

1 **Lineage-specific plasmid acquisition and the evolution of specialized**
2 **pathogens in *Bacillus thuringiensis* and the *Bacillus cereus* group**

3
4 Guillaume Méric¹, Leonardos Mageiros², Ben Pascoe^{1,3}, Dan J. Woodcock⁴, Evangelos
5 Mourkas¹, Sarah Lamble⁵, Rory Bowden⁵, Keith A. Jolley⁶, Ben Raymond^{7,8*}, Samuel K.
6 Sheppard^{1,3,6*}

7
8 ¹The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath;
9 ²Swansea University Medical School, Institute of Life Science, Swansea; ³MRC CLIMB Consortium,
10 University of Bath; ⁴Mathematics Institute and Zeeman Institute for Systems Biology and Infectious
11 Epidemiology Research, University of Warwick, Coventry; ⁵Wellcome Trust Centre for Human Genetics,
12 University of Oxford, Oxford; ⁶Department of Zoology, University of Oxford, Oxford; ⁷Department of Life
13 Sciences, Faculty of Natural Sciences, Imperial College London, Ascot; ⁸Department of Biosciences, University
14 of Exeter, Exeter, United Kingdom.

15
16 *Corresponding authors: Samuel K. Sheppard; s.k.sheppard@bath.ac.uk; Ben Raymond;
17 B.Raymond@exeter.ac.uk.

18
19 Keywords: *Bacillus cereus*, *Bacillus thuringiensis*, pangenome, mobile genetic elements,
20 insecticidal toxins

21
22 Running title: Pangenome and plasmids within *B. cereus* group
23

24 **Abstract**

25

26 Bacterial plasmids have roles that range from large secondary chromosomes to small selfish
27 genetic elements. Distinct, but not necessarily mutually exclusive theories have been
28 proposed to resolve plasmid bacteria relationships: plasmids may facilitate evolutionary
29 novelty and maintain beneficial genes via hitchhiking, while plasmid mobility may be
30 opposed by coevolutionary relationships with chromosomes or encouraged via the infectious
31 sharing of genes encoding public goods. Here, we sought to explore a range of these
32 hypotheses through a large-scale examination of the association between plasmids and
33 genomes in the phenotypically diverse *Bacillus cereus* group. This complex group is rich in
34 plasmids, many of which encode essential virulence factors (Cry toxins) that are known
35 public goods. We aimed to characterize population genomic structure, examine the dynamics
36 of plasmid distribution and gene content and the role of mobile elements in diversification..
37 We analysed coding sequence within the core and accessory genome of 190 *B. cereus* group
38 isolates, including 23 novel sequences, including plasmid genes from a reference collection
39 of 410 plasmid genomes. While *cry* genes were widely distributed, those with invertebrate
40 toxicity were predominantly associated with one sequence cluster (clade 2) and
41 phenotypically defined *Bacillus thuringiensis*. Cry toxin plasmids in clade 2 showed evidence
42 of recent horizontal transfer and dynamic gene content, a pattern of plasmid segregation
43 consistent with transfer during infectious cooperation. Nevertheless, comparison between
44 clades suggests that coevolutionary interactions may drive association of plasmids and
45 chromosomes and limit wider transfer of key virulence traits. Proliferation of successful
46 plasmid and chromosome combinations is a feature of specialized pathogens with
47 characteristic niches (*Bacillus anthracis*, *B. thuringiensis*) and has occurred multiple times in
48 the *B. cereus* group.

49

50

51

52

53

54

55

56

57

58
59
60
61
62
63
64

65 The *Bacillus cereus* group includes phenotypically diverse pathogens, including the agents of
66 anthrax and lethal food poisoning, and *Bacillus thuringiensis*, an important biopesticide and
67 source of insecticidal Cry toxins. The taxonomy of the group is complex, partly because of
68 the variety of plasmids, many of which encode essential virulence factors. In this study we
69 sought to better characterize the population genomic structure of this group and the role of
70 mobile elements in evolution and diversification. We analysed coding sequence within the
71 core and accessory genome of 190 *B. cereus* group isolates, including 23 novel sequences
72 from phenotypically-confirmed *B. thuringiensis* and 10 purified plasmids. The distribution of
73 Cry toxins and population structure and gene content of Cry-positive and Cry-negative
74 isolates was examined, including plasmid genes from a reference collection of 410 plasmid
75 genomes. While *cry* genes were widely distributed, those with clear invertebrate toxicity were
76 predominantly associated with one sequence cluster (clade 2) and phenotypically defined *B.*
77 *thuringiensis*. Cry-positive isolates had reduced core genome allelic diversity, but a higher
78 diversity of plasmid genes. *B. thuringiensis* isolates in clade 2 also contained clade-specific
79 plasmids. Cry toxin plasmids in clade 2 showed evidence of recent horizontal transfer and
80 dynamic gene content, a pattern of plasmid segregation consistent with transfer during
81 infectious cooperation. Nevertheless, comparison between clades suggests that
82 coevolutionary interactions may drive association of plasmids with changeable gene content
83 and host chromosomes and limit wider transfer of key virulence traits. Proliferation of
84 successful plasmid and chromosome combinations is a feature of specialized pathogens with
85 characteristic niches (*B. anthracis*, *B. thuringiensis*) and has occurred multiple times in the *B.*
86 *cereus* group.

