CMI
CLINICAL MICROBIOLOGY
AND INFECTION

ESCMID

New genetic biomarkers to differentiate pathogenic and clinically relevant Bacillus cereus strains

Journal:	*Clinical Microbiology and Infection*
Manuscript ID	CLM-21-20559.R3
Article Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Kavanaugh, Devon; INRAE, Micalis Glasset, Benjamin; INRAE, Micalis Dervyn, Rozenn; INRAE, Micalis Guérin, Cyprien; INRAE, Maiage Plancade, Sandra; INRAE, Maiage Herbin, Sabine; Anses, Food Safety Brisabois, Anne; Université Paris-Est, ANSES, Laboratory for Food Safety, Maisons-Alfort Cedex, France, Food safety Nicolas, Pierre; INRAE, Maiage Rama Rao, Nalini; INRAE, Micalis
Key Words:	Bacillus cereus, pathogenicity, genetic biomarkers, diagnostic tool, AUC
Abstract:	Objectives. Bacillus cereus is responsible for food poisoning in France and rare but severe clinical infections. The pathogenicity of strains varies from harmless to lethal strains. However, there are currently no markers, either alone or in combination, to differentiate pathogenic from non-pathogenic strains. The objective of the study was to identify new genetic biomarkers to differentiate pathogenic from clinically relevant Bacillus cereus strains. Methods. A first set of 15 B. cereus strains were compared by RNAseq. A logistic regression model with lasso penalty was applied to define combination of genes whose expression was associated with strain pathogenicity. The identified markers were checked for their presence/absence in a collection of 95 B. cereus strains with varying pathogenic potential (FBO, clinical and non-pathogenic). ROC-AUC analysis determines the combination of biomarkers, which best differentiate between the “disease” versus ‘non-disease’ groups. Results. 7 genes were identified during the RNAseq analysis with a prediction to differentiate between pathogenic and non pathogenic strains. The validation of the presence/absence of these genes in a larger collection of strains coupled with AUC prediction showed that a combination of 4 biomarkers was sufficient to accurately discern clinical strains from harmless strains, with an AUC of 0.955, sensitivity of 0.9 and specificity of 0.86. Conclusions. These new findings help in the understanding of B. cereus pathogenic potential and complexity and may provide tools for a better assessment of the risks associated with B. cereus contamination to improve patient health and food safety.

New genetic biomarkers to differentiate pathogenic and clinically relevant *Bacillus cereus* strains

Devon W. Kavanaugh[1+], Benjamin Glasset[1+], Rozenn Dervyn[1], Cyprien Guérin[3], Sandra Plancade[3], Sabine Herbin[2], Anne Brisabois[2], Pierre Nicolas[3] and Nalini Ramarao[1*]

[1]Université Paris-Saclay, INRAE, Micalis Institute, 78350, Jouy-en-Josas, France

[2]Université Paris-Est, Anses, Laboratory for Food Safety, 94700 Maisons-Alfort, France

[3]Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

[+] these authors contributed equally to this work

* Correspondence should be addressed to Nalini Rama Rao, INRAE, Micalis Institute, 78350, Jouy-en-Josas, France

Email: nalini.ramarao@inrae.fr

Keywords

Bacillus cereus, *pathogenicity*, *genetic biomarkers*

Abstract

Objectives. Bacillus cereus is responsible for food poisoning in France and rare but severe clinical infections. The pathogenicity of strains varies from harmless to lethal strains. However, there are currently no markers, either alone or in combination, to differentiate pathogenic from non-pathogenic strains. The objective of the study was to identify new genetic biomarkers to differentiate pathogenic from clinically relevant *Bacillus cereus* strains.

Methods. A first set of 15 *B. cereus* strains were compared by RNAseq. A logistic regression model with lasso penalty was applied to define combination of genes whose expression was associated with strain pathogenicity. The identified markers were checked for their presence/absence in a collection of 95 *B. cereus* strains with varying pathogenic potential (FBO, clinical and non-pathogenic). ROC-AUC analysis determines the combination of biomarkers, which best differentiate between the “disease” versus ‘non-disease’ groups.

Results. 7 genes were identified during the RNAseq analysis with a prediction to differentiate between pathogenic and non pathogenic strains. The validation of the presence/absence of these genes in a larger collection of strains coupled with AUC prediction showed that a combination of 4 biomarkers was sufficient to accurately discern clinical strains from harmless strains, with an AUC of 0.955, sensitivity of 0.9 and specificity of 0.86.

Conclusions. These new findings help in the understanding of *B. cereus* pathogenic potential and complexity and may provide tools for a better assessment of the risks associated with *B. cereus* contamination to improve patient health and food safety.

Introduction

Bacillus cereus is the third causative agent of food-borne-outbreaks (FBO) in Europe [1]. *B. cereus* can induce two types of gastrointestinal diseases, leading to generally mild and self-limiting emetic or diarrhoeal syndromes, although several cases of severe infections have been reported [2]. *B. cereus* also induces systemic infections leading to patient death in approximately 10% of cases [3-7]. *B. cereus* is also a source of central nervous system infections and other systemic infections especially in newborns [3, 8]. Recent epidemiological studies show that the number of cases of serious *B. cereus* infections is largely underestimated [9]. The pathogenic potential of *B. cereus* is extremely variable, with some strains being harmless and others lethal.

B. cereus possesses several toxin genes, such as *nhe*, *hbl* and *cytK* [2, 10]. These toxins provide an indication of the strain toxicity potential but are not sufficient, alone, to discriminate hazardous from harmless strains [9, 11-13]. Indeed, several studies have shown that Nhe production by hazardous strains is variable and that non-pathogenic strains can also produce it in large quantities [1, 12]. Moreover, these toxins do not appear to be suitable markers for strains causing non-gastrointestinal infections [9]. *B. cereus* produces other toxins such as haemolysin II (HlyII), the metalloproteases InhA1, InhA2 and the cell wall peptidase FM (CwpFM), which may also be involved in pathogenicity [14-18]. The emetic form of *B. cereus* food poisoning is caused by the peptide cereulide [19], which represent less than 1% of the FBO strains of *B. cereus* [1, 19, 20].

To date, the above described determinants were not sufficient to completely explain the virulence of *B. cereus* [21] and there are currently no markers, either alone or in combination, to differentiate pathogenic from non-pathogenic strains. In this work, we took advantage of a well characterized collection of 95 *B. cereus* strains and compared pathogenic (FBO and clinical) with non-pathogenic strains. We identified a combination of four as yet undescribed biomarkers, wherein their presence/absence allows an accurate identification of clinical *B. cereus* strains. These findings constitute a huge step in the understanding of the *B. cereus* pathogenic potential and complexity and may provide tools to better assess the risks associated with *B. cereus* contamination.

Methods

Isolate information

This study includes 39 *B. cereus* strains associated with foodborne illness [1], 35 strains isolated from human patients following systemic or local infections [9] and 21 non-pathogenic strains [11, 22] (Sup Table 1). We have previously shown a correlation between cytotoxicity and virulence [21]. Nevertheless, although these strains had previously been shown to be weakly cytotoxic to human cells and to have reduced virulence in an insect infection model, this does not rule out their potential ability to produce symptoms in specific vulnerable populations.

RNA extraction

The transcriptome study by RNAseq was carried out on 15 strains representative of the three collections (Sup Table 2) in triplicates. Bacterial cultures were incubated in BHI medium at 30°C in microaerophilic condition (5% O_2–15% CO_2–80% N_2) at pH 7 until entry into stationary growth phase. Samples were centrifuged at 12,000 g for 3 min at 4°C and placed immediately at -80°C until processing. The bacterial pellets were re-suspended with 200 µl of 10 mM Tris-HCl at pH 8 + 4 µl of lysozyme at 50 mg/ml and incubated at 37°C. Total RNA was extracted with the HPRNA kit (High Pure RNA Isolation Kit; Roche) as previously described [23]. The RNA integrity was measured by the RIN (RNA Integrity Number) and were between 7 and 10. The mRNA were enriched with the RiboZero Kit (Illumina). The sequencing of the mRNA was carried out by the I2BC platform (CNRS, Gif-sur-Yvette). Directional and paired libraries were prepared with the Illumina scriptseq kit and the sequencing was performed on an Illumina Nextseq machine.

