L has 11 extra genes for carbohydrate polymer degradation compared with *B. anthracis* Ames. Most of these are located on the large plasmid pE33L466. One of the most significant differences between BcE33L and other isolates in *Bacillus cereus* group is the large number of sugar degradation gene clusters organized as operons on the pE33L466 plasmid (Figure 3).

The host range and virulence of Bt9727 and BcE33L may be increased through the presence of two lantibiotic resistance operons that are not present in *B. cereus* or *B. anthracis*. This includes a mersacidin resistance operon consisting of mrsR2, mrsK2, mrsF, mrsG, mrsE and a salivaricin resistance operon consisting of salY, salK and salR. Bt9727 and BcE33L have all of the genes in the mersacidin operon, while *B. anthracis* strains A2012, Ames and Sterne only have mrsF. Although Bt9727 and BcE33L have all of the genes in the salivaricin and mersacidin resistance operons, they do not encode the mrsA gene to produce mersacidin, or the salA gene to produce salivaricin. Therefore, these organisms can detect the presence of mersacidin and salivaricin produced by other bacteria, but do not encode the capability to produce these lantibiotics themselves. Instead, the response may include increased expression of genes encoding other lantibiotics or virulence factors as previously suggested (*30*).

There is considerable debate in regard to the systematic classification of members of the *B. cereus* group. Historically, these organisms were classified into three species (*B. cereus, B. thuringiensis,* and *B. anthracis*) on the basis of distinct phenotypic differences that defined them. For example, the isolation of an organism from an animal with anthrax resulted in the designation of *Bacillus anthracis*. While the relationship between these organisms is still not clearly understood, recent molecular approaches (*6, 7, 9, 10*) reveal extensive similarities between genomes and relatively few consistent differences

warranting the segregation of isolates into discrete species classified as *B. anthracis*, *B. cereus* and *B. thuringiensis.* One unifying concept that has emerged from nucleic acid sequence analyses (*5, 6, 9, 31*) is that the *B. cereus* group has evolved as asexually derived clonal populations. This has allowed most of the vast number of isolates from this group to be subdivided into consistent phylogenetic clusters.

In this classification scheme (9), Bt9727 and BcE33L are both members of the Anthracis lineage and are descended from ancestral clones that are very distinct from the Tolworthi, Kurstaki, Sotto and Thuringiensis lineages. Importantly, the Anthracis lineage provides a molecular based distinction that separates Bt9727 from the bulk of the commercially important insecticidal lineages. The Anthracis lineage also contains other pathogenic isolates such as a number of known B. cereus food pathogens (*9*).

The Bt9727 strain lacked the typical *cry*, *cyt* and *vip* genes encoding insecticidal proteins characteristic of strains that are known to produce entomopathic toxins. The original *B. thuringiensis* designation for this isolate was due to the discovery of crystals in the initial characterization of the strain (*8*).Certainly, the Bt9727 is distinct from other known *B. thuringiensis* isolates in that it is suspected of causing human morbidity (*8*) resulting in severe tissue necrosis. It was, subsequently, demonstrated to cause lethal infection in laboratory mice (*32*). Both phylogenetic lineage placement and subsequent laboratory diagnostics suggest that Bt9727 is more like a pathogenic *B. cereus* than an insecticidal strain.

Both Bt9727 and BcE33L have homologs of chromosomal virulence genes found in other members of the *B. cereus* group. Specifically, suspected virulence genes in *B. cereus* ATCC 14579 and *B. anthracis* Ames were detected by sequence similarity. Consequently, the isolation of Bt9727 from a rare case of disease, and the presence of common *B. cereus* group chromosomal virulence genes, make it likely that this organism is an opportunistic pathogen. While BcE33L came from a carcass swab, it is possible an environmental isolate and not the cause of death. The relationships between members of the *B. cereus* group are non-linear and complex likely resulting from cycles of isolation and niche expansion facilitated, at least in part, by horizontal gene transfer mechanisms. While the germination of *Bacillus anthracis* spores or its vegetative growth may be limited to nutritionally rich environments like that found in a mammalian host, the rapid

death of the host resulting from vegetative growth would limit the opportunity for genetic exchange and would result in the homogeneity observed in sequenced strains of this species. In contrast, the capacity for vegetative growth outside of an infected host or non-lethal infection provides an opportunity for genetic exchange and niche expansion. The sequences of the two *B. cereus* group members presented here provide fertile ground to study the evolution of host range and virulence.

REFERENCES

1. G. B. Jensen, B. M. Hansen, J. Eilenberg, J. Mahillon, *Environmental Microbiology* **5**, 631 (2003).
2. F. A. Drobniewski, *Clinical Microbiology Reviews* **6**, 324 (1993).
3. E. Schnepf *et al.*, *Microbiology and Molecular Biology Reviews* **62**, 775 (1998).
4. P. Keim *et al.*, *Journal of Bacteriology* **182**, 2928 (2000).
5. T. Pearson *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 13536 (2004).
6. E. Helgason *et al.*, *Applied and Environmental Microbiology* **66**, 2627 (2000).
7. L. Radnedge *et al.*, *Applied and Environmental Microbiology* **69**, 2755 (2003).
8. E. Hernandez, F. Ramisse, J. P. Ducoureau, T. Cruel, J. D. Cavallo, *Journal of Clinical Microbiology* **36**, 2138 (1998).
9. K. K. Hill *et al.*, *Applied and Environmental Microbiology* **70**, 1068 (2004).
10. T. D. Read *et al.*, *Nature* **423**, 81 (2003).
11. N. Ivanova *et al.*, *Nature* **423**, 87 (2003).
12. D. A. Rasko *et al.*, *Nucleic Acids Research* **32**, 977 (2004).
13. D. M. Guttmann, D. J. Ellar, *FEMS Microbiology Letters* **188**, 7 (2000).

14. J. Pannucci, R. T. Okinaka, R. Sabin, C. R. Kuske, *Journal of Bacteriology* **184**, 134 (2002).

15. R. T. Okinaka *et al.*, *Journal of Bacteriology* **181**, 6509 (1999).

16. D. J. Beecher, J. L. Schoeni, A. C. L. Wong, *Infection and Immunity* **63**, 4423 (1995).

17. S. Salamitou *et al.*, *Microbiology-Sgm* **146**, 2825 (2000).

18. H. Agaisse, M. Gominet, O. A. Okstad, A. B. Kolsto, D. Lereclus, *Molecular Microbiology* **32**, 1043 (1999).

19. J. W. Ezzell, S. L. Welkos, *Journal of Applied Microbiology* **87**, 250 (September 7-10, 1998, 1999).

20. A. R. Hoffmaster *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 8449 (2004).

21. P. J. Piggot, J. G. Coote, *Bacteriological Reviews* **40**, 908 (1976).

22. M. Honjo *et al.*, *Journal of Bacteriology* **172**, 1783 (1990).

23. J. Zhang, P. C. Fitz-James, A. I. Aronson, *Journal of Bacteriology* **175**, 3757 (1993).

24. P. Sylvestre, E. Couture-Tosi, M. Mock, *Molecular Microbiology* **45**, 169 (2002).

25. M. O. Clements, A. Moir, *Journal of Bacteriology* **180**, 6729 (1998).

26. Y. Hachisuk, *Japanese Journal of Microbiology* **13**, 199 (1969).

27. M. Weiner, T. D. Read, P. C. Hanna, *Journal of Bacteriology* **185**, 1462 (2003).

28. S. C. Warren, G. W. Gould, *Biochimica et Biophysica Acta* **170**, 341 (1968).

29. P. J. Barlass, C. W. Houston, M. O. Clements, A. Moir, *Microbiology (Reading)* **148**, 2089 (2002).

30. M. Upton, J. R. Tagg, P. Wescombe, H. F. Jenkinson, *Journal of Bacteriology* **183**, 3931 (2001).

31. F. G. Priest, M. Barker, L. W. J. Baillie, E. C. Holmes, M. C. J. Maiden, *Journal of Bacteriology* **186**, 7959 (2004).

32. E. Hernandez, F. Ramisse, T. Cruel, R. le Vagueresse, J. D. Cavallo, *FEMS Immunology and Medical Microbiology* **24**, 43 (1999).

ACKNOWLEDGEMENTS

We thank Peter Turnbull for carefully reading the manuscript, for his very constructive input, and for answering all of our questions about BcE33L. We also thank Martin Hugh-Jones for providing information about the natural history of *B. anthracis*. This program is supported by the U. S. Department of Energy under the contract No. W-7405-ENG-36.

AUSPICES

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.

Table 1. Major phenotypic characteristics of *B. cereus* group genomes

Genotype	*B.anthracis* Ames	*B.cereus* ATCC 10987	*B. cereus* ATCC 14579	*B. thuringiensis 97-27*	*B. cereus E33L*
Plasmid	pXO1 (189 kb) pXO2 (96 kb)	pBc10987 (208 kb)	Linear phage-like (15 kb) pBClin15	pBT9727(77 kb)	pE33L466 (466 kb) pE33L54 (54 kb) pE33L9 (9 kb) pE33L8 (8 kb) pE33L5 (5 kb)
Tripartite lethal toxin	Present	Absent	Absent	Absent	Absent
rRNA	33 copies	12 copies	39 copies	39 copies	39 copies
Urease gene cluster	Absent	Present	Absent	Absent	Absent
Xylose utilization genes	Absent	Present	Absent	Absent	Absent
Capsule synthesis					
Polysaccharide capsule	Absent	Present[a]	Present[a]	Present[a]	Absent
Polyglutamic acid capsule	Present[a]	Absent	Absent	Absent	Absent
Flagellar genes	fragmental	intact	intact	intact	intact
N-acetylgalactosamine degradation	Absent	Present	Absent	Absent	Present
Functional PlcR	Absent[e]	Present[e]	Present[e]	Present[e]	Present[e]
Phage[j]	4	4 (1 degenerate)	6 (1 linear plasmid)	7	18
arginine degradation					
Arginase genes	Present	Present	Present	Present	Present
Arginine deiminase genes	Absent	Present	Present	Present	Absent
B.cereus repeat 1 (bcr1)[k]	10 copies	72 copies	56 copies	19 copies	22 copies
Restriction enzymes	Absent	Present[h]	Present[h]	Partial	Absent
enterotoxin					
hemolytic enterotoxin HBL	Absent	Present	Absent	Present	Absent
nonhemolytic enterotoxin NHE	Present	Present	Present	Present	Present
diarrheal toxin BceT	Absent	Absent	Present	Present	Absent
cry genes	Absent	Absent	Absent	Absent	Absent

[a]The capsule biosynthesis region of 20 kb is partially similar in both *B. cereus* genomes (Figure 7).

[e]plcR in *B. anthracis* Ames contains a frameshift that results in a truncated and non-functional protein. The plcR gene in the *B.cereus* strains is full length and apparently functional and can act as a regulatory protein.

[h]There are four unique restriction–modification systems in *B. cereus* ATCC 10987 and three in *B.cereus* ATCC 14579. BT97 and BA have a CDS weakly similar to the 5-methylcytosine-specific Mrr endonuclease

[j]The phages are not conserved in sequence or genomic location in the three genomes studied.

[k]bcr1 is an 160 bp repeated DNA sequence with unknown function overwhelmingly over-represented in intergenic regions of the *B. cereus* group organisms.

Table 2. Number of protein and sugar utilization genes in *B. cereus* group

Amino acid and peptide utilization	*B. subtilis*	*B. anthracis* Ames	*B cereus* 14579	*B. cereus* E33L	*B.thuringiensis* 97-27
peptide ABC transporter, ATP-binding protein	7	18	23	20	18
branched chain Amino-acid trasporter	4	11	11	13	14
LysE/RhtB/CadD amino-acid efflux system	2	6	8	9	8
peptidase	30	64	91	92	90
protease	24	50	49	61	52
Amino acids and amines catabolism	34	52	55	55	55
(BA0242) tyrosine degradation	n	y	y	y	y
Epr Bpr AprX protease	y	n	n	n	n
Sugar utilization					
PTS-sugar	25	19	18	23	20
Carbohydrate polymer degradation	41	12	12	23	12
Mannose, Arabinose, Rhamnose catbolic pathway	y	n	n	n	n

FIGURE LEGENDS

Figure 1. An AFLP-based tree of *B.anthracis*, *B.cereus* and *B.thuringiensis* isolates . These 48 isolates are representative of the branches identified when over 300 isolates of *B.anthracis* , *B.thuringiensis* and *B.cereus* were examined by AFLP. Yellow highlighted isolates have fully sequenced genomes, blue are *B.thuringiensis* isolates, black are *B.cereus* isolates and red are *B.anthracis* isolates.

Figure 2A. Comparison of the gerI and hbl operon region in *B. cereus*, *B. thuringiensis*, and *B. anthracis*. The light blue area between the two groups indicate that these regions share a high level of identity. A conserved region consists of five contiguous genes in *B. anthracis* Ames, including L-asparaginase (BA3137), ans operon repressor (BA3138), degenerate ISRso11 transposase (BA3139), UvrC-like protein (BA3140) and amino acid permease(BA3141).

Figure 2B. Flanking region of insertion boundary. The orthologs of genes are shown as arrows of the same color. BT9727_2896/BA3140 encode uvrC like proteins. BT9727_2885.1/BA3139 encode degenerate ISRso11 transpoase. Yellow blocks denote the direct repeats found around the insertion boundary. The red triangle is the genes of C-terminal uvrC like protein fragment.

