Clinical Microbiology and Infection xxx (xxxx) xxx

Contents lists available at ScienceDirect

Clinical Microbiology and Infection

journal homepage: www.clinicalmicrobiologyandinfection.com

Original article

New genetic biomarkers to differentiate non-pathogenic from 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], Nalini Ramarao [1, *]

[1)] *INRAE, Micalis, University Paris-Sacly, Jouy-en-Josas, France*
[2)] *ANSES, Universtité Paris-Est, Laboratory of Food Safety, Maisons-Alfrot, France*
[3)] *INRAE, MaIAGE, Université Paris-Saclay, Jouy-en-Josas, France*

ARTICLE INFO

Article history:
Received 10 February 2021
Received in revised form
18 May 2021
Accepted 22 May 2021
Available online xxx

Editor: S. Malhotra-Kumar

Keywords:
Bacillus cereus
CombiRoc analysis
Genetic biomarkers
Pathogenicity
Transcriptomic

ABSTRACT

Objectives: *Bacillus cereus* is responsible for food poisoning 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 non-pathogenic from clinically relevant *B. 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 (food-borne outbreaks, clinical and non-pathogenic). Receiver operating characteristic area under the curve (AUC) analysis was used to determine the combination of biomarkers, which best differentiate between the "disease" versus "non-disease" groups.

Results: Seven 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 four 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. **Devon W. Kavanaugh, Clin Microbiol Infect 2021;▪:1**

Introduction

Bacillus cereus is the third causative agent of food-borne outbreaks (FBOs) 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 showed 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

* Corresponding author. Nalini Rama Rao, INRAE, Micalis Institute, 78350, Jouy-en-Josas, France.
E-mail address: nalini.ramarao@inrae.fr (N. Ramarao).
† These authors contributed equally to this work.

https://doi.org/10.1016/j.cmi.2021.05.035
1198-743X/

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.

Materials and 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] (Table S1). We have previously shown a correlation between cytotoxicity and virulence [21]. Nevertheless, although these non-pathogenic 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 (Table S2) in triplicates. Bacterial cultures were incubated in BHI medium at 30°C in microaerophilic condition (5% O_2 to 15% CO_2 to 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 seven 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 (Table S3).

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.

Please cite this article as: Kavanaugh DW et al., New genetic biomarkers to differentiate non-pathogenic from clinically relevant *Bacillus cereus* strains, Clinical Microbiology and Infection, https://doi.org/10.1016/j.cmi.2021.05.035

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 receiver operating characteristic (ROC) area under the curve (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 (Fig. 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 seven biomarkers by logistic regression analysis

A Mann–Whitney–Wilcoxon non-parametric rank test with a classical 5% of q value 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 seven genes were selected (Table 1).