Transcriptome sequencing analysis

Sequencing quality was assessed using FastQC, and adapter sequences and low-quality base pairs were removed using cutadapt (version 1.9) [24]. Reads were further trimmed in 3' using sickle (version 1.33, option "-x" and default values for all other parameters, implying a Phred quality cutoff of 20). In absence of whole genome sequences for the 15 strains, the cleaned reads were mapped against a repertoire of allelic variants for 23,815 genes aiming at accounting for the pangenome of *B. cereus* group. This repertoire was obtained by single-linkage clustering based on the results of an all-against-all blastn comparison (version 2.2.26, e-value cut-off 1e-5) [25] of 519,931 CDSs extracted from the 91 annotated complete genomes available at the time of analysis for *B. cereus* group in Genbank. Pairs of CDSs that aligned over at least 70% of the length of the shortest sequence and with at least 75% nucleotide sequence identity were grouped in the same cluster, which resulted in 23,815 clusters representing distinct genes. Reads were mapped using bowtie2 (version 2.2.6, options "-N 1 -L 16 -R 4") [26] whose results were converted to bam format using SAMtools version 1.9 [27]. Read counts on each allelic variant were obtained using HTSeq-count (version 0.6.1) [28] and summed over allelic variants to obtain a single read count per gene per sample. To cope with sequence similarity between allelic variants of a same gene and fragmentation of the reference according to gene boundaries, R1 and R2 reads were aligned independently and use of HTSeq-count option "-a 0" allowed to count reads that aligned equally well on several allelic variants of a same gene. Of note, since bowtie2 mapped each read on a single allelic variant, reads could not be counted more than once in the sum. Expression levels expressed as $\log_2$ scaled rpkm (reads per kilobase per million mapped reads) were produced by the R package "edgeR" (version 3.11) using the mean length of the genes in the cluster and a prior count of 1.

Raw transcriptomic data and differential expression analysis are accessible through GEO Series accession number GSE168681 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE171128).

Statistical model

The strategy for statistical analysis of RNAseq data was to select genes to predict whether a strain is pathogenic *y=1* or not *y=0* and evaluate the prediction accuracy. We considered the logistic regression model with lasso penalty implemented in the R-package "glmnet", which allows the selection of a limited subset of genes whose expression is associated with strain pathogenicity [29]. The package glmnet provides an interval cross validation procedure to select the penalty constant, which determines the number of selected genes.

The prediction accuracy of the procedure was evaluated in a cross-validation framework where splitting in training and validation sets preserves the matching of the three replicates of each strain. For each replicate, the model provides a probability $\hat{z}_i$ to be pathogenic, and we considered the average value over the three replicates as the prediction probability of the strain. The predicted pathogenicity status is set to zero if the prediction probability is smaller than 0.5 and 1 otherwise.

Biomarker screen by PCR

The 7 marker genes were retrieved from at least 20 sequenced *B. cereus* strains from NCBI databases and aligned by CLC Main workbench7 software to identify two regions conserved across the strains. Within these regions, 20 bp primers were designed using the Beacon Designer software. For the majority of the selected genes there were no perfectly conserved sequence and some bases had to be replaced with R (A/T), Y (C/T) or W (A/T) for primer design (Sup Table 3).

For all the strains of the collection, a single colony was picked, resuspended in 100 µL Tris-EDTA NaCl buffer (TEN) and incubated at 98°C for 10 min. After centrifugation, 1 µl of supernatant was used as DNA matrix. The PCR mixture contained 1 µl DNA matrix, 0.5 µM primer (forward and reverse), 10 µL DreamTaq Green PCR Master Mix (2X) (Thermo Scientific) in a final volume of 20 µL. PCR fragment sizes were revealed on 1.5% agarose gels containing Midori Green, and visualised by a UV imaging device.

AUC analysis to select combinations of biomarkers

The PCR data were pooled into a presence (1) /absence (0) table, which was then used as input for ROC-AUC analysis facilitated by the web-based suite of tools hosted at www.combiroc.eu. The ROC-AUC analysis determines the combination of biomarkers, which will best differentiate the classes of samples input ('disease' versus 'non-disease' groups). Sets of biomarkers were selected based on their performance in sensitivity or specificity alone, or in combination as the AUC metric. Potential hits were filtered at 85% specificity and 85% sensitivity.

Results

RNAseq analysis

We obtained between 9-15 million reads per samples with 90% correctly paired. The overall alignment rate was over 85%. The analysis enabled the creation of a read counts table based on gene expression levels for each sample (Figure 1). The dispersion of the sample count values was homogeneous and the biological triplicates clustered well together. We identified 3276 genes in the core transcriptome, which represents approximately 65% of the genes in each strain.

Identification of 7 biomarkers by logistic regression analysis

A Mann-Whitney-Wilcoxon nonparametric rank test with a classical 5% of qvalue did not allow the prediction of significant differences in gene expression among the strain collections (not shown). Thus, to identify markers that could potentially differentiate pathogenic from non-pathogenic strains, we performed a penalized conditional logistic regression with the lasso method on the entire counting table to select relevant genes for the prediction of pathogenic potential. By applying the prediction model to the 11,179 genes with the selected penalty constant of 0.01, only 7 genes were selected (Table 1).

With the RPKM values of these 7 genes (Sup Table 4), a prediction in a cross-validation framework among the 15 strains, leads to 13 well classified strains (estimated probability $\hat{z}_i$ value below 0.5 for non-pathogenic and above 0.5 for pathogenic strains) and two misclassified strains, one false positive (NP strain PF predicted as pathogenic) and one false negative (pathogenic FBO strain 12CEB01BAC predicted as NP) (Table 2).

Validation of the biomarkers on a large strain collection

Initially, for the first 15 strains, the presence of the 7 selected genes was further assed by PCR (Table 3). These data revealed that when a gene showed no expression by transcriptomic analysis, the gene was actually absent from the strain. Thus, the identification of these 7 biomarkers was based on gene presence/absence, rather than mRNA expression. As such, an approach centred on gene detection was chosen for the screening of the large bacterial collection with the 7 genes selected (Table 3) and to determine the area under the curve (AUC), specificity, and sensitivity of possible combinations of the selected biomarkers.

1-FBO vs NP

For the FBO strains, the best combination of biomarkers able to differentiate NP from FBO strains was obtained with 4 biomarkers (Figure 2A). With this combination, the best AUC was 0.768, the sensitivity 0.69 and the specificity 0.773. Therefore, we obtained some false positive (NP strains that appear pathogenic), and some false negative (FBO strains that appear NP). Taken together, the general trend for the FBO identification was an overall low AUC among the tested combinations, thus preventing their accurate differentiation.

Nevertheless, we identified that several FBO strains were lacking almost all biomarkers. These FBO strains primarily belong to the phylogeny group IV (table 3). We thus performed an additional

AUC analysis after the removal of all strains of the phylogeny group IV of the collection (FBO and NP). The results were significantly improved and the best combination resulted in an AUC above 0.9 and with significantly improved sensitivity or improved specificity. But a combination resulting in sensitivity and specificity above 0.9 was not determined (Figure 2B).

2-NP vs clinical strains

Regarding the clinical strains, the best results were achieved with a combination of 4 biomarkers with an AUC of 0.955, sensitivity of 0.9 and specificity of 0.86. Therefore, the analysis concludes that an accurate differentiation between clinical and non-pathogenic strains can be obtained by using these biomarkers (Figure 2C). These two combinations allowed the accurate discrimination between the two strain populations. Some markers have the same occurrence within the strain collection (5, 6, 7) and were therefore interchangeable during the AUC analysis. Thus, the best combinations of biomarkers are: 1, 2, 3, 5 (or 6 or 7). The genes are named, adhB, agrC, thiJ, BCQ_PI180 (or gshAB or BCQ_PI181).

As a conclusion, a suitable combination of 4 biomarkers has been found to create a robust and accurate test to differentiate clinical from non-pathogenic strains, with an AUC of 0.955, given that test results above 0.9 are considered excellent.

Discussion

The emergence of *B. cereus* as a foodborne pathogen and as an opportunistic pathogen has intensified the need to distinguish strains of public health concern. The pathogenic potential of *B. cereus* is extremely variable, with some strains being harmless and others lethal. Currently, due to the lack of validated and standardized analytical methods, only the presence of *B. cereus* is usually investigated in foods or clinical samples at a species-level. Over the years, new methods have been developed with the leading principle to detect and distinguish *B. cereus* from others *Bacillus* group members by a time-saving and *in-situ* analysis [30], genotyping using high-resolution melting analysis [31], the use of multi-locus sequence (MLST) [32] or the classification of the strains according to their affiliation to a phylogenetic group that offers a first useful indicator of risk [11]. Nevertheless, MLST analysis of the 53 strain sequences included in this study revealed that 21% belonged to the sequence type ST26, and approximately 11% to an undetermined ST (not shown), while >40% of the strains were identified as belonging to PanC clade III (Table 3). As such, the ST types and PanC classifications were unable to completely explain the grouping of the strains.