87

88 **Introduction**

89

90 A recurring feature of the genome organization of many pathogenic bacteria is that important
91 virulence factors are often encoded on horizontally mobile genetic elements (MGEs) (Hacker
92 and Carniel 2001, Sansonetti et al 1981, Smith 2001). A simplistic argument for the location
93 of the genes in the ‘accessory genome’ is that the products they encode are beneficial
94 periodically, as might be the case for opportunistic pathogens with facultative environment
95 niches (Eberhard 1990). However, theory indicates that if genes are beneficial overall, then
96 selection will favour transfer of these genes onto the chromosome to avoid the costs of
97 plasmid carriage (Bergstrom et al 2000). Moreover, many pathogenic bacteria carry essential
98 virulence genes on plasmids, sometimes even when there is ecological and genomic evidence
99 indicating that they are obligate pathogens or compete and replicate poorly in the
100 environment (Hugh-Jones and Blackburn 2009, Keim et al 2009, Yang 2005, Yara et al
101 1997).

102

103 There are several competing, although not necessarily mutually exclusive hypotheses, that
104 explain why some genes are carried on mobile elements and why bacterial virulence factors,
105 in particular, tend to be mobile. This includes the theory that hot spots for recombination
106 occur on the accessory genome. Non-homologous recombination in the accessory genome
107 may have less costly consequences for overall fitness of the cell and there is widespread
108 evidence of substantial recombination in the evolution of bacterial virulence genes (de Maagd
109 et al 2003, Lawrence 2005). Furthermore, genes may be able to persist in plasmids through
110 hitch-hiking with beneficial genes or alleles ensuring that plasmids are maintained by
111 periodically rising to high frequencies via selection on these traits (Bergstrom et al 2000).
112 Both the recombination and hitch-hiking theories may be pertinent for pathogenic bacteria,
113 which are expected to be subject to intense and ongoing selection pressure via host parasite
114 coevolution (Lawrence 2005). Another explanation would be that plasmid genes are present
115 in generally higher copy numbers than chromosomal genes, which may result in the
116 persistence of fitness-enhancing genes that would be beneficial during highly selective
117 events. This has been demonstrated to some extent for antibiotic resistance genes carried on
118 plasmids (Huang et al 2013, San Millan et al 2015).

119

120 One theory that explains the particular mobility of bacterial virulence genes is ‘infectious
121 cooperation’. Many bacterial virulence factors are secreted, and costly. Secreted virulence

122 factors can be exploited by social ‘cheaters’ that fail to invest in virulence, and these cheaters
123 can outcompete more virulent producers within hosts (West et al 2007). Infecting cheating
124 bacteria with plasmids or MGEs carrying virulence genes can convert cheaters to co-
125 operators, a process that can theoretically improve transmission and alter population structure
126 to favour cooperative virulence (Rankin et al 2010, Smith 2001). Synthetic experiments
127 (Dimitriu et al 2014) and the recent evolutionary origin of genes for secreted products
128 provides some support for this theory (Nogueira et al 2009) and major classes of virulence
129 factors can be cooperative public goods, including Cry toxins, quorum-sensing signals and
130 quorum regulated virulence factors in the *Bacillus cereus* group (Deng et al 2015, Raymond
131 et al 2012, Zhou et al 2014).

132

133 The *B. cereus* group has adapted and radiated to exploit environmental niches and a
134 taxonomically broad array of hosts to an extent that can be matched by few known pathogens
135 (Raymond and Bonsall 2013). Hosts for *B. cereus sensu stricto* (*Bc*), *B. thuringiensis* (*Bt*) and
136 *B. anthracis* (*Ba*) include vertebrates, insects and nematodes (Raymond et al 2010a,
137 Raymond and Bonsall 2013, Ruan et al 2015, Turnbull 2002), while plants have been
138 implicated as vectors of entomopathogenic strains (Raymond et al 2010b). This adaptive
139 radiation means that this group is of broad significance, containing strains important for
140 insect pest management, food production and human health. This provides an opportunity for
141 studying how ecology in diverse pathogenic niches shapes bacterial genomes, especially as a
142 large number of *B. cereus* genotypes are associated with well characterized environmental
143 and host niches (Guinebretière et al 2008, Guinebretière et al 2010, Raymond et al 2010b,
144 Raymond and Bonsall 2013). Importantly several characteristic and essential virulence
145 factors are encoded on plasmids in *B. cereus sensu lato*, a group which includes *B. cereus*
146 *sensu stricto* (*Bc*), *Bt*, *Ba*, and collectively referred to as the *B. cereus* group (Gonzalez et al
147 1982, Okinaka et al 1999).