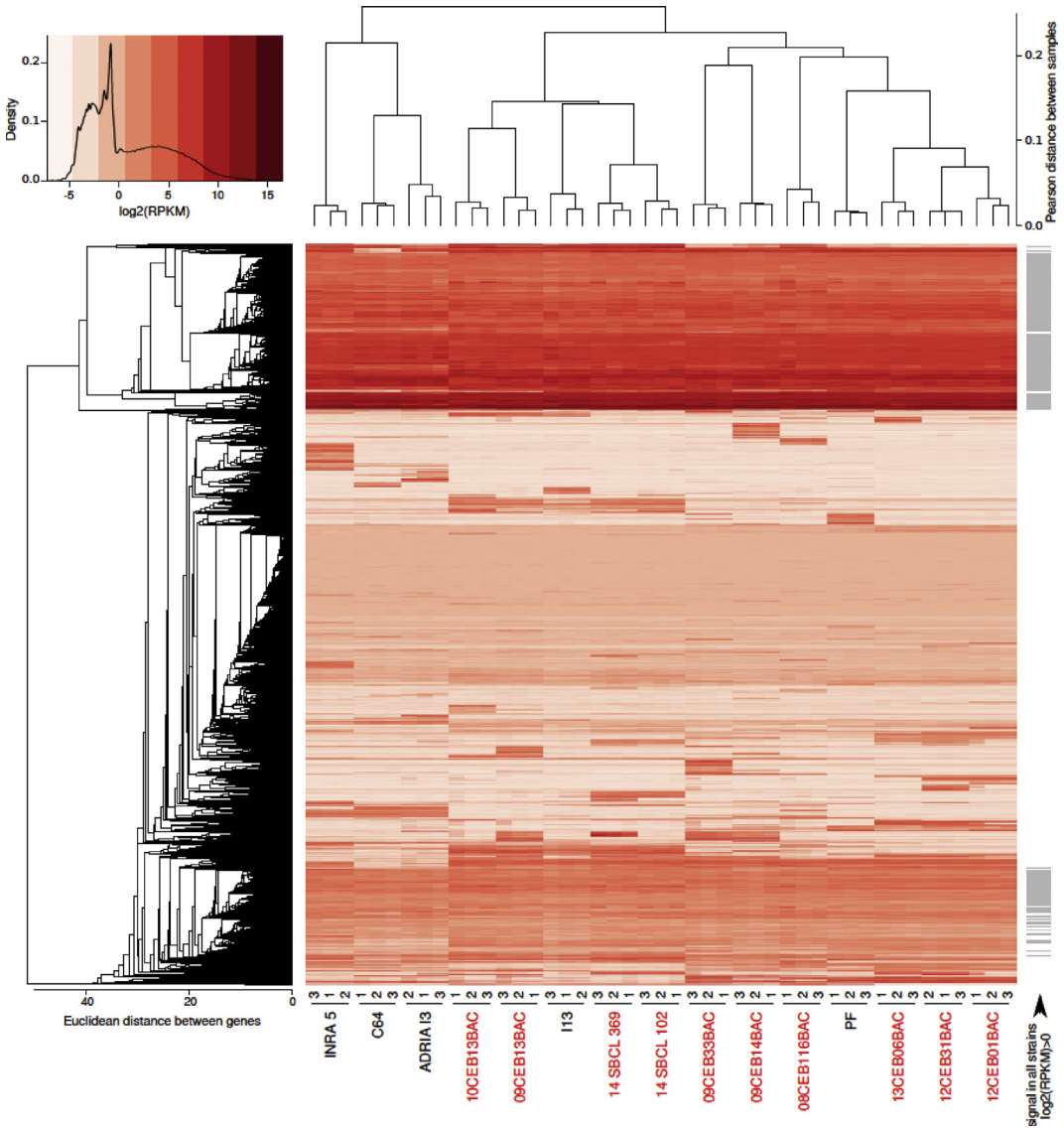


Fig. 1. RNAseq heatmap. Heatmap representation of expression levels (log_2 rpkm) across the pangenomic repertoire of 23 815 genes (rows) and the 45 samples (columns). Dendrograms are built by hierarchical clustering with average-link. The 3272 genes with signal in all strains are indicated by grey bars. Non-pathogenic strains are indicated in black and pathogenic strains in red.

Please cite this article as: Kavanaugh DW et al., New genetic biomarkers to differentiate non-pathogenic from clinically relevant *Bacillus cereus* strains, Clinical Microbiology and Infection, https://doi.org/10.1016/j.cmi.2021.05.035

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

With the RPKM values of these seven genes (Table S4), 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 (non-pathogenic strain INRA-PF predicted as pathogenic) and one false negative (pathogenic FBO strain 12CEB01BAC predicted as non-pathigenic) (Table 2).

Validation of the biomarkers on a large strain collection

Initially, for the first 15 strains, the presence of the seven 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 seven 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 seven genes selected (Table 3) and to determine the 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 non-pathogenic (NP) from FBO strains was obtained with four biomarkers (Fig. 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 (Fig. 2B).

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

Non-Pathogenic	Prob mean
INRA-5	0.153328340753618
INRA-C64	0.0752423643321016
ADRIAI3	0.0437357685829226
I13	0.5
INRA-PF	0.599889993544854
Food-borne outbreaks	
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

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 (Fig. 2C). These two combinations allowed the accurate discrimination between the two strain populations. Some markers have the same occurrence within the strain collection [5–7] 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 multilocus sequence (MLST) [32] or the classification of the strains according to their affiliation to a phylogenetic group that offers a first useful indicator of

Please cite this article as: Kavanaugh DW et al., New genetic biomarkers to differentiate non-pathogenic from clinically relevant *Bacillus cereus* strains, Clinical Microbiology and Infection, https://doi.org/10.1016/j.cmi.2021.05.035