Figure 3. Schematic presentation of phosphotransferase system-catalyzed sugar uptake and phosphorylation in BCE33L showing possible metabolic pathways catalyzed by the products of genes in this polymorphic locus. Steps along the pathways are catalyzed by the gene products specified near the corresponding arrow. Genes with homologues in in *B. subtilis* are in bold print.

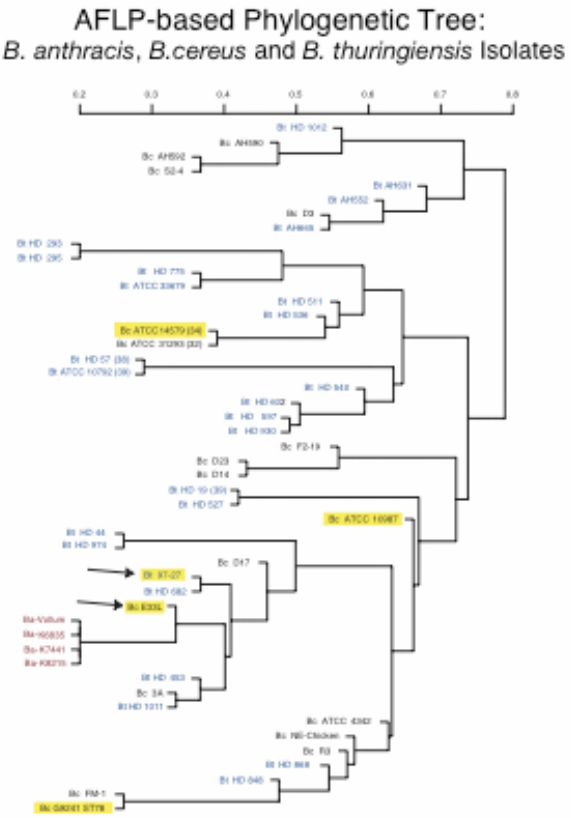

AFLP-based Phylogenetic Tree:
B. anthracis, B.cereus and B. thuringiensis Isolates

Figure 1.

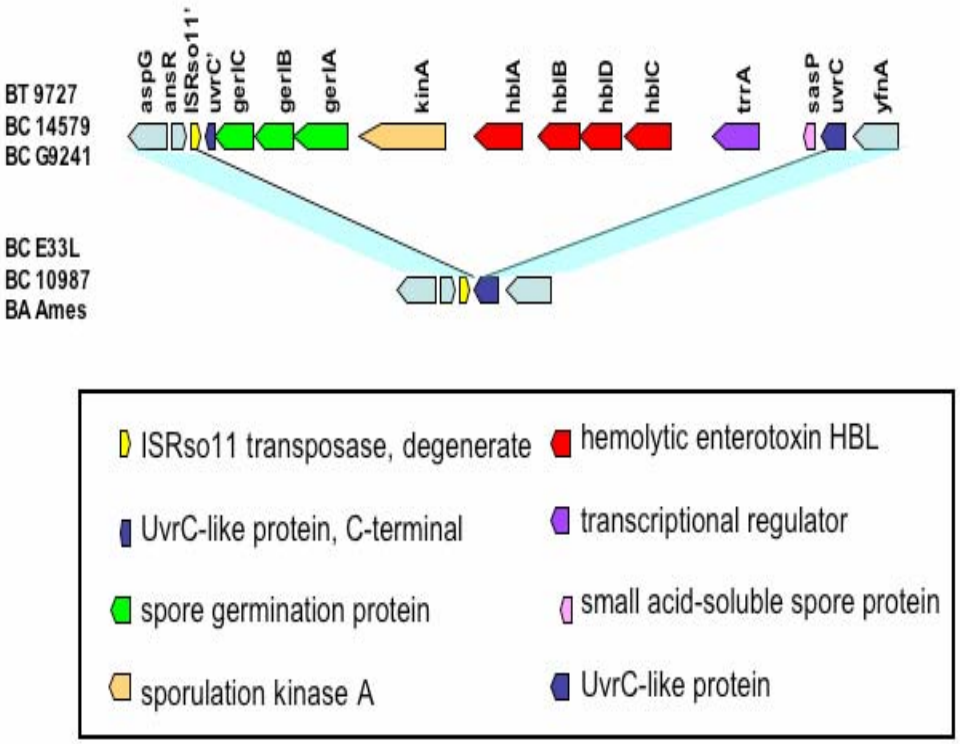

BT 9727
BC 14579
BC G9241
BC E33L
BC 10987
BA Ames
aspG
ansR
ISRso11'
uvrC'
gerIC
gerIB
gerIA
kinA
hblA
hblB
hblD
hblC
trrA
sasP
uvrC
yfnA
ISRso11 transposase, degenerate
UvrC-like protein, C-terminal
spore germination protein
sporulation kinase A
hemolytic enterotoxin HBL
transcriptional regulator
small acid-soluble spore protein
UvrC-like protein

Figure 2A.

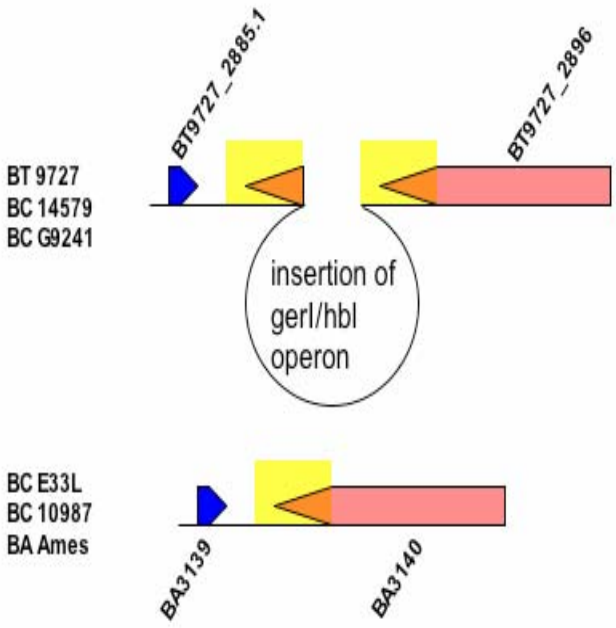

BT9727_2885.1
BT9727_2896
BT 9727
BC 14579
BC G9241
insertion of
gerI/hbl
operon
BC E33L
BC 10987
BA Ames
BA3139
BA3140

Figure 2B.

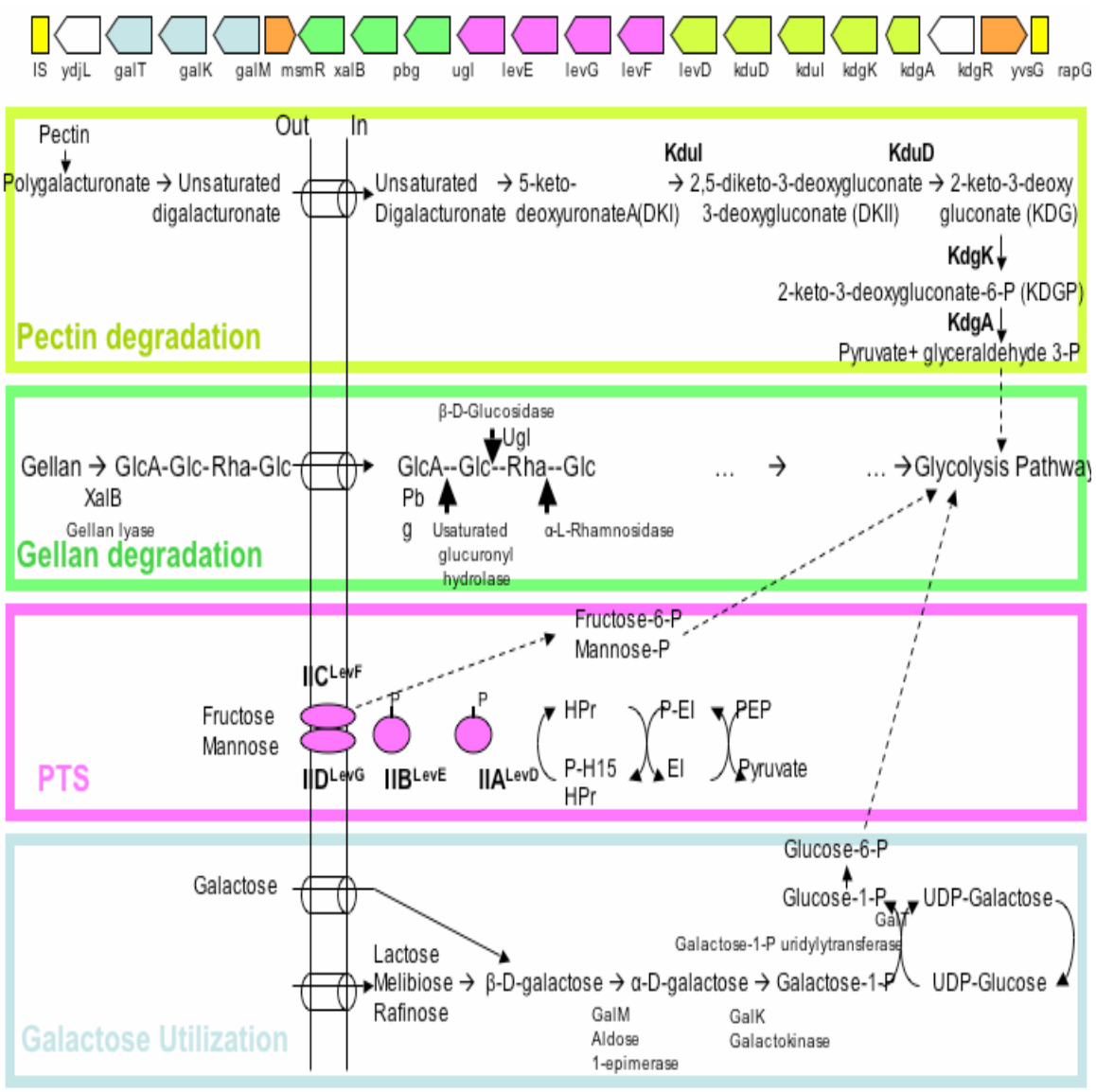

IS ydjL galT galK galM msmR xalB pbg ugl levE levG levF levD kduD kduI kdgK kdgA kdgR yvsG rapG
Pectin
Polygalacturonate → Unsaturated digalacturonate
Out
In
Unsaturated Digalacturonate → 5-keto-deoxyuronateA(DKI)
KduI
→ 2,5-diketo-3-deoxygluconate 3-deoxygluconate (DKII)
KduD
→ 2-keto-3-deoxy gluconate (KDG)
KdgK
2-keto-3-deoxygluconate-6-P (KDGP)
KdgA
Pyruvate+ glyceraldehyde 3-P
Pectin degradation
β-D-Glucosidase
Ugl
Gellan → GlcA-Glc-Rha-Glc
XalB
Gellan lyase
GlcA--Glc--Rha--Glc
Pb
g
Usaturated glucuronyl hydrolase
α-L-Rhamnosidase
... →
... →Glycolysis Pathway
Gellan degradation
Fructose-6-P
Mannose-P
IIC LevF
Fructose
Mannose
IID LevG
IIB LevE
IIA LevD
HPr
P-H15 HPr
P-EI
EI
PEP
Pyruvate
PTS
Glucose-6-P
Galactose
Glucose-1-P
UDP-Galactose
GalT
Galactose-1-P uridylytransferase
Lactose
Melibiose → β-D-galactose → α-D-galactose → Galactose-1-P
UDP-Glucose
Rafinose
GalM
Aldose 1-epimerase
GalK
Galactokinase
Galactose Utilization

Figure 3.

Supporting online material

MATERIALS AND METHODS

Sequencing of the *B. thuringensis* 97-27 and *B. cereus* E33L genomes

The random shotgun method of cloning, sequencing and assembly were used. Large (40 kb, BT97 only), median (8 kb) and small (2.5-3.5 kb) insert random sequencing libraries were sequenced for this genome project with average success rate of 90% and average high-quality read lengths of 643 and 621 nucleotides, for *B. thuringensis* 97-27 and *B. cereus* E33L, respectively. The completed genome sequences of *B. thuringensis* 97-27 and *B. cereus* E33L contained 134054 and 141352 reads respectively, achieving an average of 19.3 and18.7-fold sequence coverage per base. After assembly, gaps between contigs were closed by editing, walking library clones or linking contigs by PCR. The sequences of *B. thuringensis* 97-27 and *B. cereus* E33L genomes and plasmids can be accessed using the GenBank accession numbers AE017355, CP000001, CP000040, CP000041, CP000042, CP000043, CP000044, and CP000047.

Annotation

Gene predictions were obtained using Glimmer and tRNAs were identified using tRNAScan-SE. Basic analysis of the gene predictions was performed by comparing coding sequences against the PFam, BLOCKS and Prodom databases. Gene definitions and functional classes were added manually by a team of annotators, using BLAST results in addition to information from the basic analysis.

Sequence analyis

We compared the genomes at the nucleotide level using genome-alignment tools such as MUMmer2 (1), ACT (http://www.sanger.ac.uk/Software/ACT/) and Pipmaker (2). To obtain a list of orthologs in the *B. thuringensis* 97-27 and *B. cereus* E33L genomes, we wrote a perl script that determines bidirectional best hits as follows. Genes *g* and *h* are considered orthologs if *h* is the best BLASTP hit for *g* and vice versa, with *e*-values less than or equal to 10^{-15}. A gene is considered strain-specific if it has no hits with an *e*-value 10^{-15} or less. Analysis results can be visualized by the LANL developed Bugspray program (http://biosphere.lanl.gov/bugspray_std/cgi-bin/wc.cgi).