148

149 Within the *B. cereus* group, the species designation *Bt* is defined by the possession of
150 proteinaceous inclusion bodies, mainly formed of the essential virulence factors known as
151 Cry (Crystal) toxins. These are large, pore-forming proteins that enable orally ingested
152 bacteria to invade the invertebrate haemolymph from the midgut (Schnepf et al 1998). These
153 toxins cause paralysis and are lethal at high doses, but are relatively host specific and have no
154 known toxicity to vertebrates, hence their widespread incorporation into genetically modified
155 insect resistant crops (Bravo et al 2011). The *B. cereus* group possesses a rich diversity of

156 accessory genome elements with numerous large conjugative plasmids (Hu et al 2009b, Van
157 der Auwera and Mahillon 2008, Zheng et al 2013). *B. cereus* group isolates can contain a
158 large number of plasmids, and this plasmid complement can vary substantially both within
159 and between serotypes (Hu et al 2009a, Reyes-Ramirez and Ibarra 2008), indicating that the
160 accessory genome has the potential to respond rapidly to ecological change.

161

162 Defining a species based on the possession of horizontally mobile *cry* genes is problematic.
163 Unsurprisingly, *Bt* is not a monophyletic group and several divergent clades defined by multi-
164 locus sequence typing (MLST) or genomic data contain *Bt* isolates with Cry inclusions
165 (Raymond et al 2010b). The taxonomy of the *B. cereus* group, and of *Bt* within it, is
166 controversial; while accurate and informative species delineation has important economic
167 implications (EFSA 2016, Raymond and Federici 2017). The licensing and ‘safe’ status of *Bt*
168 as a biological control agent that can be applied to vegetable crops, is partly dependent on its
169 biological distinctiveness from human pathogenic *Bc* and *Ba*. Although *Bt*-based products are
170 considered to be among the safest insecticides on the market (Federici and Siegel 2007,
171 Siegel 2001), this reputation can be damaged by uncertain taxonomy and lack of rigour in
172 interpreting epidemiological evidence (Raymond and Federici 2017). Moreover, the possible
173 horizontal mobility of virulence factors from vertebrate pathogens within the *B. cereus* group
174 to invertebrate pest control agents also has potential safety implications for the use of *Bt* in
175 biocontrol (EFSA 2016).

176

177 The aims of this study are three-fold. First, to use a revised pan-genomic analysis to assess
178 the phylogenetic status of the *B. cereus* group. Second, to explore the mobility of key
179 virulence gene and virulence plasmids across the group. Third, to assess if patterns of
180 plasmid/chromosome association in this group are consistent with current evolutionary
181 ecology theory for plasmids and plasmid gene content.

182

183 **Methods**

184

185 **Isolate sampling and plasmid extraction**

186 *Bt* isolates with diverse host toxicity were chosen for whole genome sequencing (WGS) and
187 plasmid purification. These included isolates available from the *Bacillus* Genetic Stock
188 Centre (BGSC) the Agricultural Research Service (NRRL) culture collection, supplemented
189 with isolates sampled for this study. Prior to sequencing, the identity of isolates with Cry

190 inclusions was confirmed by light microscopy of sporulated cultures and cross-checked by
191 Sanger sequencing of flagellin genes (*Bthag*, *fliC*) using primers and conditions described in
192 Xu and Cote (2006) and BLAST (Altschul et al 1990) searches of at least 500bp of both
193 genes against the *nr* database from NCBI. One isolate, *Bt* serovar *brasiliensis* BGSC 4AY1,
194 was excluded because production of Cry inclusions could not be confirmed. Plasmid
195 extractions used High Speed midi kits (Qiagen) with 200ml of bacterial culture and
196 subsequent digestion with plasmid-safe ATP-dependent exonuclease (Epicentre) to remove
197 linear DNA fragments, both as per manufacturer's directions.