Here, we report new markers characteristic of pathogenic *B. cereus* strains, which detection requires only PCR, and is thus independently of growth conditions. We could indeed show that the simple presence/absence of the gene was as discriminant as its expression value by transcriptomic analysis. We further calculated the AUC, specificity and sensitivity obtained using the combination of these 4 biomarkers to discriminate between our large *B. cereus* collection inducing various pathologies. CombiROC results demonstrate that clinical strains were more efficiently separated from the non-pathogenic strains than the FBO strains.

Regarding the FBO strains, to improve the analysis, strains belonging to the phylogenetic group IV were removed, thus allowing a significant improvement in strain differentiation. This might prove very useful for food industries to better communicate the risks of *B. cereus* food contamination and to take the appropriate measures for decontamination while preventing or minimizing economic loss. Nevertheless, this implies a two step-test with a first *panC* phylogenetic attribution followed by a biomarker test.

By contrast, regarding the clinical strains, the combination of 4 biomarkers allowed the identification of a strong differentiation test with an AUC of 0.955, sensitivity of 0.9, and specificity of 0.86. Thus, a global test with a strong AUC (above 0.9) and increased sensitivity (rare false negative) could be proposed to accurately discriminate between clinical and harmless strains. As such, our new findings may be relevant to gain additional knowledge on the strains found in hospitals and healthcare settings.

Acknowledgments

We want to thank Sandrine Auger and Mylène Sperry for their help in the primer design. We thank Valentin Loux for bioinformatics support.

COI statement

The authors declare no conflict of interest.

Funding

This work was supported by the European EJP Toxdetect project from the European Union’s Horizon 2020 research and innovation program under Grant Agreement No 773830 and by the Comue Paris Saclay Idex Program n°CDE-2018-002323 - IRE 2018-0021.

Author Contributions

DK, BG, RD: performed experiments, analyzed data, manuscript writing; CG, SP, PN: analyzed data; SH, AB: supervision; NR: initial concept, supervision, analyzed data, writing of manuscript, funding sources.

Legends of figures and tables

Figure 1. RNAseq heatmap. Heatmap representation of expression levels ($\log_2$ rpkm) across the pangenomic repertoire of 23,815 genes (rows) and the 45 samples (columns). Dendrograms are built by hierarchical clustering with average-link. The 3,272 genes with signal in all strains are indicated by grey bars. Non-pathogenic strains are indicated in black and pathogenic strains in red.

Figure 2. CombiROC analysis results. The presence/absence matrix resulting from PCR detection of biomarker sequences was analyzed by CombiROC. (A) Foodborne outbreak strains (FBO) versus non-pathogenic; (B) FBO versus non-pathogenic strains, excluding phylogenetic group IV. Links best sensitivity performance, right highest specificity; (C) clinical versus non-pathogenic strains.

Table 1. List of 7 selected biomarkers with gene position (on the reference genome pAH187_270 - NC_011655.1) and putative function.

Table 2. Estimated probability $\hat{z}_i$ for the 15 strains. A logistic regression model with lasso penalty was applied to select the penalty constant, which determines the number of selected genes. Then prediction accuracy of the procedure was evaluated in a cross-validation framework. For each replicate, the model provides a probability $\hat{z}_i$ to be pathogenic, and we considered the average value over the three replicates as the prediction probability of the strain. The predicted non-pathogenicity corresponds to a $\hat{z}_i$ smaller than 0.5 and the predicted pathogenicity corresponds to $\hat{z}_i$ above 0.5.

Table 3. Presence/absence of biomarkers among non-pathogenic (green), FBO (blue) and clinical (beige) strains. The presence of each biomarker gene was assessed by PCR in all strain of the collection. If the gene was present, a score of 1 was attributed (green boxes), if the gene is absent, a score of 0 is attributed (red boxes).

References

1. Glasset B, Herbin S, Guiller L, Cadel-Six S, Vignaud ML, Grout J, et al. Large-scale survey of Bacillus cereus-induced food-borne outbreaks: epidemiologic and genetic characterization EuroSurveillance. 2016;21(48).
2. Fagerlund A, Brillard J, Fürtst R, Guinebretiere MH, Granum PE. Toxin production in a rare and genetically remote cluster of strains of the Bacillus cereus group. BMC Microbiol. 2007;7:43.
3. Bottone EJ. Bacillus cereus, a volatile human pathogen. Clin Microbiol Rev. 2010;23(2):382-98.
4. Ramarao N, Belotti L, Deboscker S, Ennahar-Vuillemin M, de Launay J, Lavigne T, et al. Two unrelated episodes of Bacillus cereus bacteremia in a neonatal intensive care unit. Am J Infect Control. 2014;42(6):694-5.
5. Gaur AH, Patrick CC, McCullers JA, Flynn PM, Pearson TA, Razzouk BI, et al. Bacillus cereus bacteremia and meningitis in immunocompromised children. Clinic Infect dis. 2001;32:1456-62.
6. Lotte R, Herisse AL, Berrouane Y, Lotte L, Casagrande F, Landraud L, et al. Virulence Analysis of Bacillus cereus Isolated after Death of Preterm Neonates, Nice, France, 2013. Emerg Infect Dis. 2017;23(5):845-8.
7. Chan WM, Liu DT, Chan CK, Chong KK, Lam DS. Infective endophthalmitis caused by Bacillus cereus after cataract extraction surgery. Clin Infect Dis. 2003;37(3):e31-4.
8. Cormontagne D, Rigourd V, Vidic J, Rizzotto F, Bille E, Ramarao N. Bacillus cereus Induces Severe Infections in Preterm Neonates: Implication at the Hospital and Human Milk Bank Level. Toxins (Basel). 2021;13(2).
9. Glasset B, Herbin S, Granier S, Cavalié L, Lafeuille E, Guérin C, et al. Bacillus cereus, a serious cause of nosocomial infections: epidemiologic and genetic survey. PLoS ONE. 2018;13(5):e0194346.
10. Ramarao N, Sanchis V. The pore-forming haemolysins of Bacillus cereus: a review. Toxins. 2013;5:1119-39.
11. Guinebretière MH, Broussolle V, Nguyen-The C. Enterotoxigenic profiles of food-poisoning and food-borne *Bacillus cereus* strains. J Clin Microbiol. 2002;40(8):3053-6.
12. Martinez-Blanch JF, Sanchez G, Garay E, Aznar R. Development of a real-time PCR assay for detection and quantification of enterotoxigenic members of Bacillus cereus group in food samples. Int J Food Microbiol. 2009;135(1):15-21.
13. Ramarao N, Tran SL, Marin M, Vidic J. Advanced Methods for Detection of Bacillus cereus and Its Pathogenic Factors. Sensors (Basel). 2020;20(9).
14. Tran SL, Cormontagne D, Vidic J, Andre-Leroux G, Ramarao N. Structural Modeling of Cell Wall Peptidase CwpFM (EntFM) Reveals Distinct Intrinsically Disordered Extensions Specific to Pathogenic Bacillus cereus Strains. Toxins (Basel). 2020;12(9).
15. Tran SL, Ramarao N. Bacillus cereus immune escape: a journey within macrophages. FEMS Microbiol Lett. 2013;347:1-6.
16. Tran SL, Guillemet E, Ngo-Camus M, Clybouw C, Puhar A, Moris A, et al. Hemolysin II is a *Bacillus cereus* virulence factor that induces apoptosis of macrophages. Cell Microbiol. 2011;13:92-108.
17. Cadot C, Tran SL, Vignaud ML, De Buyser ML, Kolsto AB, Brisabois A, et al. InhA1, NprA and HlyII as candidates to differentiate pathogenic from non-pathogenic Bacillus cereus strains. J Clin Microbiol. 2010;48:1358-65.