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 independent 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 four 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

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 adhB	Marker 2 agrC	Marker 3 thiJ	Marker 4 araC	Marker 5 BCQ_PI180	Marker 6 gshAB	Marker 7 BCQ_PI181	PanC group
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

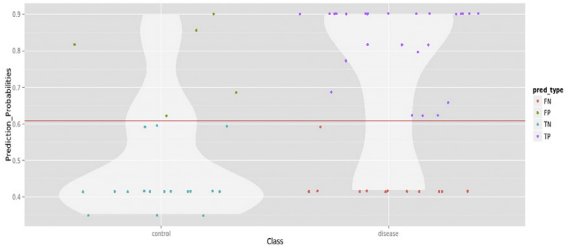
Please cite this article as: Kavanaugh DW et al., New genetic biomarkers to differentiate non-pathogenic from clinically relevant *Bacillus cereus* strains, Clinical Microbiology and Infection, https://doi.org/10.1016/j.cmi.2021.05.035

Table 3 ***continued***

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

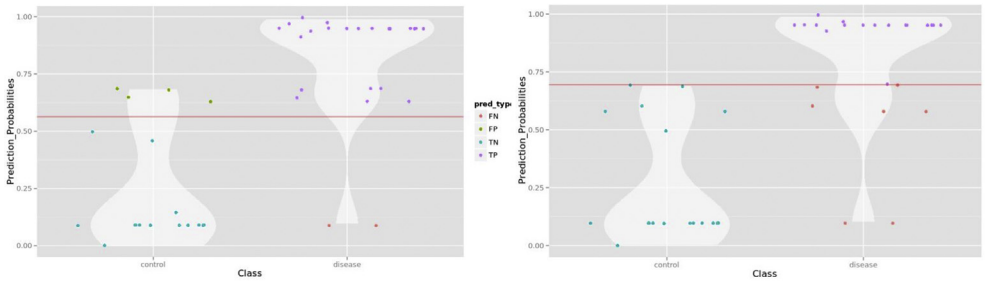
A

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

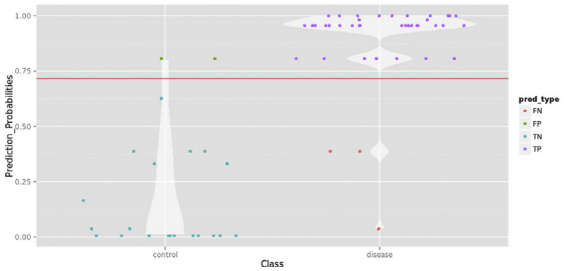


Fig. 2. CombiROC analysis results. The presence/absence matrix resulting from PCR detection of biomarker sequences was analyzed by CombiROC. (A) Foodborne outbreak strains (FBO) versus non-pathogenic; (B) FBO versus non-pathogenic strains, excluding phylogenetic group IV. Links best sensitivity performance,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

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 four 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

Please cite this article as: Kavanaugh DW et al., New genetic biomarkers to differentiate non-pathogenic from clinically relevant *Bacillus cereus* strains, Clinical Microbiology and Infection, https://doi.org/10.1016/j.cmi.2021.05.035

new findings may be relevant to gain additional knowledge on the strains found in hospitals and healthcare settings.

Transparency declaration

The authors declare no conflict of interest. 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

D.K., B.G., R.D.: performed experiments, analysed data, manuscript writing; C.G., S.P., P.N.: analysed data; S.H., A.B.: supervision; N.R.: initial concept, supervision, analysed data, writing of manuscript, funding sources.

Acknowledgements

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cmi.2021.05.035.

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;21:30413.
[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;23:382–98.
[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;42:694–5.
[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;23:845–8.
[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: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;13:123.
[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: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;40:3053–6.
[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;135:15–21.
[13] Ramarao N, Tran SL, Marin M, Vidic J. Advanced methods for detection of Bacillus cereus and its pathogenic factors. Sensors (Basel) 2020;20:2667.
[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;12:593.
[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;232:189–95.
[20] Hoton FM, Andrup L, Swiecicka I, Mahillon J. The cereulide genetic determinants of emetic Bacillus cereus are plasmid-borne. Microbiology (Reading) 2005;151:2121–4.
[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. Int J Mol Sci 2021;22:5079.
[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.