To identify IS elements in *B. thuringiensis* 97-27 and *B. cereus* E33L and compare them to IS elements present in other *B. cereus* group members, all known IS elements were used as query sequences and used with BLAST against the genomes of three strains of *B. anthracis* (Ames, A2012, and Sterne), *B. thuringiensis* 97-27, *B. cereus* E33L, and *B. cereus* (ATCC 14579).

Tandem repeats were identified in *B. thuringensis* 97-27 and *B. cereus* E33L genomes using Tandem Repeats Finder (Benson, 1999) with the threshold set for a minimum alignment score of 50.

Amplified fragment length polymorphism analysis

AFLP analysis of the microbial DNAs was accomplished as previously (3). Briefly, each of the DNA preparations was digested with EcoRI and MseI and the resulting fragments were ligated to double-stranded adapters then amplified by PCR using +0/+0 primers. Selective amplifications using the +1/+1 primer combination of FAM (6-carboxyfluorescein) labeled EcoRI-C and MseI-G resulted in products that were mixed with a solution containing DNA size standards (Genescan-500, Applied Biosystems Inc., Foster City, CA and MapMarker-400, BioVentures, Inc., Murfreesburo, TN) both labeled with TAMARA (N,N,N,N-tetramethyl-6-carboxyrhodamine). Following a two min heat denaturation at 90°C, the reactions were loaded onto a 5% Long Ranger DNA sequencing gel (Cambrex Bio Science, Rockland, ME) and visualized on an ABI 377 automated fluorescent sequencer (Applied Biosystems Inc., Foster City, CA). Each set of AFLP experiments also included as a sample, B. anthracis Vollum DNA, which was used as an internal control to allow comparison of results from different gels run at different times. Genescan analysis software (Applied Biosystems Inc., Foster City, CA) was used to determine the length of the sample fragments by comparison to the DNA fragment length size standards included within each sample.

Data analysis of the microbial DNAs was as previously described (4). DNA fragment sizes between 100 and 500 bp from triplicate data (derived from 3 lanes from 3 different gels) for each sample were combined. Fragment sizes that appeared in all 3 replicates were used to represent the sample and the peak heights for the fragment sizes were averaged. This “averaged” sample was then used to compare to other “averaged” samples. A hierachical agglomerative clustering routine using group averages was used to determine which fragments among the samples had similar lengths. A decision rule was added to this clustering routine that limited the allowable number of fragments within a cluster to equal the number of samples being compared and limited the maximum acceptable range of fragments sizes for a cluster to a preset value. Similarities between samples were measured using the Jaccard coefficient. Dendrograms were produced using the similarity matrix and the unweighted pair-group mean average method (UPGMA). (F.J. Rohlf, NTSYS-PC numerical taxonomy and multivariate analysis system, version 1.8; Exeter Software, Setauket, N.Y.)

Supplemental Text

General features of the Chromosomes of Bt9727 and BcE33L

Comparison of the genomes of *Bacillus thuringiensis 97-27* (Bt9727) and *Bacillus cereus E33L* (BcE33L) with *B. cereus* 14579 and *B. anthracis* Ames identified 357 common genes among Ames, Bt9727 and *B. cereus* E33L; there were no bidirectional best hits for these genes in *B. cereus* 14579 (see Figure S2). These 357 genes are either absolutely unique (there are no homologs in *B. cereus* 14579) or relatively unique (there may be homologs present in *B. cereus* 14579, but Ames, Bt9727 and BcE33L have paralogus genes)**.** The complete data are available upon request. These loci are often associated with strain-specific genotypes (Table 1). In BcE33L, sixteen of the 357 genes are located on the pE33L466 plasmid. The predicted products of these genes are diverse and include proteins involved in bacitracin synthesis and transport, oligopeptide transport, iron transport, flagellar biosynthesis, myo-inositol catabolism, glutamate biosynthesis, fatty acid and phospholipid metabolism, the oxidative branch of the pentose phosphate pathway, and prosthetic group cofactor biosynthesis. Fifteen of the 357 genes appear to encode transcriptional regulators.

Comparison of the genomes of BcE33L and Bt9727 with *B. anthracis* Ames and *B. cereus* 14579 also identified strain specific genes in each organism. Of the 5681 predicted BcE33L proteins, 93 of the chromosomally encoded genes and 149 of the plasmid genes were unique among this group. Of these, 5 of the chromosomal genes and 18 of the plasmid genes have an assigned function, whereas 219 are hypothetical or conserved hypothetical. Of the 5199 predicted Bt9727 proteins, 83 of the chromosomally encoded genes and 21 of the plasmid genes are unique in the group. Of these, 7 of the chromosomal genes and 1 of the plasmid genes have an assigned function, whereas 96 are hypothetical or conserved hypothetical.

In addition, approximately 30 pseudogenes and gene remnants are present in the genome of *B. thuringiensis* 97-27, and 23 pseudogenes and gene remnants are present in the BcE33L chromosome, not including fragmented IS elements. As many as 19 pseudogenes in Bt9727 chromosome are represented by intact orthologs in *B. cereus* E33L; 12 pseudogenes in BcE33L chromosome are represented by intact orthologs in Bt9727 chromosome. Furthermore, Bt9727 and BcE33L have 11 common pseudogenes; and suggest they represent elements that were present in a common ancestor. A complete list of pseudogene and gene remnants in both genomes is provided in Table S17. Analysis of these genes in Bt9727 and BcE33L indicates that half of them are redundant genes that have active paralogs in their respective genome suggesting a lack of strong selective pressure for their retention. The other half does not have paralogs, and are either putative enzymes or hypothetical proteins.

The replicative origin in Bt9727 and BcE33L chromosomes was identified by similarity to the *B. subtilis* origin (*5, 6*), through co-localization of the genes (rpmH, dnaA, dnaN, recF, and gyrA) often found near the Rep origin in prokaryotic genomes, and through GC nucleotide skew (G-C/G+C) analysis as illustrated in Figure S2. Based on these data, we designated base-pair 1 in an intergenic region located in the putative origin of replication. We located the replication termination site (terC), also by GC skew analysis. The region lies roughly opposite oriC in the circular chromosome, and is at about 2561 kb in Bt9727 and at 2571.7 kb in BcE33L (Figure S1). Similar to *B. subtilis*

and *B. anthracis*, there are rRNA, tRNA, and ribosomal protein genes around the oriC in Bt9727 and BcE33L.

General features of the Bt9727 and BcE33L plasmids

The largest of the BcE33L plasmids, pE33L466, is 466,370 nucleotides in length encoding 466 genes (Table S18). We compared pE33L466 with other sequenced large plasmids of the *B.cereus* group, including *B. anthracis* pXO1 (182 kb), pXO2 (95 kb), *B. cereus* G9241 pBC218 (218 kb), *B. thuringiensis* subsp. *israeliensis* pBtoxis (128 kb) and *B. cereus* 10987 pBc10987 (208 kb) using the program PipMaker (REF Schwartz et al, 2000). The comparative study reveals that these plasmids have variable molecular structures, in which gene order appears to be conserved in one part of the replicon and variable in another part of the replicon. Comparison of proteins from pE33L466, pBc10987 and pXO1 using three way bidirectional best hits revealed that 17 proteins were homologous among the three plasmids (Figure S1b); The low number of orthologous genes and the lack of extensive gene order conservation among these three plasmids suggest that they either have different origins or underwent substantial rearrangements. Comparison of the nucleotide sequences of the plasmids using NUCmer (REF Kurtz et al, 2004) reveals that pE33L466 and pBC218 are most closely related to each other, and there are 62 proteins that are bidirectional best hits between the two plasmids. Fifty one percent of the nucleotide sequence of pBC218 aligns with the pE33L466 DNA sequence at 90% identity, whereas pE33L466 and pBc10987 have only a minimum alignment of approximately 6 kbp.

Comparison of the Bt9727 plasmid pBt9727 with the BcE33L plasmids and *B. anthracis* plasmids revealed that greater than 70% of pBt9727 is strongly homologous to pXO2 (Figure S4). However, pBT9727 does not display strong homology to the other BcE33L plasmids, which suggests that pBT9727 and pXO2 share a common ancestry.

As illustrated in Figure S4, the plasmid pE33L466 reveals a bacterial chromosome-like GC skew (7). Other *B. cereus* group plasmids, such as pBC218, pXO1, pBtoxis, and pBc10987, display atypical bacterial chromosome-like GC skew, whereas pXO2 and pE33L54 do not show a GC skew at all. It has been suggested that GC skew and gene strand bias result from the mechanism used in bacterial chromosome replication (Rocha, 2004). Therefore, pE33L466 might share common replication machinery with host chromosomes.

The six plasmids in the two genomes can be partitioned into three groups according to the likely replication mechanism. There are no chromosomal or plasmid type replication genes on pE33L466. Although, a fragmental DNA polymerase III containing gamma and tau subunits has been detected. This observation is similar to pBC218, probably reflecting an ancient pseudogene in a common ancestor. Comparing the replication-related genes on the other Bt9727 and BcE33L plasmids with other known plasmids, we found that the 4 of the smaller plasmids in *B. cereus* E33L, pE33L5, pE33L8, pE33L9 and pE33L54, belong to a group which utilize a rolling circle replication mechanism. While pBT9727 belongs to the same group as pAMbeta1, and likely employs a theta replication mechanism. Each plasmid has 1-2 copies per cell with the exception of pE33L5, which has approximately 0.6 copies per cell.

There are 466 potential protein coding sequences on the large pE33L466 plasmid. Of these, 362 are either relatively unique or absolutely unique to BcE33L based on our

bidirectional best hit BLAST study. Most of these specialized genes are hypothetical proteins. A majority of the known genes on pE33L466 have homologs on other *B. cereus* group chromosomes. The 466 genes on pE33L466 were compared with genes in the genomes of *B. cereus* E33L, *B. thuringiensis* 97-27, *B. subtilis*, *B. anthracis*, *B. cereus* 14579 and *B. cereus* 10987 with blastp. There were between 78-132 bidirectional best hits in each of the comparisons. A majority of the homologus genes in these lists are known, a total of 140 genes, while there are only 35 hypothetical genes. However, as indicated, most of the potential coding sequences not found in *Bacillus* genomes are hypothetical genes, a total of 170 genes. The remainder of this set includes transposon or IS related genes (46 coding sequences) and in cludes genes known to be involved in cellular processes (85 sequences), such as energy metabolism, DNA replication and gene transcription.

Mobile DNA

The Bt9727 appears to contain seven putative prophage genes, each is isolated and not flanked by other identified phage-like genes. There are no identified phage-like genes on the plasmid pBT9727. The BcE33L genome appears to contain nine isolated putative prophage genes on the chromosome, six isolated putative prophage genes on the large pE33L466 plasmid, and three isolated putative prophage genes on pE33L54 plasmid (See Table S19). BcE33L has two regions that contain clusters of inserted phage genes, including one candidate lambdoid prophage on the chromosome (BCE33L3407 - BCE33L3461) containing genes that are also present on the *B. anthracis* Ames chromosome. Moreover genes BCE33L3407 - BCE33L3429 are ordered from lysis genes, tail fiber genes, tapemeasure gene, tail-related gene through capsid, portal and terminase genes matching the conserved gene order for temperate phage morphogenetic operons (33). Another phage gene cluster on the large plasmid (pE33L466_0162 - pE33L466_0165) contains genes also present on *B. anthracis* Ames and *B. cereus* ATCC14579.

There are at least two genes containing type I introns in both Bt9727 and BcE33L genomes. BT9727_3520 and BCE33L3538 are N-terminal homologues, while BT9727_3521 are BCE33L3539 are C-terminal homologues of recA, respectively. BT9727_1240 and BCE33L1242 encode N-termini while BT9727_1241 and BCE33L1243 encode C-termini of the ribonucleoside-diphosphate reductase alpha chain, respectively. The type I introns are in the intergenic space between the N- and C-terminal parts of these genes in both Bt9727and *B. cereus* E33L. In addition, BcE33L has coding sequences for the N-terminal and C-terminal regions of the terminase large subunit, which are interrupted by a GIY-YIG intron.

The *B.cereus* group has a large and diverse population of insertion sequence (IS) elements, which may have had a critical role in shaping these genomes (Table S20). Analysis of variation of the copy numbers, insertion location, sequence polymorphisms of inverted repeat, and target-site duplication revealed few differences among various strains of *B. anthracis*, suggesting that all *B. anthracis* strains have evolved recently(8). However, other *B. cereus* group species exhibited variation in copy numbers and insertion locations. For example, the Bt9727 chromosome contains 13 pairs of intact IS5 family elements compared to *B. cereus*, which have none, and the *B. anthracis* Ames pXO1 plasmid which has only an IS5 fragment. It is interesting that only BcE33L has

some intact IS elements that are distributed on both the plasmid and chromosome, including some with identical sequence that are present on both molecules, suggesting intra-replicon transposition. Examples include BCE33L2461 (IS4 family) and BCE33L2264 (IS605 family), which are present on the chromosome and have several homologs on the large plasmid of *B. cereus* E33L.