18. Haydar A, Tran SL, Guillemet E, Darrigo C, Perchat S, Lereclus D, et al. InhA1-Mediated Cleavage of the Metalloprotease NprA Allows Bacillus cereus to Escape From Macrophages Front Microbiol. 2018;23:1063.
19. Ehling-Schulz M, Fricker M, Scherer S. Identification of emetic toxin producing Bacillus cereus strains by a novel molecular assay. FEMS Microbiol Lett. 2004;232(2):189-95.
20. Hoton FM, Andrup L, Swiecicka I, Mahillon J. The cereulide genetic determinants of emetic Bacillus cereus are plasmid-borne. Microbiology (Reading). 2005;151(Pt 7):2121-4.
21. Glasset B, Sperry M, Dervyn R, Herbin S, Brisabois A, Ramarao N. The cytotoxic potential of Bacillus cereus strains of various origins. Food Microbiol. 2021;98:103759.
22. Kamar R, Gohar M, Jéhanno I, Réjasse A, Kallassy M, Lereclus D, et al. Pathogenic Potential of Bacillus cereus Strains as Revealed by Phenotypic Analysis. J Clin Microbiol. 2013;51:320-3.
23. Porrini C, Guérin C, Tran SL, Dervyn R, Nicolas P, Ramarao N. Implication of a Key Region of Six Bacillus cereus Genes Involved in Siroheme Synthesis, Nitrite Reductase Production and Iron Cluster Repair in the Bacterial Response to Nitric Oxide Stress International Journal of Molecular Sciences. 2021;22(10):5079.
24. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. Embnet J. 2011;17:10.
25. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389-402.
26. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357-9.
27. Li H, Handsaker B, Wysoke rA, Fennell T, Ruan J, Homer N. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078-9.
28. Anders S, Pyl P, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. . Bioinformatics. 2015;31:166-9.
29. Engebretsen S, Bohlin J. Statistical predictions with glmnet. Clin Epigenetics. 2019;11:123.
30. Manzano M, Giusto C, Iacumin L, Cantoni C, Comi G. Molecular methods to evaluate biodiversity in Bacillus cereus and Bacillus thuringiensis strains from different origins. . Food Microbiol. 2009;26:259-64.
31. Antolinos V, Fernandez P, Ros-Chumillas M, Periago P, Weiss J. Development of a high-resolution melting-based approach for efficient differentiation among Bacillus cereus group isolates. Foodborne Pathog Dis. 2012;9:777-85.
32. Didelot X, Barker M, Falush D, Priest F. Evolution of pathogenicity in the Bacillus cereus group. Syst Appl Microbiol. 32:81-90.

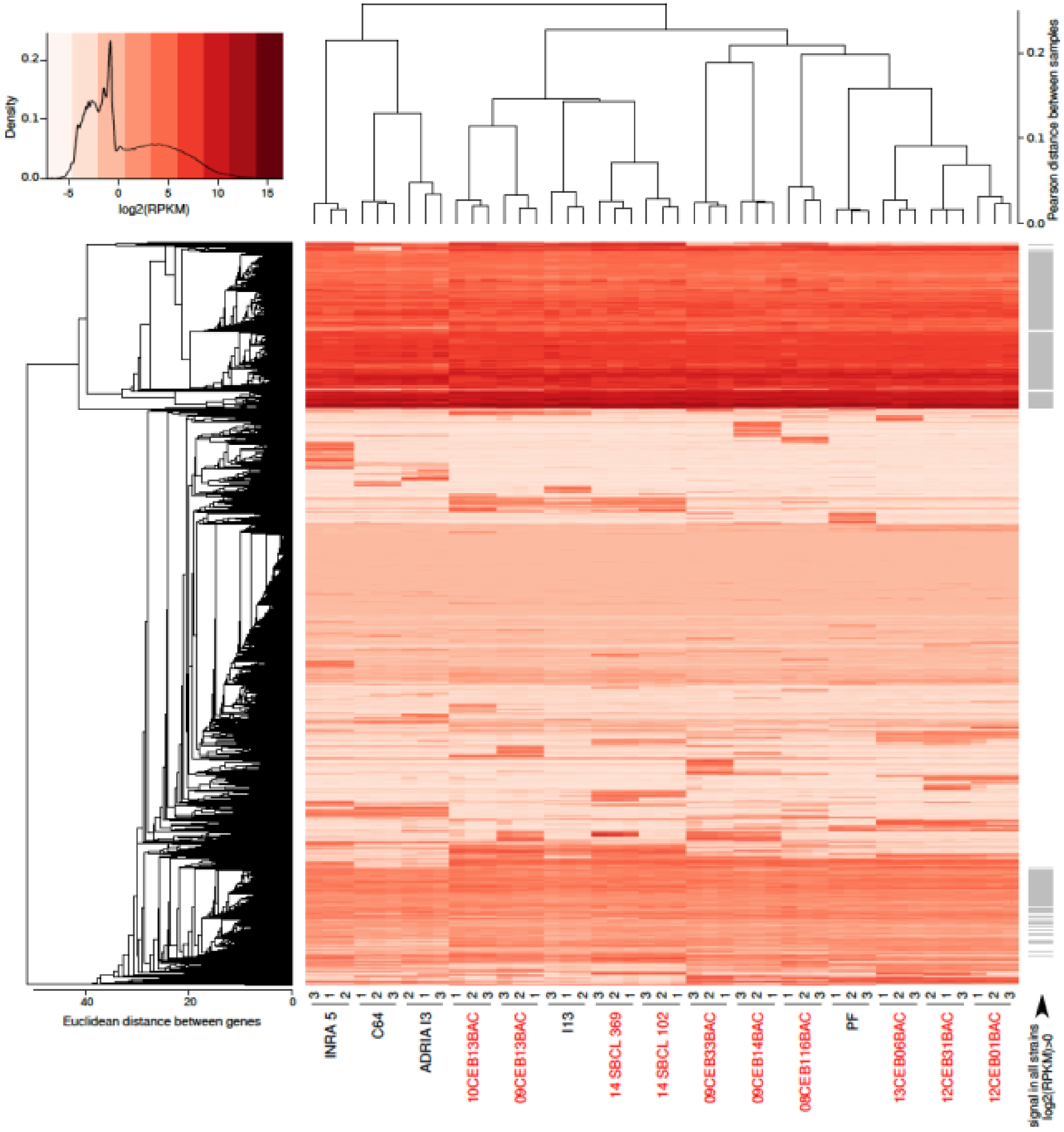
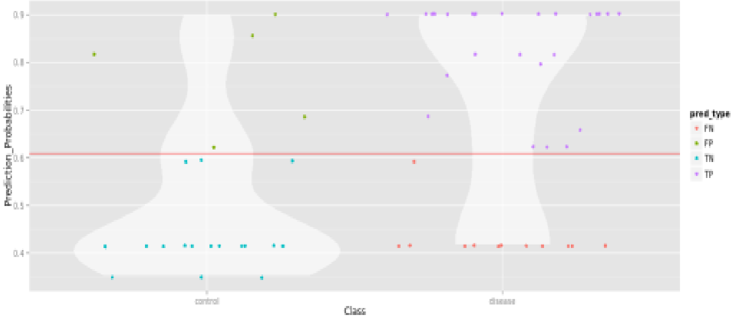

0.2
0.1
0.0
Density
-5
0
5
10
15
log2(RPKM)
0.2
0.1
0.0
Pearson distance between samples
40
20
0
Euclidean distance between genes
INRA 5
C64
ADRIA I3
10CEB13BAC
09CEB13BAC
I13
14 SBCL 369
14 SBCL 102
09CEB33BAC
09CEB14BAC
08CEB116BAC
PF
13CEB06BAC
12CEB31BAC
12CEB01BAC
signal in all strains log2(RPKM)>0

Figure 1

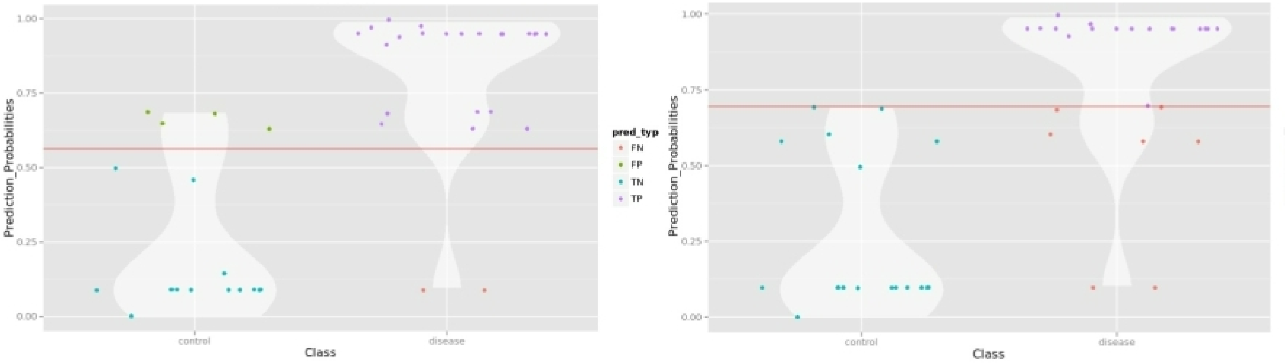
A

Biomarkers	Genes	AUC	SE	SP
Marker2-Marker3-Marker4-Marker6	agrC, thiJ, araC, gshAB	0.768	0.692	0.773