198

199 **Genome sequencing**

200 A total of 190 *Bacillus* group genomes were used, including 23 *Bt* isolates that were
201 sequenced as part of this study (**Table S1**). Plasmid and chromosomal DNA were extracted
202 using the QIAamp DNA Mini Kit (QIAGEN, Crawley, UK), using manufacturer's
203 instructions. DNA was quantified using the Quant-iT DNA Assay Kit (Life Technologies,
204 Paisley, UK) and a Nanodrop spectrophotometer before sequencing using an Illumina HiSeq
205 2500 analyzer (Illumina, San Diego, CA, USA). 100bp short read paired-end data was
206 assembled using the Velvet version 1.2.08 *de novo* assembly algorithm (Zerbino and Birney
207 2008), incorporating the VelvetOptimiser protocol (version 2.2.4)
208 (<https://github.com/dzerbino/velvet>) for all odd k-mer values from 21 to 99. Scaffolding was
209 disabled and the minimum output contiguous sequence assembly setting was 200bp. The
210 average number of contiguous sequences for the 23 isolates and 10 plasmid extractions
211 sequenced from this study was 407 ± 225 and 85 ± 32 respectively. The average assembly
212 sequence length was $6,162,692 \pm 348,490$ bp for isolate whole genomes and $475,459 \pm 395,713$
213 bp for plasmid extractions (**Table S2**). This is consistent with published estimates of the
214 genome size of members of the *B. cereus* group. Isolates sequenced in this study were
215 augmented with 182 genomes from public databases (available in April 2013), including
216 reference genomes from *Bt* strain YBT020 (Zhu et al 2011), *Bc* strain ATCC 14579 (Ivanova
217 et al 2003), and *Ba* strain Ames (Read et al 2003) to give a total of 190 isolate genomes.
218 Metadata for published isolate genomes was variable and sometimes lacked detailed
219 sampling information, but these genomes were included to provide as much information as
220 possible on the genomic diversity within the *Bacillus* group (**Table S1**). Functional
221 predictions were made using the WebMGA COG server using rpsblast 2.2.15 on the NCBI
222 COG database (<http://weizhong-lab.ucsd.edu/metagenomic-analysis/server/cog/>). Sequence

223 type (ST) assignment from assembled genome sequences was performed using the mlst
224 software (<https://github.com/tseemann/mlst>).

225

226 **Creation of a reference pan-genome from bacterial genomes**

227 As in recent publications on other species (Meric et al 2016, Monteil et al 2016, Morley et al
228 2015, Murray et al 2017, Yahara et al 2017), a reference pan-genome approach was used with
229 gene-by-gene alignment, consistent with whole genome MLST (Jolley and Maiden 2010,
230 Maiden et al 2013, Meric et al 2014, Sheppard et al 2012), implemented in BIGSdb open
231 source software. Briefly, the reference pan-genome was constructed by combining the
232 genomes of several reference strains (*Bt* strain YBT020 (Zhu et al 2011), *Bc* strain ATCC
233 14579 (Ivanova et al 2003), and *Ba* strain Ames (Read et al 2003)) with whole-genome
234 annotations from all the other genomes of this study to derive a single gene list. To achieve
235 this, all assembled genomes were submitted to the online automatic annotation pipeline
236 RAST (Aziz et al 2008). Rapid annotations of bacterial genomes provided by RAST are
237 accurately relying on the curated database system SEED, in which novel annotations are
238 provided directly by the annotations from the RAST user community (Overbeek et al 2014).
239 Allelic variants of unique genes were identified as duplicates, found in more than one isolate,
240 and were removed to create the reference pan-genome of the whole dataset. Gene homology
241 was defined using BLAST, with those found to have >70% nucleotide identity over >10% of
242 the sequence length, considered to be homologous. This conservative sequence length
243 threshold to distinguish genes from their allelic variants was deliberately set lower than the
244 threshold of >50% sequence length commonly used to identify gene presence/absence, false
245 negatives being considered less problematic than false positives in terms of characterising
246 pan-genomes. Indeed, from a purely quantitative perspective, overestimating the size of the
247 core genome is potentially equally as bad as underestimating it. However, from a
248 methodological point of view, when defining the pan-genome, the rigorous elimination of
249 duplicates reduces the number of potential BLAST gene mismatches for each draft genome
250 that is compared to the reference pan-genome list. In real terms, this leads to more accurate
251 quantification of the total genome size based on coding sequences. Furthermore,
252 overestimating alleles by considering them as distinct genes is particularly problematic for
253 downstream analyses where putative gene function is investigated. For example, bias could
254 be introduced into broad analyses of COG/KEGG functional groups and more detailed
255 analyses of individual gene function would be confounded by an inflated number of paralogs.
256 The resulting *B. cereus* reference pan-genome was based upon all genomes listed in **Table**

257 **S1**, which included isolates from *Bc*, *Bt* and *Ba* within the *B. cereus* group. The total number
258 of unique genes in the pan-genome from all these isolates was 27,016.

259

260 **Creation of a reference plasmid gene list from 410 reference plasmid sequences**

261 Discriminating plasmid genes from chromosomal genes is challenging using the data from
262 high-throughput short read sequencing that typically use total genomic DNA as a sample.
263 This is because the reads are assembled and therefore do not produce a single read for each
264 amplicon. To account for this we conducted purification and separation of chromosomal and
265 plasmid DNA prior to sequencing for 10 isolates, resulting in the sequencing of 10 plasmid
266 sequences (**Table S3**). Additionally, diversity and possible genomic rearrangements among
267 plasmids, or even their possible chromosomal integration, make gene prediction difficult
268 without informed comparative approach to a curated reference database of known plasmid
269 genes.