Although the *B. cereus* group genome is rich in IS elements, half of these elements appear to be nonfunctional due to INDELs and frameshift mutations that result in degenerate transposase genes. Despite the large number of IS elements in the Bt9727 and BcE33L genomes, IS insertions disrupt only one gene in Bt9727 (encoding putative phosphohydrolase (BT9727_2996)) and truncate two genes, one in Bt9727 and one in BcE33L (encoding gluconate permease (BT9727_4574) and L-lactate permease (BCE33L4922)). This suggests selection against insertions into most of the Bt9727 and *B. cereus* genes, or the occurrence of some form of repair mechanism to remove these IS elements.

Tandem repeat analysis

As observed in several other genomes, both Bt9727 and BcE33L tandem repeat pattern lengths that are a multiple of 3 far outnumbered repeat lengths that are not a multiple of three (9). This includes 62% of Bt9727 repeats and 64% of BcE33L repeats. In addition, BcE33L and Bt9727 display differences in the hypervariable regions previously identified by Keim et al.(10). The 12 nucleotide sequence (CAGCAATACTCA) present in the *vrrA* gene is found tandemly repeated 2.75 times in *B. cereus* E33L. This fits nicely within the 2-6 copies found in 426 isolates of *B. anthracis* previously analyzed (10), while the pattern in the *vrrA* gene repeat in Bt9727 is truncated to approximately 1.8 copies. In contrast, the repeat is absent in the pathogenic *B. cereus* strains ATCC 14579 and ATCC 10987. In the *vrrB* gene the BcE33L has two tandem repeats of 9 bp length with copy numbers of 3.8 and 14.2, while Bt9727 also has two 9 bp length repeats with copy numbers of 17.2 and 10.2. The *vrrC* region within the *ftsK*/*spoIIIE* gene is 216 bp shorter in BcE33L than in *B. thuringiensis* 97-27, and most of this difference occurs within two 12 bp tandem repeats.

A *B. cereus* group specific repeated element, *bcr1,* was identified previously in the *B. cereus* genome (*11, 12*). The *bcr1* sequences were identified by blastn searching against all finished *Bacillus* genome sequences. So far this repeated element appears to be present on the chromosomes of all *B. cereus* group species, but not on their plasmids. Analysis of tandem repeats in the Bt9727 plasmid, pBt9727, revealed a density of elements approximately 3.5 times greater than the frequency of repeats in the Bt9727 chromosome.

Motility

Both *B. thuringiensis* 97-27, and BcE33L are motile ((13) and PCB Turnbull, personal communication). Our analysis found a number of common flagellar genes in both Bt9727 and BcE33L (see Table S21). Similar to *B. cereus* 10987 both Bt9727 and BcE33L encode 2 flagellin subunits (*flaA* and *flaB*). By comparison, *B. cereus* 14579 has four of these flagellin subunits and *B. cereus* G9241 has five flagellin subunits (14).

One recent study (15) provides evidence for a link between virulence and swarming motility, indicating that coordinated export of flagellar and virulence proteins

exists in *B. thuringiensis*. Specifically, *flhA* plays a crucial role in the secretion of flagellin, hemolysin BL, and phosphatidylcholine-preferring phospholipase C(15). In comparison to *B. cereus* ATCC 14579 (16), the region I cluster of flagellar genes, *flgB, flgC, fliE, fliF, fliG, fliI, flgD,* and *flgE,* present in both of these organism does not appear to be regulated by PlcR, due to lack of the upstream recognition motif. Since the region I flagellar gene cluster in both Bt9727 and BcE33L may not be regulated by PlcR, these flagellar genes may not be expressed in coordination with other PlcR regulated virulence genes. Also of note is the failure to identify the sigma 28 form of RNA polymerase (SigD) which controls transcription of the flagellin, chemotaxis and motility genes (17) in *B. subtilis*. However, the *sigD* gene is not present in the *B. cereus* group species, including Bt9727 and *B. cereus* E33L. Furthermore, *L. monocytogenes*, which is motile and carries a flagellar operon lacks *sigD (18).* Together these observations suggest a sigD independent regulatory mechanism function in the *B. cereus* group organisms to regulate the transcription of motility genes.

Metabolic Potential

Like *B. cereus* 14579 and *B. anthracis* (*16, 19*), Bt9727, and BCE33L appear predisposed to an environment rich in protein, having fewer genes for carbohydrate catabolisn and more genes for amino acid metabolism (Table 2). For example, there are 12 carbohydrate polymer degradation genes in Ames, *B. cereus* 14579, and Bt9727 and 23 in *B. cereus* E33L, compared to 41 in *B. subtilis*. The *B. cereus* group also appears to have reduced numbers of sugar specific phosphoenolpyruvate-dependent phospho-transferase system (PTS). In contrast, members of the *B. cereus group* have an expanded capacity for amino acid and peptide utilization. For example, there are 18-23 genes encoding peptide/amino acid ABC transporter-ATP binding proteins in the *B. cereus* group, compared to 7 in *B .subtilis*. There are 6-9 genes for LysE family amino-acid efflux system proteins in *B. cereus* group members and only 2 in *B. subtilis*. In addition to the expanded number of peptidase and proteases in *B. cereus* group species, 52-55 genes encode proteins involved in amino-acid and amine catabolic pathways compared to 34 in *B. subtilis*. These observations suggest that proteins, peptides and amino acids may be a preferred nutrient source for all members of the *B. cereus* group, which are consistent with the observations made previously (*16, 19*).

Although the *B. cereu*s group species are closely related, variation in sugar catabolism pathways is observed between them. Of particular note, BCE33L has 11 additional genes for carbohydrate polymer degradation compared with *B. anthracis* Ames. Most of these are located on the large plasmid pE33L466. One of the most significant differences between BcE33L and other *Bacillus cereus* group species is the large number of sugar degradation gene clusters contained in an operon organization on the pE33L466 plasmid (Figure 1). These carbohydrate utilization genes provide activities for myo-inositol degradation, galactose ultization, pectin degradation, and gellan degradation

The BcE33L plasmid pE33L466 has the genes encoding the complete myo-inositol degradation pathway, suggesting that BcE33L can utilize myo-inositol as a source of both carbon and energy. It has been recently noted that inositol is present in soil and serves as an important substrate for *Rhizobium loti* and *Rhizobium fredii(20)*. Given

the likelihood that BcE33L is a soil bacterium, it is not surprising that it can metabolize inositol.

Many prokaryotes can metabolize galactose via two possible pathways depending on the mode of transport of the sugar. When lactose or galactose is transported via a phosphoenolpyruvate-dependent phospho-transferase system (PTS), the resulting galactose 6-phosphate is metabolized by the enzymes of the tagatose 6-phosphate pathway (21). In the case of galactose or lactose entry via a permease, the sugar is not phosphorylated, and the galactose is degraded via the Leloir pathway. While BcE33L encodes both pathways, *B. subtilis*, *B. anthracis* Ames, and Bt9727 only have the tagatose 6-phosphate pathway (Figure S5).

Both *B. subtilis* and the BcE33L plasmid 466 encode a pectin degradation gene cluster, however these genes are absent in *B. anthracis* Ames, *B. cereus* 14579 and *B. thuringiensis* 97-27. This finding suggests that, like *B. subtilis* (*22, 23*), BcE33L may use pectin as a source of carbon. Pectin is a major component of the plant cell wall. While *B. cereus* isolates have been found on oak leaf litter (24), these strains were unable to hydrolyze pectin but instead hydrolyzed starch.

Gellan is a heteropolysaccharide produced by the gram-negative bacterium *Sphingomonas paucimobilis* (25). The gellan degradation pathway is not widely distributed, and genes for all of the pathway components have been identified in only *Bacillus* GL1 (26) and pE33L466. This pathway is not present in *B. subtilis*, *B. anthracis* Ames, and *B. thuringiensis* 97-27. One of the key enzymes of this pathway is a galactosidase, highly similar to beta-galactosidases from animal, plants, and fungi, implying an evolutionary relationship between this novel gene and those of eukaryotic origins (27)**.** Although several of these enzymes can participate in the xanthan degradation pathway, it is not likely that BcE33L has a functional xanthan pathway, since alpha-mannosidase, an important enzyme required for xanthan degradation, is absent in *B. cereus* E33L.

It is worth noting that the region of pE33L466 containing these four interlocked sugar metabolism gene clusters is flanked by IS elements that were probably involved in their mobilization and integration into pE33L466. Figure 3 illustrates the metabolic pathways in which the products of these genes participate.

Other variations in carbohydrate utilization pathways among the *B. cereus* group organisms involve the ability (or lack of) to use other carbon sources such as sucrose, myo-inositol, chitosan and xylose. While the *bgl* operon (*bglG, bglP*and *bglH*) is not represented in *B. anthracis* Ames and *B. cereus* 14579, this cluster of 3 genes is present in Bt9727 and BcE33L encoding a transcription antiterminator, a beta-glucoside-specific enzyme II of PTS system, and a 6-phospho-beta-glucosidase. The products of these genes likely play an important role in carbohydrate catabolism The BcE33L plasmid pE33L466 also encodes an endoglucanase, indicating that BcE33L may be able to hydrolyze 1,4-beta-D-glucosidic linkages in cellulose, lichenin and cereal beta-D-glucans to yield glucose (28).

In addition, *B. cereus* E33L, similar to *B. cereus* ATCC 10987, contains a 17.9 kb sequence segment encoding the functions for the transport and utilization of tagatose. It has been suggested previously that the tagatose gene cluster may be an environmental adaptation (12). The comparative organization of this operon in these two isolates is illustrated in Figure S5. In contrast, *B. thurindiensis* 97-27 and *B. anthracis* encode a

gene for the carbon starvation protein cstA, which encodes a peptide transporter (29) and is involved in peptide utilization (30).

In *B. subtilis*, the most important source of nitrogen is glutamine (31). Glutamine functions as an amino acid, and is easily converted to glutamate, which is the main nitrogen donor for synthesis of amino acids and nucleotides (32). Currently, not much is known about nitrogen metabolism in *B. cereus* and *B. anthracis* strains. However, Bt9727 and BcE33L have the genes encoding enzymes GlnA and GltAB (33) which are required to use glutamine and glutamate respectively. In some organisms, glutamate is derived from ammonium by glutamate dehydrogenase. In *B. subtilis*, there are 2 glutamate dehydrogenase enzymes, rocG and gudB, but only rocG is functional (34). In *B. thuringiensis* 97-27, *B. cereus* E33L, 14579, 10987, G9241, and *B. anthracis* Ames, Sterne, and A2012 two glutamate dehydrogenase genes are present, one with 75% identity to gudB and a rocG with 85% identity to their *B. subtilis* homologs.

Arginine, another important source of nitrogen, is degraded by both arginase-dependent and arginine deiminase-dependent pathways in *Bacillus* species. In *B. licheniformis*, the arginase pathway is preferentially utilized during aerobic growth (35)**.** In contrast, the arginine deiminase pathway is favored during anaerobic growth(36), where its primary role is to generate ATP in the absence of oxygen(35). As in *B. cereus* ATCC 14579 (16) and *B. licheniformis* (36), arginine is degraded by two alternative routes in *B. thuringiensis* 97-27: the arginase-dependent and arginine deiminase-dependent pathways. Only the arginase-dependent pathway is encoded by *B. subtilis, B. anthracis*, and *B. cereus* E33L.

Ammonium, which is produced by the arginine deiminase and urease pathways, is another source of nitrogen and can be taken up from the environment by an ammonium transporter NrgA in *B. subtilis* (31). We found that the *ure* genes encoding components of the urease pathway, which are encoded in *B. cereus* ATCC 10987 (12), are absent in *B. thuringiensis* 97-27, *B. cereus* E33L, *B. anthracis* Ames and *B. cereus* ATCC 14579. Whereas, the *nrgA* gene is encoded by *B. thuringiensis* 97-27, *B. cereus* E33L, *B. cereus* ATCC 14579, *B. anthracis* Sterne and A2012it is absent from *B. anthracis* Ames.

1. A. L. Delcher, A. Phillippy, J. Carlton, S. L. Salzberg, *Nucleic Acids Res* **30**, 2478 (Jun 1, 2002).
2. S. Schwartz *et al.*, *Genome Research* **10**, 577 (2000).
3. K. K. Hill *et al.*, *Applied and Environmental Microbiology* **70**, 1068 (2004).
4. L. O. Ticknor *et al.*, *Applied and Environmental Microbiology* **67**, 4863 (2001).
5. F. Kunst *et al.*, *Nature (London)* **390**, 249 (1997).
6. E. P. C. Rocha, *Microbiology (Reading)* **150**, 1609 (2004).
7. J. R. Lobry, *Molecular Biology and Evolution* **13**, 660 (1996).
8. P. Keim *et al.*, *Journal of Bacteriology* **179**, 818 (1997).
9. G. Benson, *Nucleic Acids Research* **27**, 573 (1999).
10. P. Keim *et al.*, *Journal of Bacteriology* **182**, 2928 (2000).
11. O. A. Okstad, I. Hegna, T. Lindback, A. L. Rishovd, A. B. Kolsto, *Microbiology-Sgm* **145**, 621 (1999).