B

Biomarkers	Genes	AUC	SE	SP
Marker1-Marker2-Marker4-Marker5-Marker6	adhB, agrC, araC, BCQ_PI180, gshAB	0.917	0.917	0.778
Marker1-Marker3-Marker4-Marker5-Marker6	adhB, thiJ, araC, BCQ_PI180, gshAB	0.919	0.708	1.000



C

Biomarkers	Genes	AUC	SE	SP
Marker1-Marker2-Marker3-Marker6	adhB, agrC, thiJ, gshAB	0.955	0.909	0.864
Marker1-Marker2-Marker3-Marker5	adhB, agrC, thiJ, BCQ_PI180	0.955	0.909	0.864

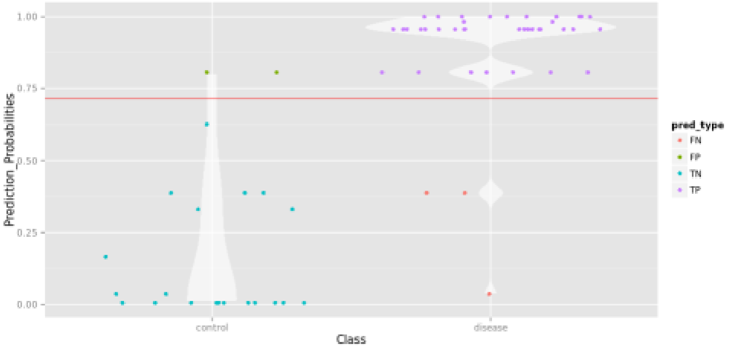


Figure 2

Table 1: list of the 7 selected biomarkers with gene position and putative function

	Marker 1	Marker 2	Marker 3	Marker 4	Marker 5	Marker 6	Marker 7
Marker name	adhB	agrC	thiJ	araC	BCQ_PI180	gshAB	BCQ_PI181
Gene name	BCAH187_RS12895	BCAH187_RS25230	BCAH187_RS22545	BCAH187_RS28400	BCAH187_RS28565	BCAH187_C0244	BCAH187_RS28570
Gene position	2465992 | 2466918	4769459 | 4769686	4287180 | 4287869	131495 | 132340	164163 | 164519 (complement)	167109 | 169376	164642 | 165757
Gene length	927 nt	228 nt	690 nt	846 nt	357 nt	2268 nt	1116 nt
Potential function	alcohol dehydrogenase catalytic domain-containing protein	hypothetical protein	type 1 glutamine amidotransferase domain-containing protein	AraC family transcriptional regulator	helix-turn-helix transcriptional regulator	bifunctional glutamate--cysteine ligase GshA/glutathione synthetase GshB	S-(hydroxymethyl)glutathione dehydrogenase/class III alcohol dehydrogenase
Start codon	ATG	ATG	ATG	ATG	ATG	ATG	TTG

Table 2: Estimated probability $\hat{z}_i$ for the 15 strains. A logistic regression model with lasso penalty was applied to select the penalty constant, which determines the number of selected genes.

Then prediction accuracy of the procedure was evaluated in a cross-validation framework. For each replicate, the model provides a probability $\hat{z}_i$ to be pathogenic, and we considered the average value over the three replicates as the prediction probability of the strain. The predicted non-pathogenicity corresponds to a $\hat{z}_i$ smaller than 0.5 and the predicted pathogenicity corresponds to $\hat{z}_i$ above 0.5.

NP	Prob mean
INRA 5	0.153328340753618
C64	0.0752423643321016
ADRIAI3	0.0437357685829226
I13	0.5
PF	0.599889993544854
FBO	
10CEB13BAC	0.993824252074421
08CEB116BAC	0.675323289631434
14SBCL102	0.953746924319411
14SBCL369	0.950799749333682
12CEB01BAC	0.382731024964747
Clinical	
09CEB13BAC	0.975134675591066
09CEB14BAC	0.890033149139494
09CEB33BAC	0.788491148616572
12CEB31BAC	0.977652814613013
13CEB06BAC	0.986545096552651

Table 3. Presence/absence of biomarkers among non-pathogenic (green), FBO (blue) and clinical (beige) strains. The presence of each biomarker gene was assessed by PCR in all strain of the collection. If the gene was present, a score of 1 was attributed (green boxes), if the gene is absent, a score of 0 is attributed (red boxes).

	Marker 1	Marker 2	Marker 3	Marker 4	Marker 5	Marker 6	Marker 7	PanC group
	adhB	agrC	thiJ	araC	BCQ_PI180	gshAB	BCQ_PI181	
INRA-PF_**S09**	0	0	1	0	0	0	0	III
I13_**S10**	1	0	0	0	0	0	0	IV
INRA-5_**S11**	0	0	0	0	0	0	0	VI
INRA-C64_**S12**	0	0	0	0	0	0	0	VI
ADRIA-I3_**S13**	0	0	0	0	0	0	0	VI
INRA-BN_**S36**	1	1	1	0	0	0	0	II
INRA-PA_**S37**	0	1	1	1	1	1	1	III
INRA-A3_**S38**	1	1	1	1	0	0	0	IV
I23_**S39**	0	0	1	0	0	0	0	IV
SB_**S40**	0	0	1	0	0	0	0	V
I11_**S41**	1	1	0	1	0	0	0	V
INRA-C1_**S42**	0	0	0	0	1	1	1	VI
INRA-C46_**S43**	0	0	0	0	0	1	0	VI
INRA-SL_**S44**	0	0	0	0	0	0	0	VI
INRA-SO_**S45**	0	0	0	0	0	0	0	VI
INRA-BC_**S47**	1	0	1	0	0	0	0	II
I2_**S48**	0	0	0	0	0	0	0	IV
INRA-BL_**S49**	0	0	0	0	0	0	0	VI
ADRIA I21_**S50**	0	0	0	0	0	0	0	VI
INRA-SV_**S51**	0	0	0	0	0	0	0	VI
WSBC-10204_**S52**	0	1	0	0	0	0	0	VI
08CEB116BAC _**S1**	1	1	1	0	1	1	1	II
10CEB13BAC _**S2**	1	1	1	1	1	1	1	IV
12CEB01BAC _ **S3**	1	1	1	1	0	0	0	III
14 SBCL 102 _**S4**	1	1	1	1	1	1	1	IV
14 SBCL 369 _**S5**	1	1	1	1	1	1	1	IV
09CEB01BAC_**S26**	1	1	1	1	1	1	1	III
09CEB04BAC_**S27**	1	1	1	1	1	1	1	VII
09CEB26BAC_**S28**	1	0	0	1	1	1	1	II
09CEB40BAC _**S29**	1	1	1	0	0	0	0	II
10CEB46BAC _**S30**	0	0	0	0	0	0	0	IV
10CEB88BAC_ **S31**	1	1	1	1	1	1	1	III
14 SBCL 013_ **S32**	1	1	1	1	1	1	1	III
14 SBCL 038 _**S33**	1	1	1	0	0	0	0	IV
14 SBCL 281_ **S34**	1	1	1	1	0	0	0	IV
14 SBCL 714 _**S35**	1	1	1	0	0	0	0	II
07CEB21BAC_ **S65**	1	1	1	1	1	1	1	III
07CEB48BAC _**S66**	1	1	1	1	0	0	1	III
07CEB53BAC _**S67**	0	1	1	1	1	1	1	III
08CEB121BAC _**S68**	0	0	0	0	0	0	0	IV
08CEB145BAC_ **S69**	0	0	0	0	0	0	0	IV
08CEB037BAC _**S70**	0	0	0	0	0	0	0	IV