270

271 We assembled a collection from 410 full plasmid genomes, most of which were all plasmids
272 available from NCBI in September 2016 (**Table S3**), and were assigned as having been
273 isolated from one of the 3 “species” of the *B. cereus* group based on the presumptive typed
274 identity of the corresponding host bacteria (**Table S3**). Briefly, the collection comprised 81
275 plasmids attributed to *Bc*, 249 to *Bt* and 87 to *Ba* (consisting only of variants of pXO1 and
276 pXO2) (**Table S3**). All automatically annotated genes were assembled in a single reference
277 gene list, without any filtering of allelic variants. Indeed, mega-plasmids, consisting of
278 assemblages of various otherwise described plasmids, have been described in the *B. cereus*
279 group (Zheng et al 2013). Therefore, all genes from all plasmids were kept in the reference
280 list to investigate whether the sequence of certain plasmids was distributed differentially
281 among the isolates. By not filtering for allelic variants, we did not create a plasmid
282 pangenome list of unique genes but maintained the plasmid sequence integrity for each
283 plasmid, making observation of rearrangements possible, as well as being able to assess the
284 prevalence of particular plasmid genes in given isolates. The plasmid gene list comprised
285 48768 genes, some of which represented allelic variants of the same gene, for example,
286 origins of replications, conjugation proteins and other members of the “core” plasmid
287 genomes.

288

289 **Core and accessory genome variation, and predicted insecticidal toxin detection**

290 All isolate genomes were compared to the reference pangenome list with a locus match
291 defined with the BLAST parameters for a positive match being >70% nucleotide identity
292 over >50% of the sequence length (Jolley and Maiden 2010, Meric et al 2014, Sheppard et al
293 2012). This whole genome MLST approach produced a matrix of gene presence/absence with
294 different allele numbers assigned to all genes based upon nucleotide identity, as previously
295 described (Meric et al 2014, Meric et al 2015). The prevalence of plasmid genes, inferred
296 from an assembled list of all genes present in 410 plasmids from NCBI, in 190 bacterial
297 genomes was determined using BLAST as above.

298

299 Genes encoding *Bt* toxins (Cry, Cyt, Vip and Sip) were predicted via BtToxin_scanner, a tool
300 designed to identify new candidate toxin genes from sequence data using three different kinds
301 of prediction methods (Ye et al 2012). This approach identified sets of candidate toxin genes
302 in a complementary approach to the RAST/SEED pipeline presented above. Briefly,
303 BtToxin_scanner specifically addresses challenges set by the detection of *Bt* toxins by
304 combining a BLAST approach with additional hidden Markov model (HMM) and a support
305 vector machine (SVM) approaches to accurately predict the presence of toxin genes and
306 annotate them (Ye et al 2012). While the RAST/SEED approach is well-suited for bacterial
307 whole genomes, care was taken for *Bt* toxins due to specific challenges such as repeats and
308 low-homology between members of the toxin families (Ye et al 2012). These are addressed
309 by BtToxin_scanner, that specifically predicts and annotates *Bt* toxins, either as previously
310 known variants or novel candidate unknown toxins (Ye et al 2012). To examine candidate
311 genes as potential novel toxins or false positives, we proceeded as follows: low-homology
312 BtToxin_scanner hits of less than 45% amino-acid sequence homology were considered good
313 candidates, as previously described (Noguera and Ibarra 2010), and were used a queries in
314 protein-protein BLAST against the database of non-redundant proteins (nr) on NCBI on the
315 27/06/2017. When a hit in the first 50 was found to have a match with an entry annotated
316 as Cry or more generally any reference to predicted insecticidal activity, the hit was
317 considered a good candidate insecticidal toxin. When no obvious predicted insecticidal-
318 related annotated hit was found, the protein was considered a false positive (**Table S4**).
319

320 **Allelic diversity calculations**

321 To avoid sampling bias, the number of unique alleles per isolate was calculated for randomly
322 selected isolates. Briefly, for comparisons involving *Ba* (**Figure 3A** and **3B**), for which only
323 17 isolates are included in our study, the number of unique alleles was determined for 17

324 randomly-selected Cry-positive isolates and 17 randomly-selected Cry-negative isolates. This
325 step was repeated 50 times and the 50 values for each group were averaged to give the final
326 value of unique alleles per isolate in the two groups. For comparisons not involving *Ba*
327 (**Figure 3C, 3D and 3E**), the number of unique alleles was determined for 50 randomly-
328 selected Cry-positive isolates and 50 randomly-selected Cry-negative isolates. This step was
329 repeated 50 times and the 50 values for each group were averaged to give the final value of
330 unique alleles per isolate in the two groups.