12. D. A. Rasko *et al.*, *Nucleic Acids Research* **32**, 977 (2004).
13. E. Hernandez, F. Ramisse, J. P. Ducoureau, T. Cruel, J. D. Cavallo, *Journal of Clinical Microbiology* **36**, 2138 (1998).
14. A. R. Hoffmaster *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 8449 (2004).
15. E. Ghelardi *et al.*, *Journal of Bacteriology* **184**, 6424 (2002).
16. N. Ivanova *et al.*, *Nature* **423**, 87 (2003).
17. D. B. Mirel, M. J. Chamberlin, *Journal of Bacteriology* **171**, 3095 (1989).
18. P. Glaser *et al.*, *Science (Washington D C)* **294**, 849 (2001).
19. T. D. Read *et al.*, *Nature* **423**, 81 (2003).
20. M. Wood, A. P. Stanway, *Soil Biology and Biochemistry* **33**, 375 (2001).
21. D. L. Bissett, R. L. Anderson, *Biochemical and Biophysical Research Communications* **52**, 641 (1973).
22. A. Jauneau, O. Morvan, C. Morvan, M. Demarty, G. Devauchelle, *Comptes Rendus de L Academie Des Sciences Serie Iii-Sciences de La Vie-Life Sciences* **302**, 641 (1986).
23. J. M. Labavitch, L. E. Freeman, P. Albersheim, *Journal of Biological Chemistry* **251**, 5904 (1976).
24. B. Brunel, C. Perissol, M. Fernandez, J. M. Boeufgras, J. Lepetit, *FEMS Microbiology Ecology* **14**, 331 (1994).
25. I. Sa-Correia *et al.*, *Journal of Industrial Microbiology and Biotechnology* **29**, 170 (2002).
26. W. Hashimoto *et al.*, *Archives of Biochemistry and Biophysics* **339**, 17 (1997).
27. W. Kuchinke, *Journal of Molecular Evolution* **29**, 95 (1989).
28. S. R. Chhabra, K. R. Shockley, D. E. Ward, R. M. Kelly, *Applied and Environmental Microbiology* **68**, 545 (2002).
29. A. K. Dubey *et al.*, *Journal of Bacteriology* **185**, 4450 (2003).
30. J. E. Schultz, A. Matin, *Journal of Molecular Biology* **218**, 129 (1991).
31. S. H. Fisher, M. Debarbouille, in *Bacillus subtilis and its closest relatives: from genes to cells* A. L. Sonenshein, J. A. Hoch, R. Losick, Eds. (ASM Press, Washington, DC., 2002) pp. 181-191.
32. C. Detsch, J. Stulke, *Microbiology-Sgm* **149**, 3289 (2003).
33. B. R. Belitsky, in *Bacillus subtilis and its closest relatives: from genes to cells* A. L. Sonenshein, J. A. Hoch, R. Losick, Eds. (ASM Press, Washington, DC., 2002) pp. 203-231.
34. B. R. Belitsky, A. L. Sonenshein, *Journal of Bacteriology* **180**, 6298 (1998).
35. K. Broman, N. Lauwers, V. Stalon, J. M. Wiame, *Journal of Bacteriology* **135**, 920 (1978).
36. A. Maghnouj, T. F. D. Cabral, V. Stalon, C. Vander Wauven, *Journal of Bacteriology* **180**, 6468 (1998).

Supplementary Table S15. Categories of chromosomal genes in *B. thuringiensis* 9727 and *B. cereus* E33L regulated by PlcR

Category	BCE33L	BT9727	BCER
Toxins	BCE33L1698(nheA) BCE33L1699(nhe) BCE33L1700(nheC) BCE33L2999 (perfringolysin O) BCE33L1009(cytotoxin K)	BT9727_1727(nheA) BT9727_1728(nhe) BT9727_1729(nheC) BT9727_2893(hblC) BT9727_2892(hblD) BT9727_2891(hblB) BT9727_2890(hblA) BT9727_3096 (perfringolysin O) BT9727_1008(cytotoxin K)	BC1809(nhe) BC1810 BC1811 BC3104(hbl) BC3103(hbl) BC3102(hbl) BC3101(hbl) BC5101 (Perfringolysin O precursor) BC1110(cytotoxin K)
Proteases	BCE33L3514 (serine protease/ collagenase) BCE33L2464(bacillolysin) BCE33L3092(enhancin) BCE33L3091(bacillolysin) BCE33L5056 (aminopeptidase) BCE33L3898 (peptidase T)	BT9727_3502 (serine protease/ collagenase) BT9727_2499(bacillolysin) BT9727_3171(enhancin) BT9727_3170(bacillolysin) BT9727_5040 (aninopeptidase) BT9727_3891 (peptidase T)	BC3161(collagenase) BC3762(collagenase) BC2735(bacillolysin) BC3384(enhancin) BC3383(Bacillolysin) BC5351(bacillolysin) BC5359 (aminopeptidase Y) BC4143 (peptidase T)
Cell wall and surface proteins	BCE33L3515 (cell wall hydrolase) BCE33L3516(cons hypo) BCE33L4258 (poss. acid phosphatase) BCE33L4550 (s-layer protein)	BT9727_3503 (cell wall hydrolase) BT9727_3504(cons hypo) BT9727_4246 (poss. acid phosphatase) BT9727_4533 (s-layer protein)	BC0991 (S-layer/ endopeptidase) BC0992(cons hypo) BC4511 (acid phosphatase) BC2464(s-layer protein/endo…)
Metabolic enzymes, etc.	BCE33L1950(cell division and morphogenesis-related protein ScdA) BCE33L1949(nitrite reductase [NAD(P)H], large subunit) BCE33L1948(nitrite reductase [NAD(P)H]) BCE33L1947(uroporphiryn-III C-methyltransferase) BCE33L1946(CbiX-like) BCE33L1945(probable protein involved in nitrite reduction)	BT9727_1971(cell division and morphogenesis-related protein ScdA) BT9727_1970(nitrite reductase [NAD(P)H], large subunit) BT9727_1969(nitrite reductase [NAD(P)H]) BT9727_1968(uroporphiryn-III C-methyltransferase) BT9727_1967(CbiX-like) BT9727_1966(probable protein involved in nitrite reduction)	BC2137(dnrN; scdA)) BC2136(Nitrite reductase [NAD(P)H] large subunit) BC2135(Nitrite reductase [NAD(P)H] small subunit) BC2134(Uroporphyrin-III C-methyltransferase) BC2133(CbiX) BC2132(Ferrochelatase) BC1350(hypo) BC1351(hypo) BC1352(hypo) BC1353(NrdI) BC1354(Ribonucleoside-diphosphate reductase alpha chain) BC1355(Ribonucleoside-diphosphate reductase beta chain) B5335(Fructose-

			bisphosphate aldolase) BC4957(hypo) BC4958(NAD(P)H dehydrogenase [quinone]) BC3184(Phosphoglycerate mutase)
Motility and chemotaxis proteins	BCE33L0488 (methyl-accepting chemotaxis protein) BCE33L0489 (sensor kinase) BCE33L0490 (response regulator) BCE33L0491 (Na/Malate symporter) BCE33L0492 (NAD-malic enzyme) BCE33L3093 (methyl-accepting chemotaxis protein)	BT9727_0486 (methyl-accepting chemotaxis protein) BT9727_0487 (sensor kinase) BT9727_0488 (response regulator) BT9727_0489 (Na/Malate symporter) BT9727_0490 (NAD-malic enzyme)	BC0576 (methyl-accepting chemotaxis protein) BC0577 (sensor kinase) BC0578 (response regulator) BC0579 (Na/Malate symporter) BC0580 (NAD-malic enzyme) BC1641(flgB) BC1642(flgC) BC1643(fliE) BC1644(fliF) BC1645(fliG) BC1646(hypo) BC1647(fliI) BC1648(hypo) BC1650(flgD) BC1651(flgE) BC1652(hypo) BC3385 (methyl-accepting chemotaxis protein)
Sporulation	 BCE33L4986(SpoIID) BCE33L4985(ABC transporter, ATP-binding protein) BCE33L4984(BcrA) BCE33L4983(ABC trans permease) BCE33L4982(SpoIIQ)	 BT9727_4968(SpoII D) BT9727_4967(Transcriptional regulator, ABC trans, LytR) BT9727_4966(BcrA) BT9727_4965(ABC trans permease) BT9727_4964(SpoIIQ)	BC4794(spore ger. protein PF) BC3528(sporulation kinase) BC5287(SpoII D) BC5286(Transcriptional regulator, ABC trans, LytR) BC5285(BcrA) BC5284(ABC trans permease) BC5283(SpoIIQ)
Transcriptional regulators	BCE33L5049(plcR) BCE33L3497(adaA) BCE33L2498(DeoR family) BCE33L2497(acetyltransferase) BCE33L2496(BC group-specific) BCE33L2495(sigma-54-dependent transc.activ) PapR not found by Glimmer – but blast of BC5349 seq suggests it may be there between 5048 and 5049 BCE33L3497(AraC family)	BT9727_5033(plcR) BT9727_3485(adaA) No site upstream of BT9727_1731(DeoR) or BT9727_2354(transcr. regulator) PapR not found by Glimmer – but blast of BC5349 seq suggests it may be there between 5032 and 5033 BT9727_0944(poss. trans regul)	BC5350(PlcR) BC3740(AdaA) BC2770(DeoR family) BC2769(acetyltransferase) BC2768(hypo cytosolic) BC2767(hypo membrane) BC2766(sigma-54-dependent transc.activ) BC5349(PapR) BC3194(MarR family) BC1715(trans. regulator) BC1716(Na+ driven multidrug efflux pump)
Phospholipases	BCE33L0588(plc) BCE33L0589	BT9727_0587(plc) BT9727_0588	BC0670(plc) BC0671

	(sphingomyelin phosphodiesterase) BCE33L0590 (cons hypo) BCE33L0591 (FAD-dependent oxidoreductase) BCE33L3513(pi-plc)	(sphingomyelin phosphodiesterase) BT9727_0589 (cons hypo) BT9727_0590 (FAD-dependent oxidoreductase) BT9727_3501(pi-plc)	(sphingomyelin phosphodiesterase) BC0672 (hypo) BC0673 (flavin-dep dehydrog) BC3761(pi-plc)
Transporters/ antibiotic efflux	BCE33L0373(terD) BCE33L0374(terD) BCE33L0375(terD) BCE33L0376(terC-like) BCE33L0377 BCE33L0378(tellurite resistance protein) BCE33L4257(ABC transporter, ATP-binding protein; sodium export ATP-binding protein) no permease BCE33L1608(multidrug ABC transporter, ATP-binding protein) BCE33L1609(multidrug ABC transporter, permease) BCE33L1610(multidrug ABC transporter, permease)	BT9727_0376(terD) BT9727_0377(terD) BT9727_0378(terD) BT9727_0379(terC-like) BT9727_0380 BT9727_0381(tellurite resistance protein) BT9727_4245(ABC transporter, ATP-binding protein; sodium export ATP-binding protein) no permease BT9727_1639(multidrug ABC transporter, ATP-binding protein) BT9727_1640(multidrug ABC transporter, permease) BT9727_1641(multidrug ABC transporter, permease)	BC0442(terD) BC0443(terD) BC0444(terD) BC0445(terC-like) BC0446 BC0447(tellurite resistance protein) BC4510 (Sodium export ATP-binding protein) BC4509 (Sodium export permease protein) BC2410(TetR family) BC2411(macrolide efflux protein) BC1734(export ABC transporter, ATP-binding) BC1735(permease) BC1736(permease) BC5093 (xanthine permease) BC4140(Mg(2+) transport ATPase, P-type) BC4141(Mg(2+) transport ATPase protein C)
DNA metabolism	BCE33L3496(exodeoxyribonuclease III) BCE33L3497(AdaA) BCE33L3498(methylated-DNA--protein-cysteine S-methyltransferase) BCE33L3499(DNA-3-methyladenine glycosidase) BCE33L3897(DNA polymerase IV)	BT9727_3484(exodeoxyribonuclease III) BT9727_3485(AdaA) BT9727_3486(methylated-DNA--protein-cysteine S-methyltransferase) BT9727_3487(DNA-3-methyladenine glycosidase) BT9727_3889(DNA polymerase IV)	BC3739(Exodeoxyribonuclease III) BC3740(Ada regulatory protein) BC3741(O6-methylguanine-DNA methyltransferase) BC3742(DNA-3-methyladenine glycosylase II) BC4142(DNA polymerase IV)

Supplementary Table S15. Categories of chromosomal genes in *B. thuringiensis* 9727 and *B. cereus* E33L regulated by PlcR

Category	BCE33L	BT9727	BCER
Toxins	BCE33L1698(nheA) BCE33L1699(nhe) BCE33L1700(nheC)	BT9727_1727(nheA) BT9727_1728(nhe) BT9727_1729(nheC)	BC1809(nhe) BC1810 BC1811