08CEB049BAC _**S71**	1	1	1	1	1	1	1	III
08CEB075BAC _**S72**	1	1	1	1	1	1	1	III
09CEB03BAC_**S73**	0	0	1	1	0	0	0	III
09CEB05BAC_**S74**	1	1	1	1	1	1	1	III
09CEB38BAC _**S75**	1	1	1	1	1	1	1	III
10CEB06BAC _**S76**	1	1	1	1	1	1	1	III
10CEB33BAC _**S77**	1	1	1	1	1	1	1	III
10CEB68BAC_ **S78**	1	1	1	1	1	1	0	III
14 SBCL 008 _**S79**	0	0	0	0	0	0	0	IV
14 SBCL 016_ **S80**	0	0	0	0	0	0	0	IV
14 SBCL 020 _**S81**	0	0	0	0	0	0	0	IV
14 SBCL 022 _**S82**	0	0	0	0	0	0	0	IV
14 SBCL 049_ **S83**	0	1	0	0	0	0	0	IV
14 SBCL 175 _**S84**	0	0	0	1	0	0	0	VII
14 SBCL 180 _**S85**	0	0	0	0	0	0	0	IV
14 SBCL 266 _**S86**	0	0	0	0	0	0	0	IV
14 SBCL 374 _**S87**	0	0	0	0	0	0	0	iV
14 SBCL 566 _**S88**	0	1	1	1	1	1	0	III
09CEB13BAC_**S6**	1	1	1	1	1	1	1	IV
09CEB14BAC_**S7**	1	1	1	1	1	1	1	II
09CEB33BAC_**S8**	1	1	1	1	1	1	1	III
12CEB31BAC_**S14**	1	1	1	1	1	1	1	III
13CEB06BAC_**S15**	1	1	1	1	1	1	1	III
09CEB11BAC_**S16**	1	1	1	1	1	1	1	III
09CEB16BAC_**S17**	1	1	1	1	1	1	1	III
12CEB30BAC_**S18**	1	1	1	1	1	1	1	II
12CEB40BAC_**S20**	1	1	1	1	1	1	1	III
12CEB46BAC _**S21**	1	1	1	1	1	1	1	IV
12CEB47BAC_**S22**	1	1	1	0	0	0	0	IV
12CEB51BAC_**S23**	1	1	1	1	1	1	1	II
13CEB01BAC_**S24**	1	1	0	0	0	0	0	III
09CEB12BAC_**S53**	1	1	1	1	1	1	1	III
09CEB34BAC_**S59**	1	1	1	1	1	1	1	III
09CEB36BAC_**S61**	1	1	1	0	1	1	1	III
12CEB34BAC_**S64**	1	1	1	0	0	0	0	IV
12CEB37BAC_**S90**	1	0	0	0	0	0	0	IV
12CEB38BAC_**S91**	1	1	1	1	1	1	1	III
12CEB39BAC_**S92**	1	1	1	1	1	1	1	III
12CEB42BAC_**S94**	1	1	1	1	1	1	1	III
12CEB43BAC_**S95**	1	1	1	1	0	0	0	III
12CEB44BAC_**S96**	1	1	1	1	1	1	1	IV
12CEB45BAC_**S97**	1	1	1	1	1	1	1	II
12CEB48BAC_**S98**	1	1	1	1	1	1	1	II
12CEB49BAC_**S99**	1	1	0	0	0	0	0	IV
12CEB50BAC_**S100**	1	1	1	0	0	0	0	IV
12CEB52BAC_**S101**	0	1	1	1	0	0	0	III
13CEB03BAC_**S102**	1	1	1	1	1	1	1	II
13CEB07BAC_**S105**	1	1	1	1	1	1	1	III
13CEB09BAC_**S106**	1	1	1	1	1	1	1	III
13CEB30BAC_**S107**	1	1	1	0	0	0	0	II
14CEB16BAC_**S114**	1	1	1	1	1	1	1	IV
14CEB17BAC_**S115**	1	1	1	1	1	1	1	III
14SBCL987_**S116**	1	1	1	0	0	0	0	IV

Supplementary Table 1: Strain table

A- Origin of the 21 non-pathogenic strains and their genetic signature

Strain	Source	Genetic Signature (GS)	panC group
INRA-PF_**S09**	Milk protein	10	III
I13_**S10**	Cooked rice	2	IV
INRA-5_**S11**	Pasteurized zucchini puree	8	VI
INRA-C64_**S12**	Pasteurized vegetables	8	VI
ADRIA-I3_**S13**	Cooked foods	8	VI
INRA-BN_**S36**	Vegetable	12	II
INRA-PA_**S37**	Milk protein	4	III
INRA-A3_**S38**	Starch	2	IV
I23_**S39**	Cooked apple	10	IV
SB_**S40**	Soil from a vegetable field	10	V
I11_**S41**	Cooked food	5	V
INRA-C1_**S42**	Pasteurized vegetables	8	VI
INRA-C46_**S43**	Pasteurized vegetables	8	VI
INRA-SL_**S44**	Soil	8	VI
INRA-SO_**S45**	Soil	8	VI
INRA-BC_**S47**	Vegetable	2	II
I2_**S48**	Dried fruit	2	IV
INRA-BL_**S49**	Vegetable	8	VI
ADRIA I21_**S50**	Cooked foods	8	VI
INRA-SV_**S51**	Soil	8	VI
WSBC 10204_**S52**	Pasteurized milk	8	VI

B- Epidemiological data and symptoms of the 39 selected food-borne outbreaks (FBO) strictly associated to *B. cereus* and GS of the associated strains

Key of strains	**Year**	**Incriminated food**	**Number of human cases**	**Incubation period (h)**	**Symptoms**	**Place of outbreaks**	**CFU/g**	**Genetic Signature (GS)**	**panC group**
08CEB116BAC **_S1**	2009	Semolina	40	12	Diarrhea	Staff canten	1,20E+03	1	II
10CEB13BAC **_S2**	2006	Paella	27	7	Diarrhea	Medico-social institute	2,80E+04	2	IV
12CEB01BAC **_ S3**	2006	Apricot compote	8	5-16	Vomiting	School canteen	7,00E+02	1	III
14 SBCL 102 **_S4**	2007	Lamb meat	5	8	Vomiting-diarrhea	Canteen of company	2,30E+03	2	IV
14 SBCL 369 **_S5**	2005	Vetebales soup	10	12-24	Vomiting-diarrhea	School canteen	9,10E+02	2	IV
09CEB01BAC**_S26**	2008	Tiramisu	15	1	Vomiting-diarrhea	Canteen of company	8,00E+02	9	III
09CEB04BAC**_S27**	2004	Mashed potatoes	24	not known	Vomiting-diarrhea	School or equivalent	4,00E+02	7	VII
09CEB26BAC**_S28**	2008	Quenelle of pike	15	2	Vomiting-diarrhea-other	Canteen of company	1,20E+03	6	II
09CEB40BAC **_S29**	2009	Squid-sauce	3	12	Diarrhea	Canteen of company	2,10E+05	12	II
10CEB46BAC **_S30**	2008	Taboulesh	11	not known	Abdominal pain-other	Canteen of hospital	not known	2	IV
10CEB88BAC**_ S31**	2011	Rice salad	8	1-1,5	Vomiting-diarrhea	Family	1,70E+07	3	III
14 SBCL 013**_ S32**	2002	Mashed potatoes	10	not known	Vomiting-diarrhea	School or equivalent	7,80E+04	4	III
14 SBCL 038 **_S33**	2011	Samoussa	9	1	Nausea-other	Restaurant or equivalent	not known	6	IV
14 SBCL 281**_ S34**	2012	Onion soup	5	8-12	Vomiting	School canteen	4,00E+02	2	IV
14 SBCL 714 **_S35**	2004	Polenta	25	18-24	Abdominal pains-diarrhea	Medico-social institute	9,00E+03	5	II
07CEB21BAC**_ S65**	2007	Semolina	5	2	Vomiting	Commercial catering	1,20E+07	3	III
07CEB48BAC **_S66**	2011	Shrimp	12	24	Vomiting-diarrhea	Commercial catering	6,80E+04	3	III
07CEB53BAC **_S67**	2012	Tomatoes	4	2-3	Vomiting-diarrhea	Commercial catering	7,00E+02	3	III
08CEB121BAC **_S68**	2010	Taboulesh	not known	not known	not known	Restaurant or equivalent	5,00E+03	4	II