331

332 **Phylogenetic and clustering analyses**

333 Phylogenetic trees were constructed based on 2274 core genes shared by all genomes in our
334 dataset, which were individually aligned using MAFFT (Katoh and Standley 2013) and
335 concatenated to produce contiguous sequence alignments in BIGSdb (Jolley and Maiden
336 2010). RAxML (Stamatakis 2014) was used to reconstruct phylogenies using default
337 parameters. Clustering of plasmid prevalence profiles was performed using the web-based
338 platform WebGimm (Joshi et al 2011) using the Context Specific Infinite Mixture Model
339 (Freudenberg et al 2010).

340

341 **Results**

342

343 **Phylogeny of *B. cereus* group isolates**

344 To examine the phylogenetic relationships between isolates from our dataset, we attributed
345 sequence types (STs) to each genome using the *B. cereus* MLST scheme on pubMLST
346 (<https://pubmlst.org/bcereus/>) and recreated a phylogenetic tree using RAxML (Stamatakis
347 2014). STs could not be assigned for 5 (2.6%) isolates because MLST loci were incomplete or
348 truncated in the draft genomes. There was considerable diversity among the typable isolates
349 with a total of XXX STs including 24 newly identified among isolates in this study. A total of
350 48 different STs were found in Cry-positive and candidate Cry toxin-harbouring isolates and
351 63 different STs in Cry-negative isolates (**Table S1, Table 1**). Only 2 different STs (ST-1
352 and ST-3) were detected in the *Ba* lineage (**Table S1, Table 1**), which is consistent with the
353 reported clonal nature of the population (Van Ert et al 2007). Interestingly, 8 STs (ST-8, ST-
354 56, ST-111, ST-223, ST-257, ST-506, ST-783 and ST-934) were shared by Cry-positive and
355 Cry-negative isolates, highlighting the acquisition of mobile virulence factors in divergent
356 genetic backgrounds (**Table S1**). A total of 14 genomes from our dataset, all initially
357 classified as *Bc*, clustered in a Clade 3 lineage with *B. mycoides* and *B. weihenstephanensis*

358 isolates (**Table S1**, **Figure S1**). Two of these isolates were Cry-positive, 7 were predicted to
359 harbour candidate novel Cry toxins and no toxin gene could be detected in 5 genomes (**Table**
360 **S1**).

361
362 A phylogenetic tree was generated from the concatenation of gene-by-gene alignments
363 (Sheppard et al 2012) of 2274 core genes found to be shared in all genomes (**Figure 1A**).
364 Most isolates clustered in Clade 2 (71/190; 37.4%), which also had the highest prevalence of
365 Cry-positive and candidate Cry-harbouring isolates (48/71; 67.6%) (**Figure 1A**, **Table S1**).
366 Clades 3 and 4 had comparable prevalence of Cry-positive and candidate Cry-harbouring
367 isolates (18/33; 54.5% and 7/17; 41.2%, respectively) while Clade 1, comprising *Ba* isolates,
368 had only 8/57 (14.0%) Cry-positive and candidate Cry-harbouring isolates (**Figure 1A**, **Table**
369 **S1**). Three isolates were not clustered in any of the MLST-defined clades, with isolate *Bc*
370 R309803 (ST-74) being a singleton, and Cry-harbouring *Bc* BAG2X1-1 (ST-723) and Cry-
371 negative *Bc* BAG2X1-3 clustered together between Clade 3 and 4 (**Figure 1A**).
372

373 **Detection of candidate novel insecticidal toxins in 190 *B. cereus* group genomes**

374 We used the BtToxin_scanner software (Ye et al 2012) to detect the presence of genes
375 encoding the δ-endotoxins Cry and Cyt, and genes encoding the secreted toxins Vip and Sip
376 in the whole genome sequences of 190 *Bc*, *Bt* and *Ba* isolates, including 23 new,
377 phenotypically confirmed, *Bt* isolates. In total, the dataset comprised 135 isolates identified
378 as *Bc*, 38 as *Bt* and 17 from the *Ba* lineage. Apart from *Ba*, the species nomenclature was
379 mostly inferred from records in the genome public repository and may include strains that are
380 mistyped, notably for the genomes labelled as *Bc*. Predicted insecticidal toxins and predicted
381 novel candidate Cry toxins were distributed differentially across these species designations
382 and among previously defined clades (Raymond et al 2010b), (**Table S1**, **Figure 1A**). Cry,
383 Cyt and Vip toxin genes, as well as uncharacterised candidate Cry toxins were detected in
384 84/190 (44.2%) isolates from our dataset (**Table 1**), including 36/38 (94.7%) classified as *Bt*,
385 in 48/135 (35.5%) *Bc* but never in *Ba*. Notably, only 12/135 (8.8%) *Bc* isolates harboured
386 previously-known toxin genes, while 36/135 (26.6%) harboured only uncharacterised
387 candidate toxin genes (**Table S1**). The fact that 2 strains of *Bt* (*Bt* *subsp. pondicheriensis*
388 BGSC4BA1 and *Bt* *subsp. malaysiensis* NRRL_B23152) seemed to harbour no toxin could
389 be due to a misclassification, but also to incomplete genomes or the presence of new toxin
390 variants possibly not detected by our protocol. The most common Cry/Vip protein variants
391 were Cry1Ia2, Cry2Aa9, Cry2Ab3 and Vip3A detected in 6 isolates, each time a combination