	BCE33L2999 (perfringolysin O) BCE33L1009(cytotoxin K)	BT9727_2893(hblC) BT9727_2892(hblD) BT9727_2891(hblB) BT9727_2890(hblA) BT9727_3096 (perfringolysin O) BT9727_1008(cytotoxin K)	BC3104(hbl) BC3103(hbl) BC3102(hbl) BC3101(hbl) BC5101 (Perfringolysin O precursor) BC1110(cytotoxin K)
Proteases	BCE33L3514 (serine protease/ collagenase) BCE33L2464(bacillolysin) BCE33L3092(enhancin) BCE33L3091(bacillolysin) BCE33L5056 (aminopeptidase) BCE33L3898 (peptidase T)	BT9727_3502 (serine protease/ collagenase) BT9727_2499(bacillolysin) BT9727_3171(enhancin) BT9727_3170(bacillolysin) BT9727_5040 (aninopeptidase) BT9727_3891 (peptidase T)	BC3161(collagenase) BC3762(collagenase) BC2735(bacillolysin) BC3384(enhancin) BC3383(Bacillolysin) BC5351(bacillolysin) BC5359 (aminopeptidase Y) BC4143 (peptidase T)
Cell wall and surface proteins	BCE33L3515 (cell wall hydrolase) BCE33L3516(cons hypo) BCE33L4258 (poss. acid phosphatase) BCE33L4550 (s-layer protein)	BT9727_3503 (cell wall hydrolase) BT9727_3504(cons hypo) BT9727_4246 (poss. acid phosphatase) BT9727_4533 (s-layer protein)	BC0991 (S-layer/ endopeptidase) BC0992(cons hypo) BC4511 (acid phosphatase) BC2464(s-layer protein/endo…)
Metabolic enzymes, etc.	BCE33L1950(cell division and morphogenesis-related protein ScdA) BCE33L1949(nitrite reductase [NAD(P)H], large subunit) BCE33L1948(nitrite reductase [NAD(P)H]) BCE33L1947(uroporphiryn-III C-methyltransferase) BCE33L1946(CbiX-like) BCE33L1945(probable protein involved in nitrite reduction)	BT9727_1971(cell division and morphogenesis-related protein ScdA) BT9727_1970(nitrite reductase [NAD(P)H], large subunit) BT9727_1969(nitrite reductase [NAD(P)H]) BT9727_1968(uroporphiryn-III C-methyltransferase) BT9727_1967(CbiX-like) BT9727_1966(probable protein involved in nitrite reduction)	BC2137(dnrN; scdA)) BC2136(Nitrite reductase [NAD(P)H] large subunit) BC2135(Nitrite reductase [NAD(P)H] small subunit) BC2134(Uroporphyrin-III C-methyltransferase) BC2133(CbiX) BC2132(Ferrochelatase) BC1350(hypo) BC1351(hypo) BC1352(hypo) BC1353(NrdI) BC1354(Ribonucleoside-diphosphate reductase alpha chain) BC1355(Ribonucleoside-diphosphate reductase beta chain) B5335(Fructose-bisphosphate aldolase) BC4957(hypo) BC4958(NAD(P)H dehydrogenase [quinone]) BC3184(Phosphoglycerate mutase)
Motility and	BCE33L0488	BT9727_0486	BC0576

chemotaxis proteins	(methyl-accepting chemotaxis protein) BCE33L0489 (sensor kinase) BCE33L0490 (response regulator) BCE33L0491 (Na/Malate symporter) BCE33L0492 (NAD-malic enzyme) BCE33L3093 (methyl-accepting chemotaxis protein)	(methyl-accepting chemotaxis protein) BT9727_0487 (sensor kinase) BT9727_0488 (response regulator) BT9727_0489 (Na/Malate symporter) BT9727_0490 (NAD-malic enzyme)	(methyl-accepting chemotaxis protein) BC0577 (sensor kinase) BC0578 (response regulator) BC0579 (Na/Malate symporter) BC0580 (NAD-malic enzyme) BC1641(flgB) BC1642(flgC) BC1643(fliE) BC1644(fliF) BC1645(fliG) BC1646(hypo) BC1647(fliI) BC1648(hypo) BC1650(flgD) BC1651(flgE) BC1652(hypo) BC3385 (methyl-accepting chemotaxis protein)
Sporulation	 BCE33L4986(SpoIID) BCE33L4985(ABC transporter, ATP-binding protein) BCE33L4984(BcrA) BCE33L4983(ABC trans permease) BCE33L4982(SpoIIQ)	 BT9727_4968(SpoII D) BT9727_4967(Transcriptional regulator, ABC trans, LytR) BT9727_4966(BcrA) BT9727_4965(ABC trans permease) BT9727_4964(SpoIIQ)	BC4794(spore ger. protein PF) BC3528(sporulation kinase) BC5287(SpoII D) BC5286(Transcriptional regulator, ABC trans, LytR) BC5285(BcrA) BC5284(ABC trans permease) BC5283(SpoIIQ)
Transcriptional regulators	BCE33L5049(plcR) BCE33L3497(adaA) BCE33L2498(DeoR family) BCE33L2497(acetyltransferase) BCE33L2496(BC group-specific) BCE33L2495(sigma-54-dependent transc.activ) PapR not found by Glimmer – but blast of BC5349 seq suggests it may be there between 5048 and 5049 BCE33L3497(AraC family)	BT9727_5033(plcR) BT9727_3485(adaA) No site upstream of BT9727_1731(DeoR) or BT9727_2354(transcr. regulator) PapR not found by Glimmer – but blast of BC5349 seq suggests it may be there between 5032 and 5033 BT9727_0944(poss. trans regul)	BC5350(PlcR) BC3740(AdaA) BC2770(DeoR family) BC2769(acetyltransferase) BC2768(hypo cytosolic) BC2767(hypo membrane) BC2766(sigma-54-dependent transc.activ) BC5349(PapR) BC3194(MarR family) BC1715(trans. regulator) BC1716(Na+ driven multidrug efflux pump)
Phospholipases	BCE33L0588(plc) BCE33L0589 (sphingomyelin phosphodiesterase) BCE33L0590 (cons hypo) BCE33L0591 (FAD-dependent oxidoreductase)	BT9727_0587(plc) BT9727_0588 (sphingomyelin phosphodiesterase) BT9727_0589 (cons hypo) BT9727_0590 (FAD-dependent oxidoreductase)	BC0670(plc) BC0671 (sphingomyelin phosphodiesterase) BC0672 (hypo) BC0673 (flavin-dep dehydrog)

	BCE33L3513(pi-plc)	BT9727_3501(pi-plc)	BC3761(pi-plc)
Transporters/ antibiotic efflux	BCE33L0373(terD) BCE33L0374(terD) BCE33L0375(terD) BCE33L0376(terC-like) BCE33L0377 BCE33L0378(tellurite resistance protein) BCE33L4257(ABC transporter, ATP-binding protein; sodium export ATP-binding protein) no permease BCE33L1608(multidrug ABC transporter, ATP-binding protein) BCE33L1609(multidrug ABC transporter, permease) BCE33L1610(multidrug ABC transporter, permease)	BT9727_0376(terD) BT9727_0377(terD) BT9727_0378(terD) BT9727_0379(terC-like) BT9727_0380 BT9727_0381(tellurite resistance protein) BT9727_4245(ABC transporter, ATP-binding protein; sodium export ATP-binding protein) no permease BT9727_1639(multidrug ABC transporter, ATP-binding protein) BT9727_1640(multidrug ABC transporter, permease) BT9727_1641(multidrug ABC transporter, permease)	BC0442(terD) BC0443(terD) BC0444(terD) BC0445(terC-like) BC0446 BC0447(tellurite resistance protein) BC4510 (Sodium export ATP-binding protein) BC4509 (Sodium export permease protein) BC2410(TetR family) BC2411(macrolide efflux protein) BC1734(export ABC transporter, ATP-binding) BC1735(permease) BC1736(permease) BC5093 (xanthine permease) BC4140(Mg(2+) transport ATPase, P-type) BC4141(Mg(2+) transport ATPase protein C)
DNA metabolism	BCE33L3496(exodeoxyribonuclease III) BCE33L3497(AdaA) BCE33L3498(methylated-DNA--protein-cysteine S-methyltransferase) BCE33L3499(DNA-3-methyladenine glycosidase) BCE33L3897(DNA polymerase IV)	BT9727_3484(exodeoxyribonuclease III) BT9727_3485(AdaA) BT9727_3486(methylated-DNA--protein-cysteine S-methyltransferase) BT9727_3487(DNA-3-methyladenine glycosidase) BT9727_3889(DNA polymerase IV)	BC3739(Exodeoxyribonuclease III) BC3740(Ada regulatory protein) BC3741(O6-methylguanine-DNA methyltransferase) BC3742(DNA-3-methyladenine glycosylase II) BC4142(DNA polymerase IV)

Supplementary Table 16. Composition of spore coat and germination proteins in *B. thuringiensis* 97-27 and *B. cereus* E33L, compared to *B. cereus* and *B. anthracis*.

Organism	*B. thuringiensis* 97-27	*B. cereus* E33L	*B. cereus* ATCC 14579	*B. anthracis* Ames	*B. anthracis* Sterne	*B. anthracis* A2012
# spore coat proteins	11	14	15	10	12	11
cotX	0	2	2	0	0	0

		BCE33L2603 BCE33L2605	BC2872 BC2874			
cotW	0	1 BCE33L2604	1 BC2873	0	0	0
cotY	1 BT9727_1126	1 BCE33L1119	1 BC1222	1 BA1238	1 BAS1145	1 BA_1773
cotZ	1 BT9727_1121	1 BCE33L1115	1 BC1218	1 BA1234	1 BAS1141	1 BA_1769
cotA	0	0	0	0	0	0
cotB	2 BT9727_0324 BT9727_0325	2 BCE33L0327 BCE33L0328	2 BC0389 BC0390	0	2 BAS0340 annotated as hypothetical BAS0341	2 BA_0927 annotated as hypothetical BA_0928
cotC	0	0	0	0	0	0
cotD	1 BT9727_1437	1 BCE33L1438	1 BC1560	1 BA1581	1 BAS1465	1 BA_2098
cotE	1 BT9727_3511	1 BCE33L3529	1 BC3770	1 BA3903	1 BAS3619	1 BA_4376
cotF	1 BT9727_2873	1 BCE33L2830	0	1 BA3121	1 BAS2903	1 BA_3623
cotG	1 BT9727_1861 annotated as exosporium protein B 83% identity with BC	1 BCE33L1851 annotated as exosporium protein B 96% identity with BC	1 BC2030	1 BA2045 annotated as hypothetical protein 89% identity with BC	1 BAS1898 annotated as hypothetical protein 89% identity with BC	1 BA_2545
cotH	1 BT9727_1862	1 BCE33L1852	1 BC2031	1 BA2046	1 BAS1899	1 BA_2546
cotL	0	0	1 BC3620	0	0	0
cotM	1 BT9727_3373	1 BCE33L3323	1 BC3621	1 BA3681	1 BAS3412	1 BA_4166
cotS	1 BT9727_4661	1 BCE33L4680	1 BC4954	1 BA5188	1 BAS4823	1 BA_0060
cotSA	0	0	0	0	0	0
cotK	0	0	1 BC3617	1 BA3678	1 BAS3409	0
cotT	0	0	0	0	0	0
spoIVA	1 BT9727_1391	1 BCE33L1391	1 BC1509	1 BA1530	1 BAS1419	1 BA_2050

spoVID	1 BT9727_4193	1 BCE33L4204	1 BC4467	1 BA4692	1 BAS4357	1 BA_5131
gerBC	1 BT9727_2902	1 BCE33L2851	1 BC3110	0	BAS2926 N-terminal BAS2925 C-terminal spore germination protein BC	BA_3651 N-terminal BA_3650 C-terminal
gerBB	1 BT9727_2903	1 BCE33L2852	1 BC3111	0	BAS2927	BA_3653 N-terminal BA_3652 C-terminal missing central part
gerBA	0	0	0	0	0	0
gerKA	1 BT9727_0546	1 BCE33L0546	1 BC0635	1 BA0635	1 BAS_0602	1 BA_1218
gerKB	1 BT9727_0545	1 BCE33L0545	1 BC0634	1 BA0634	1 BAS0601	1 BA_1217
gerKC	1 BT9727_0544	1 BCE33L0544	1 BC0633	1 BA0633	1 BAS0600	1 BA_1216
gerAA	1 BT9727_2904	1 BCE33L2853	0	1 BA3150	1 BAS2928	1 BA_3654
gerAB	0	0	0	0	0	0
gerAC	0	0	0	0	0	0
gerCA	1 BT9727_1394	1 BCE33L1394	1 BC1512	1 BA1533	1 BAS1422	1 BA_2053
gerCB	1 BT9727_1395	1 BCE33L1395	1 BC1513	1 BA1534	1 BAS1423	1 BA_2054
gerCC	1 BT9727_1396	1 BCE33L1396	1 BC1514	1 BA1535	1 BAS1424	1 BA_2055
gerD	1 BT9727_0143	1 BCE33L0141	1 BC0169	1 BA0148	1 BAS0148	1 BA_0731
gerF	1 BT9727_4840	1 BCE33L4855	2 BC5163 BC1723	1 BA5391	1 BAS5011	1 BA_0249
gerE	1 BT9727_4226	1 BCE33L4238	1 BC4501	1 BA4724	1 BAS4385	1 BA_5157
gerIA	1 BT9727_4464 81% identity with BC	1 BCE33L4483 85% identity with BC	1 BC4731	gerHA	gerHA	BA_5404
gerIB	1	1	1	gerHB	gerHB	BA_5405

	BT9727_4465 92% identity with BC	BCE33L4484 92% identity with BC	BC4732			
gerIC	1 BT9727_4466 92% identity with BC	1 BCE33L4485 93% identity with BC	1 BC4733	gerHC	gerHC	BA_5406
gerM	1 BT9727_4219	1 BCE33L4235	1 BC4495	1 BA4716	1 BAS4378	1 BA_5150
gerHA	gerIA BT9727_4464 93% identity with BA ames	gerIA BCE33L4483 93% identity with BA ames	gerIA BC4731 84% identity with BA ames	1 BA4984	1 BAS4630	1 BA_5404 100% amino acid identity with BA ames gerHA
gerHB	gerIB BT9727_4465 95% identity with BA ames	gerIB BCE33L4484 95% identity with BA ames	gerIB BC4732 94% identity with BA ames	1 BA4985	1 BAS4631	1 BA_5405
gerHC	gerIC BT9727_4466 93% identity with BA ames	gerIC BCE33L4485 93% identity with BA ames	gerIC BC4733 96% identity with BA ames	1 BA4986	1 BAS4632	1 BA_5406
gerQA	1 BT9727_2888 96% identity with BC	0	1 BC3099	0	0	0
gerQB	1 BT9727_2887 99% identity with BC	0	1 BC3098	0	0	0
gerQC	1 BT9727_2886 99% identity with BC	0	1 BC3097	0	0	0