08CEB145BAC_ **S69**	2012	Comosed salad (rice or corn)	2	not known	Abdominal pain-vomiting	Canteen of company	1,90E+03	4	II
08CEB037BAC _**S70**	2001	Rice salad	13	4-24	Vomiting-other	The elderly	2,00E+03	4	IV
08CEB049BAC _**S71**	2003	Semolina	4	0,5-3	Vomiting	Restaurant or equivalent	5,50E+04	3	III
08CEB075BAC _**S72**	2006	Fruit salad	70	not known	not known	Canteen of company	6,30E+03	3	III
09CEB03BAC_ **S73**	2002	Fish in coconut milk	2	2 to 3	Nausea-other	Restaurant or equivalent	1,10E+04	3	III
09CEB05BAC_**S74**	2007	Cantonese rice	2	0,5	Vomiting-other	Family	1,60E+05	3	III
09CEB38BAC _**S75**	2009	Chicken sauce	15	not known	Vomiting-diarrhea	Restaurant or equivalent	5,00E+02	3	III
10CEB06BAC _**S76**	2003	Pasta gratin	2	2	Vomiting-diarrhea	Family	1,50E+07	3	III
10CEB33BAC _**S77**	2007	Chicken	8	5	Vomiting-diarrhea	Family	6,50E+04	3	III
10CEB68BAC_ **S78**	2010	Mashed vegetables	19	not known	Vomiting-diarrhea-other	Canteen of social activities	1,20E+04	1	III
14 SBCL 008 _**S79**	2001	Carrot	3	5	Vomiting-diarrhea-other	Restaurant or equivalent	5,80E+03	2	IV
14 SBCL 016_ **S80**	2003	Tomatoes	3	15	Diarrhea	Hospital	5,50E+03	2	IV
14 SBCL 020 _**S81**	2005	Composed salad	3	2	Vomiting-diarrhea	Canteen of hospital	2,00E+03	2	IV
14 SBCL 022 _**S82**	2005	Tomatoe-corn-courgette	9	8-10	Abdominal pain-vomiting	School canteen	4,00E+03	2	IV
14 SBCL 049_ **S83**	2006	Composed salad	8	6-34	Abdominal pain-vomiting-other	Family	4,00E+02	2	IV
14 SBCL 175 _**S84**	2011	Mashed fish	18	12	Vomiting-diarrhea	Residence for the elderly	4,00E+02	7	VII
14 SBCL 180 _**S85**	2011	Diced mixed vegetables	14	1-21	Vomiting-diarrhea	Residence for the elderly	4,00E+02	2	IV
14 SBCL 266 _**S86**	2012	Millefeuille	2	4	Nausea	Restaurant or equivalent	2,00E+03	2	IV
14 SBCL 374 _**S87**	2006	Composed salad	not known	7	Abdominal pain	School canteen	5,50E+02	2	IV
14 SBCL 566 _**S88**	2008	Mix of pie	19	5-24	Vomiting-diarrhea	Canteen for social activities	4,00E+02	1	III

C- Epidemiological data and symptoms of the 35 selected *B. cereus* positive clinical samples and GS of the associated strains.

Key of strains	date of sampling	Hospital ward	Age of patients	Type of sampling	Symptoms	Outcomes	Genetic Signature (GS)	panC group
09CEB13BAC**_S6**	16/06/2009	Neonatology	Premature newborn	Blood culture	Brain abscess	Recovery	2	IV
09CEB14BAC**_S7**	05/07/2009	Neonatology	Premature newborn	Blood culture	Bacteremia	Recovery	1	II
09CEB33BAC**_S8**	03/09/2009	Neonatology	Newborn	Axilla-later feces	Skin infection	Recovery	1	III
12CEB31BAC**_S14**	08/2011	Neonatology	Premature newborn	Blood culture	Organ failure and pulmonary and cerebral abscesses	Death	4	III
13CEB06BAC**_S15**	juin-11	Intensive care unit	86	Blood culture from catheter	Heart failure, ventilator-associated pneumonia, ischemic stroke	Recovery	1	III
09CEB11BAC**_S16**	28/07/2009	Neonatology	Premature newborn	Blood culture	Meningitis, infection in the liver, both lungs	Death	1	III
09CEB16BAC**_S17**	21/07/2009	Neonatology	Newborn	Umbilical	Local colonization	Recovery	1	III
12CEB30BAC**_S18**	02/08/2011	Neonatology	Premature newborn	Blood culture	Sepsis	Recovery	4	II
12CEB40BAC**_S20**	03/03/2010	Gastroenterology	63	Blood culture	Bacteremia and central venous catheter-linked infection	Recovery	3	III
12CEB46BAC **_S21**	07/12/2010	Hematology	61	Blood culture	Sepsis (patient with an acute myeloid leukemia)	Recovery	2	IV
12CEB47BAC**_S22**	15/06/2008	Neurology	43	Blood culture	Bacteremia	Recovery	6	IV
12CEB51BAC**_S23**	16/07/2010	Cardiac surgery	60	blood culture	Sternum abscess, absent fever	Sequela of osteitis	1	II
13CEB01BAC**_S24**	07/2011	Orthopedic surgery	31	Prosthesis from tibia	No clinical sign of infection	Recovery	9	III
09CEB12BAC**_S53**	28/07/2009	Neonatology	Premature newborn	Cerebrospinal fluid	Meningitis, infection in the liver, both lungs	Death	1	III

09CEB34BAC_**S59**	17/09/2009	Neonatology	Premature-newborn	Stomach-tube feeding	Premature birth	Recovery	3	III
09CEB36BAC_**S61**	21/09/2009	Neonatology	Premature-newborn	Central venous catheter	Bacteremia	Recovery	1	III
12CEB34BAC_**S64**	06/2009	Emergency	80	Thoracentesis	Pulmonary infection	not known	2	IV
12CEB37BAC_**S90**	18/09/2011	Intensive care unit	30	Blood culture	Endocarditis	Death	2	IV
12CEB38BAC_**S91**	02/11/2009	Hematology	65	Blood culture	Sepsis	Death	1	III
12CEB39BAC_**S92**	12/09/2011	Nephrology	54	Blood culture	Sepsis	Recovery	1	III
12CEB42BAC_**S94**	26/03/2010	Gastroenterology	63	Blood culture	Bacteremia and central venous catheter-linked infection	Recovery	3	III
12CEB43BAC_**S95**	27/05/2010	Gastroenterology	63	Blood culture	Bacteremia and central venous catheter-linked infection	Recovery	1	III
12CEB44BAC_**S96**	03/06/2008	Surgery	34	Blood culture	Bacteremia	Recovery	4	IV
12CEB45BAC_**S97**	27/11/2010	Neurology	newborn	Blood culture	Kidneys and urinary infections	Recovery	1	II
12CEB48BAC_**S98**	06/10/2009	Oncology	66	Blood culture	Bacteremia (patient with a colorectal cancer)	Recovery	1	II
12CEB49BAC_**S99**	24/09/2010	Hematology	24	Blood culture+ skin infection	Sepsis and aplastic anemia caused by drugs	Recovery	2	IV
12CEB50BAC_**S100**	12/08/2009	Gynecological surgery	77	Blood culture	Bacteremia (patient with breast cancer)	Recovery	2	IV
12CEB52BAC_**S101**	20/06/2008	Hematology	40	Blood culture	Bacteremia (immunocompromised patient)	Recovery	4	III
13CEB03BAC_**S102**	oct-11	Intensive care unit	76	Blood culture	Community acquired pneumonia	Recovery	1	II
13CEB07BAC_**S105**	oct-11	Emergency	24	Blood culture	Abdominal pain, shivering, vomiting, fever, diarrhea	Recovery	3	III
13CEB09BAC_**S106**	sept-12	Gastroenterology	85	Liver abscess	Sepsis, hepatitis c and liver abscess,	Recovery	3	III

					abdominal pain, diarrhea			
13CEB30BAC_**S107**	sept-13	not known	not known	Blood culture	Nausea, abdominal pain and vomiting	not known	5	II
14CEB16BAC_**S114**	déc-13	Clinical laboratory	Premature newborn	Blood culture from peripheral veins	Septic shock, multiple organ failure, pulmonary and cerebral abscesses	Death	2	IV
14CEB17BAC_**S115**	déc-13	Clinical laboratory	Premature newborn	Bronchial aspiration (lung)	Septic shock and pneumonia pulmonary necrotic abscesses, recurrent pneumothorax	Death	4	III
14SBCL987_**S116**	2014	not known	not known	Biopsy (kidney)	Vomiting and diarrhea	Death	5	IV

Supplementary Table 2: Strains selected for the RNAseq study and representative of the three collections FBO (F), Clinical (C) and non pathogenic (NP). For each strain, the name, origin, Nhe and Hbl production as well as cytotoxicity to HeLa and Raw cells is indicated.