392 of *Bt* and *Bc* (**Table S4**). Parasporsins (Cry toxins with activity against cancer cells but not
393 invertebrates) from a range of classes were detected in 10/190 genomes (**Table S1**), while
394 many candidate Cry proteins had parasporsins as the closest match (**Table S4**). Notably, while
395 candidate *Cry* proteins were widely distributed across the group, those with clear invertebrate
396 toxicity, especially to *Diptera* and *Lepidoptera* were concentrated in clade 2 (**Figure 1B**).
397 Moreover, the host taxon targeted by the Cry toxin complement in all isolates were readily
398 identified (we identified a single generalist genome) and were typically associated with either
399 an insect Order or nematodes, consistent with specialization on a group of hosts (**Figure 1B**;
400 **Table S1**).

401
402 In 35/50 (70%) Cry/Vip-positive isolates harbouring known characterised variants, several
403 distinct toxin genes were detected in the same genome by BtToxin_scanner (**Table S1, Table**
404 **S4**). This was most common in *Bt* isolates, with an average of around 5 (4.71 ± 3.7 ; n=35)
405 toxins detected per toxin-positive genome, with 4 isolates predicted to harbour more than 10
406 detected variants (*Bt* subsp. *morrisoni* strain BGSC_4K1, sequenced as part of this study, had
407 a maximum of 14 detected toxin variants in its genome). *Bc* isolates also putatively harboured
408 several toxins, with between 2 and 3 in average per toxin-positive genome (2.46 ± 1.80 ; n=15),
409 and 4 strains with 5 detected variants. In contrast, 15/50 (30%) isolates in total seem to only
410 harbour one known characterised toxin variant in their genomes. For the remaining analyses
411 of this study, we considered *B. cereus* group isolates to have a possible insecticidal activity
412 based on the detection of known or candidate toxins rather than their assigned species in
413 genome databases.

414
415 **Pangenome variation and diversity across *B. cereus* group isolates**
416 We then performed a complete dataset-wide pangenome analysis in which the presence and
417 variation of every automatically annotated gene from every genome was examined. Gene
418 prevalence differences were compared between various groups of isolates to examine core
419 and accessory gene variation. An average of 6018 (±339) genes were detected from 190 *B.*
420 *cereus* group genomes from our dataset. A total of 2274 core genes were found to be present
421 in all genomes, which represents 37.8% of the average number of genes in a *B. cereus*
422 genome. Interestingly, the average amount of genes detected in Cry-positive including Cry-
423 candidate harbouring isolates was always observed to be larger than in Cry-negative isolates
424 (**Figure 2A**), and this difference was significant in Clade 1 (1-way ANOVA with Sidak's
425 multiple comparison tests, $t=3.998$, d.f.=175; adjusted $p=0.0005$) and Clade 2 ($t=6.710$,

426 d.f.=175; adjusted $p<0.0001$). One explanation for this observation could be the generally
427 higher prevalence of large plasmids in Cry-positive isolates.

428
429 Quantitative analysis of the prevalence of genes revealed that no genes were shared
430 specifically by all Cry-positive or by all Cry-negative isolates. Cry toxins are a family of
431 proteins rather than isoform variants of the same protein encoded by the same genes/alleles.
432 This may explain why no genes were shared by all isolates. A total of 6225 genes were found
433 only in Cry-positive (but not shared by all isolates) and not in Cry-negative isolates.
434 However, 6172 of these were found at very low prevalence ($n<10$ isolates), which left 53
435 genes present in >10 isolates (**Table S5**).
436

437 The dearth of genes shared at high prevalence between Cry-positive isolates from all clades is
438 related to the polyphyletic distribution of *cry* genes. This may be indicative of both the
439 diversity of structure and gene content among MGEs conferring insecticidal virulence in *B.*
440 *cereus* group isolates (see below), and clade-specific insecticidal virulence associated with
441 specific virulence factors. Nevertheless, some genes had increased prevalence in one group or
442 the other, but this was predominantly caused by the fact that clade 2 is more significantly
443 enriched for cry-positive isolates than any other clade. When clade-specific genes were
444 examined, we found that only 21 “clade-specific core genes” that were shared by all isolates
445 from specific clades but absent from any other clade (1 in clade 2, 8 in clade 3 and 12 in
446 clade 4) (**Table S6**). Genes specific and shared by every isolate from clade 4 notably encoded
447 a choline-binding protein A (CpbA), which has been shown to be an adhesion factor in
448 Firmicutes such as *Streptococcus pneumoniae* (Luo et al 2005) and which has been used for
449 vaccine development (Bologa et al 2012) (**Table S6**). Genes specific and shared by every
450 isolate from clade 3 included genes encoding an uncharacterised transport system as well as
451 genes involved in sporulation and respiration (**Table S6**). Notably, one gene (BC4305,
452 annotated as hypothetical protein) was shared by all 71 isolates from clade 2 and absent from
453 any other clade (**Table S6**). This gene is not located in any predicted operon in the *Bc*
454 ATCC14579 genome, nor is it flanked by genes of known function. A total of 30 genes were
455 found to be shared by all *Ba* and absent in the rest of the dataset, including in non-*anthracis*
456 clade 1 isolates (**Table S6**), while no genes were found to be present in all non-*anthracis*
457 clade 1 isolates but absent in *Ba* isolates, confirming previous analyses which indicated that
458 there are few large scale genomic variations that differentiate *Ba* from closely-related *Bc*
459 (Zwick et al 2012). It is interesting to note that 50% of the pangenome (13,501 genes)