Supplementary Table 17: Pseudogenes in *B. thuringiensis* 97-27 and *B. cereus* E33L

9727 gene_id	E33L gene_id	Definition
Genes that were pseudogenes in both 9727 and E33L		
BT9727_0248	BCE33L0251	probable UDP-glucose 4-epimerase, N-terminal region
BT9727_0249	BCE33L0252	probable UDP-glucose 4-epimerase, C-terminal region
BT9727_0622	BCE33L0622	ferrous iron transport protein B, C-terminal region
BT9727_0623	BCE33L0623	ferrous iron transport protein B, N-terminal region
BT9727_1240*	BCE33L1242*	ribonucleoside-diphosphate reductase alpha chain, N-terminal region
BT9727_1241*	BCE33L1243*	ribonucleoside-diphosphate reductase alpha chain, C-terminal region
BT9727_1253	BCE33L1255	possible enhancer of serine sensitivity, N-terminal region
BT9727_3520*	BCE33L3538*	recA protein (recombinase A), C-terminal region
BT9727_3521*	BCE33L3539*	recA protein (recombinase A), N-terminal region
BT9727_3806	BCE33L3821	stage V sporulation protein AE, C-terminal region
BT9727_3807	BCE33L3823	stage V sporulation protein AE, N-terminal region
Genes that were pseudogenes in 9727 but intact in E33L		
BT9727_0245		probable alpha/beta hydrolase, N-terminal fragment
BT9727_0816		possible syd protein, N-terminal region
BT9727_1005		S-layer protein, N-terminal
BT9727_1006		S-layer protein, C-terminal
BT9727_1606		possible glycosyl hydrolase, N-terminal fragment
BT9727_1724		possible amino acid permease, N-terminal
BT9727_1725		possible amino acid permease, Central region
BT9727_1726		possible amino acid permease, C-terminal
BT9727_2098		S-layer protein, missing N-terminal
BT9727_2947		peptidoglycan N-acetylglucosamine deacetylase, C-terminal fragment
BT9727_2951		phosphoglycerate mutase family protein, C-terminal region
BT9727_2974		possible aminoacylase (N-acyl-L-amino-acid amidohydrolase), N-terminal
BT9727_2973		possible aminoacylase (N-acyl-L-amino acid amidohydrolase), C-terminal
BT9727_2993**		putative phosphohydrolases, C-terminal region
BT9727_2996**		putative phosphohydrolases, N-terminal region
BT9727_3058		conserved hypothetical protein, N-terminal region
BT9727_3174		2,5-diketo-D-gluconic acid reductase, C-terminal region
BT9727_3182		conserved hypothetical protein, C-terminal region
BT9727_4574**		gluconate permease, C-terminal fragment, GntP family
pBT9727_0057		phage-related DNA methylase, C-terminal region
pBT9727_0058		phage-related DNA methylase, N-terminal region
Genes that were pseudogenes in BCE33L but intact in 9727		
	BCE33L0873	possible dihydroxyacetone kinase, C-terminal region
	BCE33L2322	conserved hypothetical protein, N-terminal region
	BCE33L2323	conserved hypothetical protein, C-terminal region
	BCE33L2324	conserved hypothetical protein, C-terminal region
	BCE33L2325	conserved hypothetical protein, N-terminal region

	BCE33L3048	putative sugar (and other) transporter, C-terminal region
	BCE33L3049	putative sugar (and other) transporter, N-terminal region
	BCE33L3380	N-acetylmuramoyl-L-alanine amidase, family 2, C-terminal region
	BCE33L3381	N-acetylmuramoyl-L-alanine amidase, family 2, N-terminal region
	BCE33L3427***	terminase large subunit, C-terminal region
	BCE33L3429***	terminase large subunit, N-terminal region
	BCE33L4922**	L-lactate permease, C-terminal fragment
_	pE33L466_0002	possible DNA-invertase, C-terminal region
_	pE33L466_0012	possible exonuclease; possible DNA polymerase III fragment
_	pE33L466_0043	sodium/proline symporter family protein, C-terminal region
_	pE33L466_0044	sodium/proline symporter family protein, N-terminal region
_	pE33L466_0100	neutral protease, N-terminal region
_	pE33L466_0236	possible ribosomal protein S1 fragment
_	pE33L466_0261	phosphoglycolate phosphatase, C-terminal region
_	pE33L466_0263	dihydropyrimidine dehydrogenase (NADPH+), N-terminal region
_	pE33L466_0264	dihydropyrimidine dehydrogenase (NADPH+), central domain
_	pE33L466_0265	dihydropyrimidine dehydrogenase (NADPH+), C-terminal region
_	pE33L466_0266	dihydroorotate dehydrogenase, C-terminal region
_	pE33L466_0267	dihydroorotate dehydrogenase, N-terminal region
_	pE33L466_0273	possible methyl-accepting chemotaxis protein, C-terminal region
_	pE33L466_0313	DNA polymerase III, gamma and tau subunits, N-terminal region
_	pE33L466_0319	DNA topoisomerase III, N-terminal region
_	pE33L466_0320	DNA topoisomerase III, central region
_	pE33L466_0321	DNA topoisomerase III, C-terminal region
_	pE33L466_0326	abortive lactococcal phage infection protein, N-terminal region
_	pE33L466_0381	possible enhancin fragment
_	pE33L466_0382	possible enhancin fragment, central domain
_	pE33L466_0406	stage V sporulation protein AE, C-terminal region
_	pE33L54_0028	ImpB/MucB/SamB family protein, N-terminal region
_	pE33L54_0046	impB/mucB/samB family protein, C-terminal region

* Genes are interrupted by type I intron, but still maintain its function.

** Genes are interrupted by IS element

***Genes are interrupted by GIY-YIG intron, but still probably maintain its function.

Table S18 General features of *B. thuringiensis* 97-27 and *B. cereus* BCE33L genomes

Feature	BT9727	pBT9727	BCE33L	pBCE33L466	pBCE33L54	pBCE33L9	pBCE33L8	pBCE33L5
Size (bp)	5237680	77112	5300915	466370	53501	9150	8191	5108
Coding sequence (bp)	4379823	62574	4439748	328878	40853	5700	4602	3345
Coding sequence (%)	84	81	84	71	76	62	56	65
G + C	35	33	35	33	32	31	32	31

content (%)								
Average ORF size (bp)	856	782	865	706	704	518	575	669
CDSs, total	5118	80	5134	466	58	11	8	5
CDSs with assigned function	3828	16	3880	228	27	6	4	2
Conserved hypothetical CDSs	1198	20	1136	168	9	1	2	2
hypothetical CDSs	92	44	118	70	22	4	2	1
B. cereus group-specific CDSs	496	21	513	33	2	0	1	0

Supplementary Table S19. Phage genes in *B. thuringiensis* 97-27 and *B. cereus* E33L chromosome and plasmids

BT97-27 gene_id	BCE33L gene_id	Definition
BT9727_0848	BCE33L0839	phage replication protein
BT9727_1343	BCE33L1342	site-specific recombinase, phage integrase family
BT9727_1979	BCE33L1958	Bacillus cereus group-specific phage-like fragment
BT9727_2095	-----	phage integrase
BT9727_2401	BCE33L2364	probable phage terminase, small subunit
BT9727_3455	BCE33L3468	possible phage integrase family protein
BT9727_4476	BCE33L4494	site-specific recombinase, phage integrase family
	BCE33L4224	phage-related protein
	BCE33L3407 - BCE33L3461	lambdoid prophage
	BCE33L2879	conserved hypothetical protein; possible phage protein
	pE33L54_0017	integration/recombination/invertion protein, phage-related
	pE33L54_0043	possible phage integrase family protein
	pE33L54_0054	prophage LambdaBa02, repressor protein
	pE33L466_0008	Bacillus cereus group-specific;

		possible phage-related protein
	pE33L466_0103	positive control factor, phage-related
	pE33L466_0143	possible phage-related protein
	pE33L466_0162 - pE33L466_0165	phage-related proteins
	pE33L466_0177	phage protein
	pE33L466_0182	Bacteriophage phi-105 ORF16

Table S20. Inventory of *B. anthracis* Ames IS elements in *B. cereus* (ATCC 14579), *B. thuringiensis* 97-27, and *B. cereus* E33L

Family	*B. cereus* (ATCC 14579)	*B. cereus* plasmid	*B. anthracis*	*B. anthracis* plasmids	97-27	97-27 plasmid	E33L	E33L plasmids
IS3	10(3)			3(7)	15(2)		(3)	8(5)
IS4	1(1)		(3)	2(8)	(1)	1	1(2)	9(20)
IS6	3							
IS605	10		3	(1)	11(1)		4(1)	10(10)
IS110			(2)		1		2	1
IS5				1(3)	26(1)			
Tn3				1(1)	(2)		(1)	2(7)
Novel	1						1	

According to the Mahillon and Chandler classification (J.Mahillon, M.Chandler Microbiol. Mol. Biol. Rev. 62, 725(1998))
Numbers in parenthesis indicate truncated or frameshifted transposase
Numbers not in parenthesis indicate intact transposase

Table S21. Common Flagellar genes in found in Bt9727 and BcE33L, compared to *B. cereus* 14579 and *B. anthracis*

gene	Bt9727	BcE33L	Bc14579	Ba A2012	Ba Sterne
cheA	BT9727_1518	BTZK1507	BC1628	BA_2174 N-terminal BA_2175 central BA_2176 C-terminal	BAS1542 N-terminal BAS1543 C-terminal
flgL	BT9727_1527	BTZK1516	|BC1637	BA_2185 N-terminal BA_2186 C-terminal	BAS1552 N-terminal BAS1553 C-terminal
fliF	BT9727_1534	BTZK1523	BC1644	BA_2193	BAS1560

M-ring protein				BA_2194	
cheV	BT9727_1544	BTZK1533	BC1654	BA_2206 N-terminal BA_2207 C-terminal	BAS1571 N-terminal BAS1572 C-terminal
fliC flagellin	BT9727_1546	BTZK1535	BC1656 BC1657 BC1658 BC1659	BA_2218	BAS1582 BAS1552 N-terminal BAS1553 C-terminal
fliN	BT9727_1552 BT9727_1551 BT9727_1549	BTZK1541 BTZK1540 BTZK1538	BC1664 BC1663 BC1661	BA_2220 fragment	BAS1584 fragment
fliM	BT9727_1550	BTZK1539	BT9727_1550	BA_2221 N-terminal BA_2222 C-terminal	BAS1585 N-terminal BAS1586 C-terminal

Figure S1A.

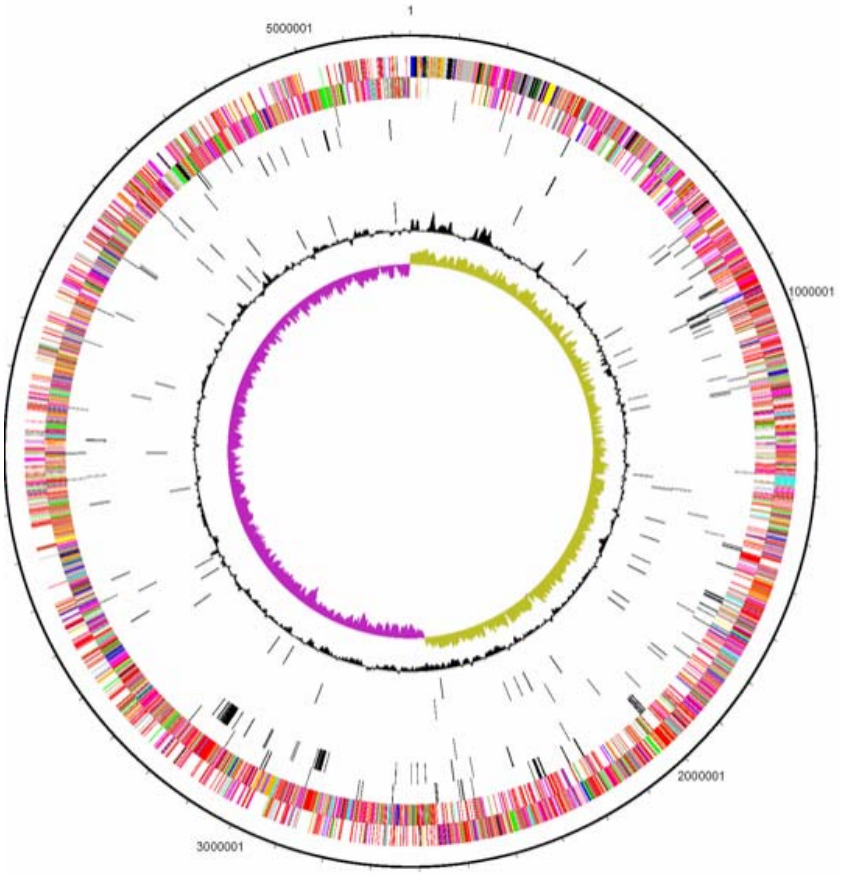


Figure S1B.