Strains		Samples	Symptoms (n)	Nhe indice	Hbl indice	cytotoxic activity on Hela cells	cytotoxic activity on Raw cells
NP	INRA-PF_S09	Milk protein	-	3-4	1/64	57%	16%
	I13_S10	Cooked rice	-	3	1/64	6%	0%
	INRA 5_S11	Pasteurized zucchini puree	-	2	1/4	11%	5%
	INRAC64_S12	Pasteurized vegetables	-	2-3	1/16	20%	2%
	ADRIA I3_S13	Cooked foods	-	2	1	7%	2%
F	08CEB116BAC_S1	Semolina	Diarrhea (40)	1	nd	7%	8%
	10CEB13BAC_S2	Paella	Diarrhea (27)	3	1/16	77%	44%
	12CEB01BAC_S3	Apricot compote	Vomit (8)	5	nd	77%	21%
	14SBCL102_S4	Ham	Diarrhea and vomit (5)	4	1/64	89%	86%
	14SBCL369_S5	Vegetables soup	Diarrhea and vomit (10)	3	1/64	76%	84%
C	09CEB13BAC_S6	blood culture	brain abscess (1)	3	1/16	77%	47%
	09CEB14BAC_S7	blood culture	bacteremia (1)	2	nd	88%	40%
	09CEB33BAC_S8	axilla later feces	skin infection (1)	4	nd	25%	12%
	12CEB31BAC_S14	blood culture	Apnea, bradycardia, and gray complexion. after that, sepsis, organ failure and pulmonary and cerebral abscesses (1)	5	nd	100%	48%
	13CEB06BAC_S15	blood culture from catheter	heart failure, ventilator-associated pneumonia, ischemic stroke (1)	5	nd	11%	7%

nd : not detected

Supplementary Table 3: Primers used in this study

Primer purpose and target gene	Primer (a)	Primer sequence (5'-3') (b)	Annealing temp (°C)	Product size (bp)	reference or source
agrC	BBC-01-F	TATCCT**R**GTTATAGCATTTTAGC	55	131	this study
	BBC-02-R	GTTAGTATGTATCC**R**AAGA**Y**GCAGTAGA	55		this study
adhB	BBC-03-F	TTATTATCTATTCTTTCGTGTGATGC	55	275	this study
	BBC-04-R	CTATTTGTAGCAGAACATTC**R**AAACC	55		this study
BCQ_PI181	BBC-05-F	TCGATGTAGAAGAGCCAAAAGC	55	289	this study
	BBC-06-R	CCTTTACCTTGTGTTTCTCG	55		this study
BCQ_PI180	BBC-07-F	ATGCAACAGCAGCT**Y**TACTTTTCAA	55	251	this study
	BBC-08-R	TGTAACAAACACCATAT**W**ATTGCTATT	55		this study
araC	BBC-09-F	GTACAGTTAAAAGC**Y**TTTCC	55	221	this study
	BBC-10-R	GG**RTY**TTCCCATGACATATCTA	55		this study
gshAB	BBC-11-F	ACGAAATGCTTTGGCCATTAAG	55	284	this study
	BBC-12-R	CCATCGATAGTGTAAATAATT	55		this study
thiJ	BBC-13-F	GCTGTTATTTATTACGCAGG	55	251	this study
	BBC-14-R	ATCTTCTGTTAAAAATGGAAC	55		this study

(a) F, forward primer; R, reverse primer

(b) R, A or G;Y, C or T; W, A or T

Supplementary Table 4. RPKM data for the 7 markers. The expression levels expressed as $\log_2$ scaled rpkm (reads per kilobase per million mapped reads) is indicated for each of the 7 marker genes and the 15 samples in biological triplicate (1, 2, 3).

	Marker1	**Marker2**	**Marker3**	**Marker4**	**Marker5**	**Marker6**	Marker7
	adhB	**agrC**	**thiJ**	**araC**	**BCQ_PI180**	**gshAB**	**BCQ_PI181**
INRA-PF_S09-1	-2,87	-1,78	0,68	-3,68	-1,49	-1,49	-4,07
INRA-PF_S09-2	-3,81	-1,78	0,60	-3,68	-2,43	-2,43	-3,04
INRA-PF_S09-3	-3,81	-1,78	1,37	-3,68	-1,65	-1,65	-4,07
I13_S10-1	1,66	-1,78	-3,37	-0,58	-1,37	-1,37	-4,07
I13_S10-2	0,47	-1,78	-2,22	-0,09	-1,28	-1,28	-2,92
I13_S10-3	0,83	-1,78	-3,37	-3,68	-2,43	-2,43	-4,07
INRA-5_S11-1	-2,33	-1,78	-3,37	-3,68	-2,43	-2,43	-4,07
INRA-5_S11-2	-3,81	-0,73	-3,37	-3,68	-2,43	-2,43	-4,07
INRA-5_S11-3	-3,81	-1,10	-3,37	-3,68	-1,75	-1,75	-4,07
INRA-C64_S12-1	-3,81	-1,78	-3,37	-1,84	-2,43	-2,43	-2,24
INRA-C64_S12-2	-3,81	-1,78	-2,25	-3,68	-2,43	-2,43	-2,33
INRA-C64_S12-3	-3,81	-1,78	-3,37	-3,68	-2,43	-2,43	-4,07
ADRIA-I3_S13-1	-3,81	-1,78	-3,37	-3,68	-2,43	-2,43	-4,07
ADRIA-I3_S13-2	-1,33	-1,78	-2,43	-3,68	-1,49	-1,49	-4,07
ADRIA-I3_S13-3	-3,81	-0,56	-3,37	-3,68	-2,43	-2,43	-4,07
08CEB116BAC_S1-1	3,46	4,52	0,06	-1,30	5,16	5,16	4,31
08CEB116BAC_S1-2	3,02	4,98	1,23	-3,68	4,83	4,83	5,18
08CEB116BAC_S1-3	1,31	3,65	1,30	-3,68	4,34	4,34	3,72
10CEB13BAC_S2-1	2,34	5,29	3,20	4,45	4,70	4,70	5,42
10CEB13BAC_S2-2	2,45	3,96	3,36	3,98	5,37	5,37	4,79
10CEB13BAC_S2-3	1,65	4,83	1,80	4,12	4,80	4,80	5,19

12CEB01BAC_S3-1	1,74	4,73	3,08	5,25	2,92	2,92	2,13
12CEB01BAC_S3-2	1,33	4,17	3,08	4,55	-2,43	-2,43	-2,64
12CEB01BAC_S3-3	1,15	4,17	1,07	5,13	-2,43	-2,43	-4,07
14SBCL102_S4-1	2,41	4,04	1,55	4,96	5,08	5,08	5,08
14SBCL102_S4-2	2,66	4,14	1,63	5,20	4,59	4,59	5,22
14SBCL102_S4-3	0,29	4,06	1,73	5,04	3,27	3,27	3,65
14SBCL369_S5-1	2,50	4,74	2,25	5,27	4,76	4,76	4,72
14SBCL369_S5-2	2,80	3,97	2,94	5,43	4,29	4,29	4,69
14SBCL369_S5-3	1,21	4,33	2,21	5,46	3,18	3,18	4,05
09CEB13BAC_S6-1	1,58	5,81	1,35	7,19	3,68	3,68	6,51
09CEB13BAC_S6-2	0,80	5,53	2,92	7,18	3,56	3,56	6,75
09CEB13BAC_S6-3	1,86	4,44	2,76	6,68	4,69	4,69	6,64
09CEB14BAC_S7-1	3,83	3,54	2,28	6,40	4,16	4,16	5,62
09CEB14BAC_S7-2	3,02	4,66	2,84	6,26	3,88	3,88	5,00
09CEB14BAC_S7-3	2,79	3,52	2,74	6,85	4,02	4,02	4,71
09CEB33BAC_S8-1	2,66	3,51	1,86	6,61	4,23	4,23	6,42
09CEB33BAC_S8-2	3,11	2,55	2,22	6,39	4,52	4,52	6,34
09CEB33BAC_S8-3	1,93	2,85	2,06	6,68	3,86	3,86	5,68
12CEB31BAC_S14-1	2,06	4,96	2,03	5,87	3,44	3,44	4,70
12CEB31BAC_S14-2	1,10	3,08	2,21	6,00	4,11	4,11	4,57
12CEB31BAC_S14-3	2,26	4,68	1,37	5,76	4,14	4,14	3,97
13CEB06BAC_S15-1	2,28	4,58	2,65	3,90	3,44	3,44	3,96
13CEB06BAC_S15-2	3,51	5,22	1,80	4,27	3,62	3,62	4,28
13CEB06BAC_S15-3	2,73	4,74	2,03	4,06	4,22	4,22	3,69