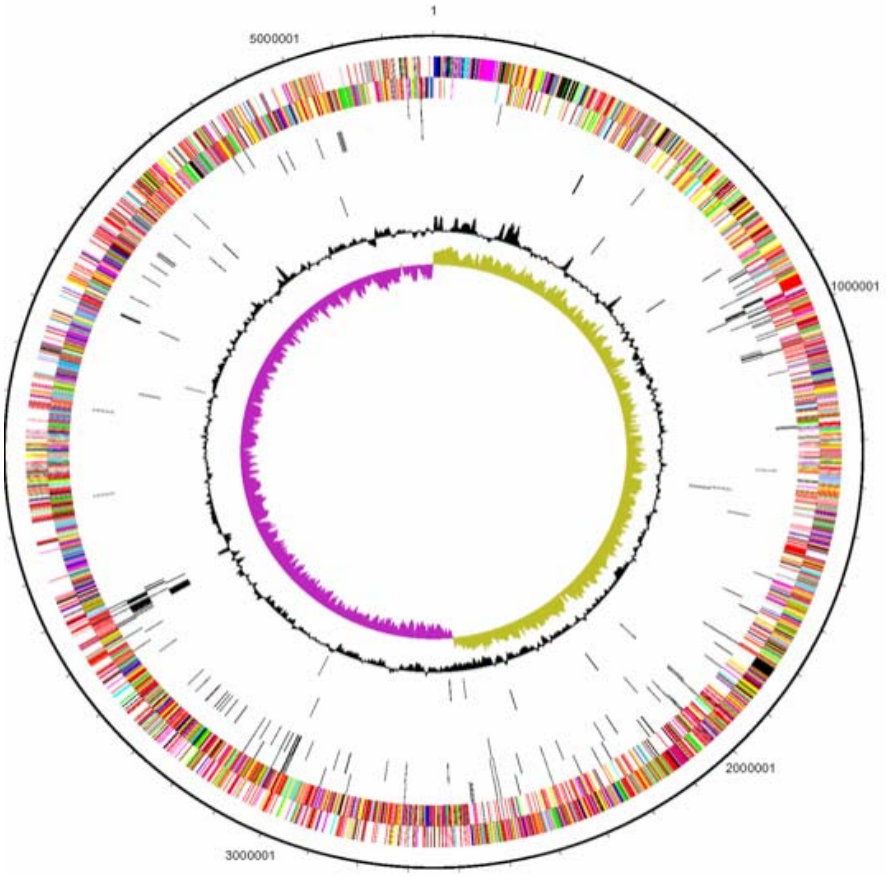


Figure S2.

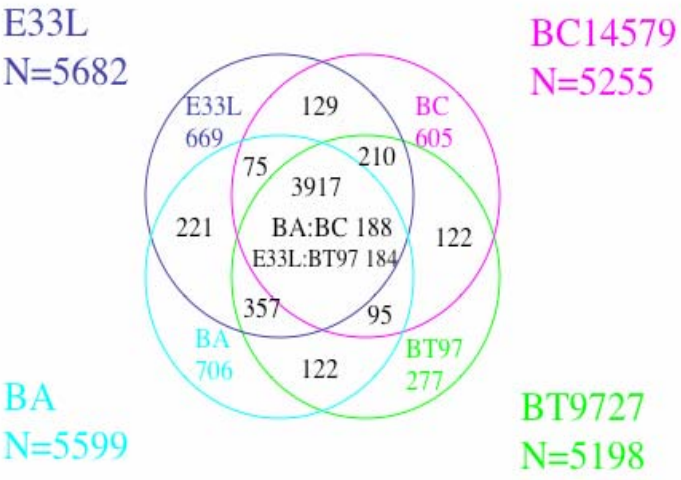

E33L
N=5682
BC14579
N=5255
E33L
669
129
BC
605
75
210
3917
221
BA:BC 188
E33L:BT97 184
122
357
95
BA
706
122
BT97
277
BA
N=5599
BT9727
N=5198

Figure S3.

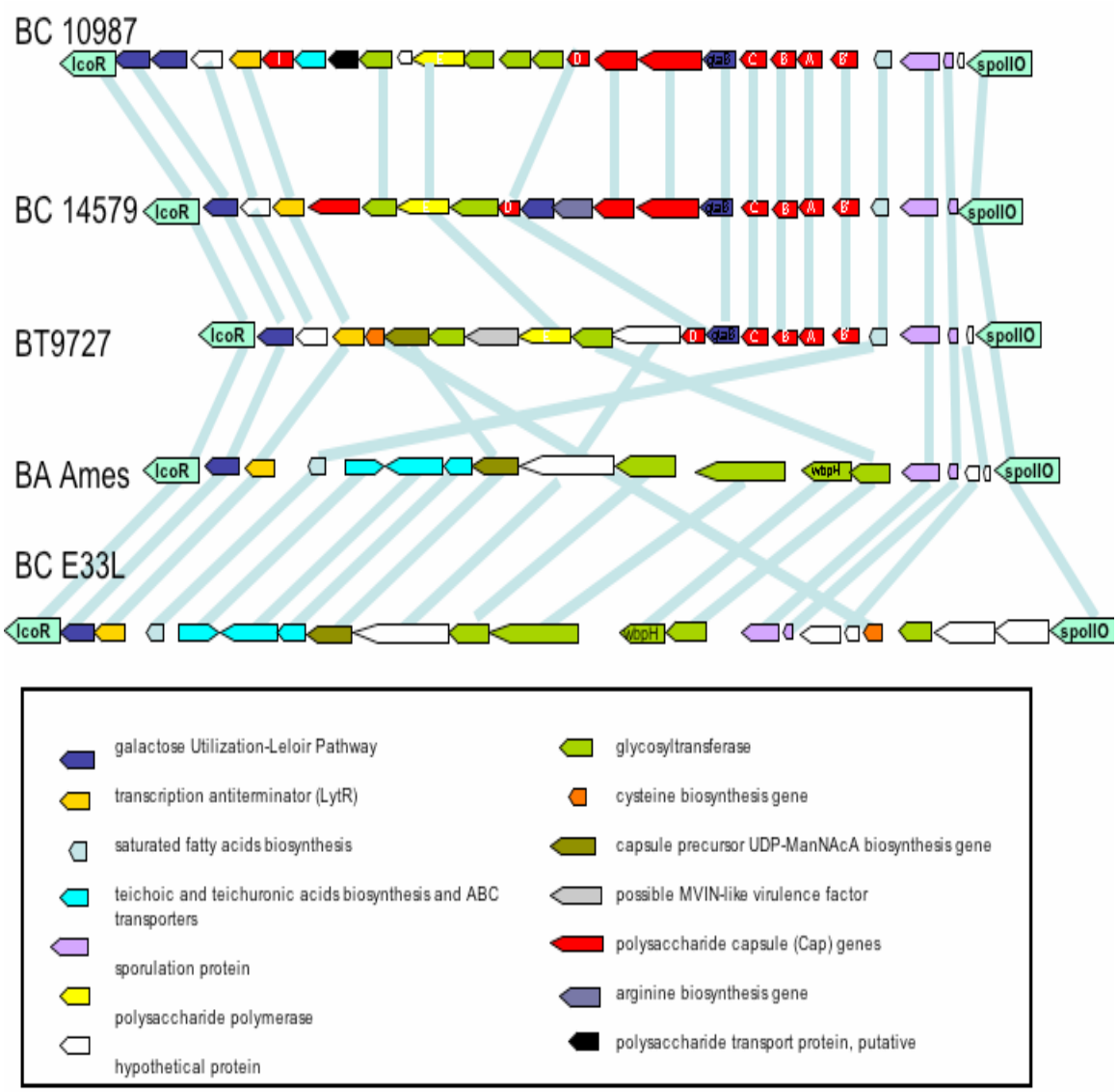

BC 10987
BC 14579
BT9727
BA Ames
BC E33L
galactose Utilization-Leloir Pathway
transcription antiterminator (LytR)
saturated fatty acids biosynthesis
teichoic and teichuronic acids biosynthesis and ABC transporters
sporulation protein
polysaccharide polymerase
hypothetical protein
glycosyltransferase
cysteine biosynthesis gene
capsule precursor UDP-ManNAcA biosynthesis gene
possible MVIN-like virulence factor
polysaccharide capsule (Cap) genes
arginine biosynthesis gene
polysaccharide transport protein, putative

Figure S4.

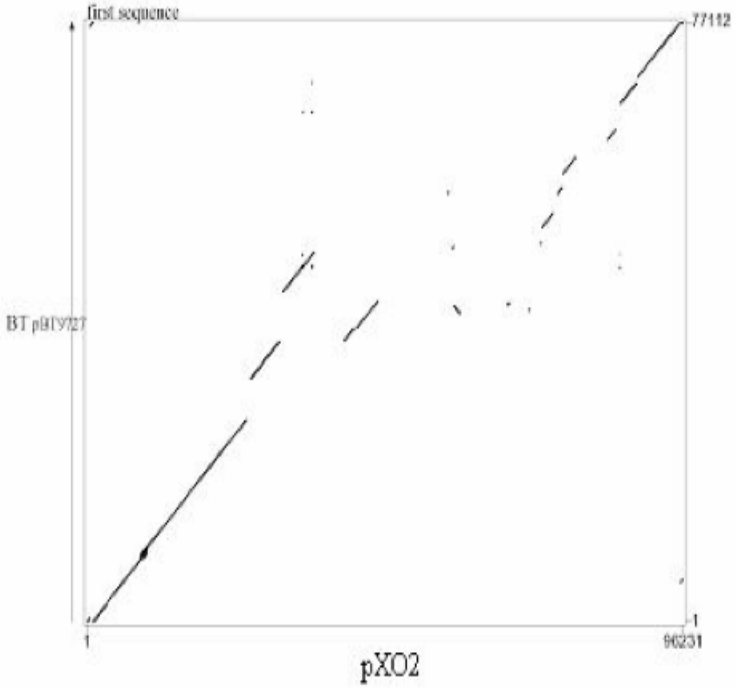
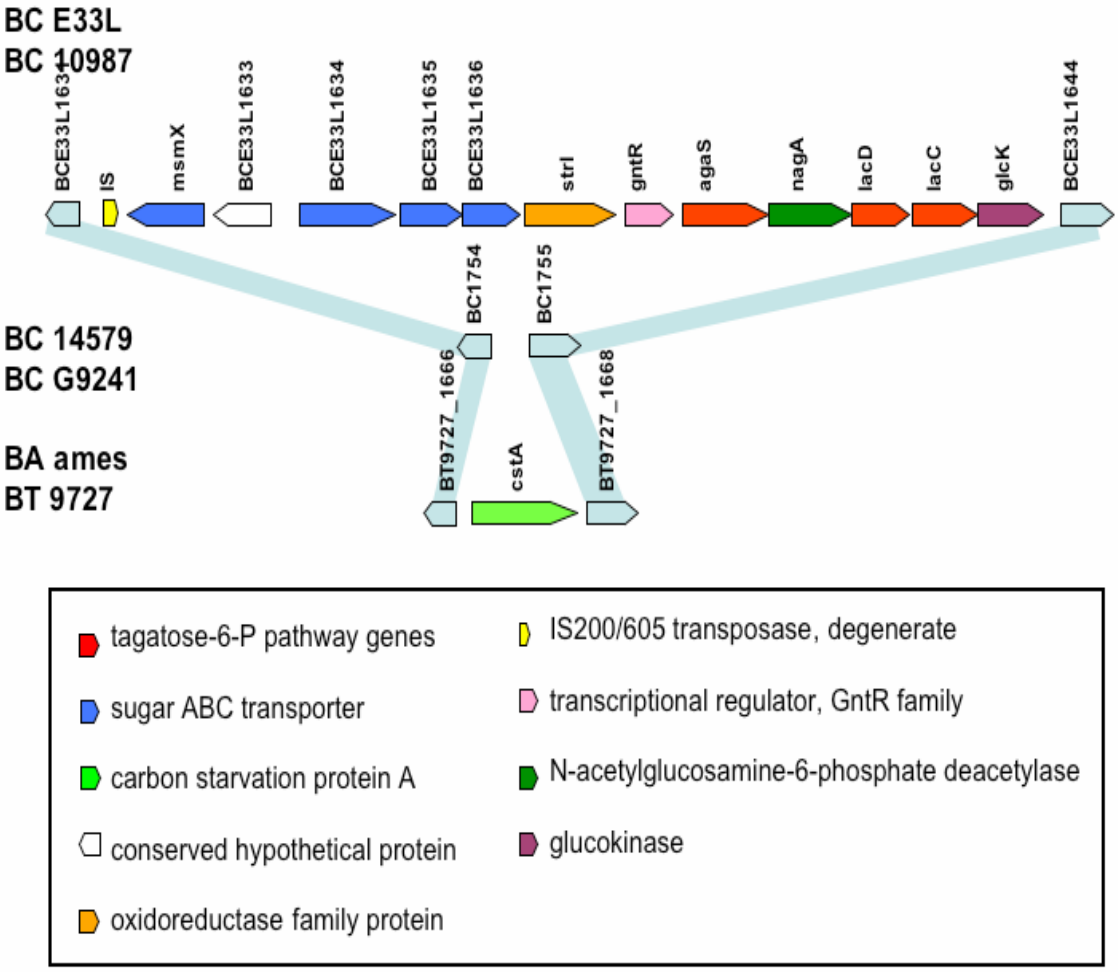


Figure S5.



LANL: LA-UR-05-6058
LBNL - 58996