460 comprised low-frequency genes that were each present in only less than 4 isolates, which
461 highlights the variability of the *Bacillus cereus* genome and is potentially related to
462 horizontal gene transfer in this species.

463

464 The comparison of functional prediction prevalence for different groups of genes showed that
465 the distribution of functional categories of accessory and plasmid-borne genes were generally
466 similar, and differed from the core genome (**Figure S4**). More specifically, accessory and
467 plasmid genes were significantly enriched in prevalence from COG class L (Replication,
468 recombination and repair) than in the core genome (Tukey multiple comparisons tests after a
469 two-way ANOVA; adjusted $p=0.0062$ and $p=0.0021$ respectively). Generally speaking,
470 although not significantly different using a stringent statistical test, there were much lower
471 proportions in metabolism-associated genes (COG classes E, P and C) in accessory and
472 plasmid genes than in the core genome (**Figure S4**).

473

474 **Lower core genome allelic diversity among Cry-positive isolates**

475 We examined the allelic diversity of various groups of isolates by calculating the number of
476 unique alleles per isolate for Cry-positive (including Cry-candidate harbouring), Cry-negative
477 and *Ba* isolates (**Figure 3**). We observed that these three groups had distinct distributions of
478 allelic diversity in their core genomes (Kruskal-Wallis test with Dunn's multiple comparisons
479 test; adjusted $p<0.0001$ for each pairwise comparison of rank differences) (**Figure 3A**). *Ba*
480 had much lower diversity, as expected from its clonal structure within clade 1 on the *B.*
481 *cereus* group phylogenetic tree (**Figure 1**). Interestingly, Cry-positive isolates had a
482 significantly lower core genome allelic diversity than Cry-negative isolates (**Figure 3AB**).
483 This was also observed when the allelic diversity of each of the 2274 core genes of Cry-
484 negative isolates was plotted against the allelic diversity of the same gene in Cry-positive
485 isolates (**Figure 3C**). Only 225/2274 (9.9%) core genes had a higher allelic diversity in Cry-
486 positive isolates, which was visualised by the circles below the proportionality line in **Figure**
487 **3C**. When we repeated this analysis at the clade-level, clade 2 (with the highest prevalence of
488 genomes harbouring predicted-insecticidal toxins) also had reduced allelic diversity with
489 respect to clades 1 and 3 (**Figure 3DE**). While this approach is sample-dependent, the
490 difference between Cry-positive and Cry-negative isolates in terms of diversity cannot be
491 explained by the clonal frame as the isolates cluster together on the tree. The contrast
492 between high diversity in predicted insecticidal virulence factor families (**Table S1**) and

493 MGEs (as inferred by **Figure 2A**) and lower diversity within the core genome of Cry positive
494 isolates is most likely explained by lateral transfer of these elements (**Figure 3**).
495

496 **Detection of plasmid genes in *B. cereus* group isolates**

497 The previous analysis, consistent with the literature on *Bt* toxins (Gonzalez et al 1982,
498 Mahillon et al 1994), suggests mobility of virulence determinants among Cry-habouring *B.*
499 *cereus* group isolates, via plasmids or transposons. The presence of each of 48768 genes from
500 a reference plasmid list was examined in the 190 genomes of our dataset, and the result
501 summarised in a heatmap (**Figure 4**) and a table (**Table 1**). The total complement of genes
502 corresponding to a particular plasmid, were detected in at least one isolate genome for 53%
503 (220/410) of reference plasmids. These included pXO1 and pXO2, but also some plasmids
504 identified in *Bt*. Our plasmid detection was consistent with previous reports of atypical
505 strains. One *B. cereus* isolate (strain G9241) was observed to carry a full *Ba* plasmid pXO1,
506 and has been described before (Wilson et al 2011). Additionally, a *Ba* strain (CDC684) was
507 found to be missing pXO2, and has been described in the literature as having attenuated in
508 virulence (Okinaka et al 2011) while another one (strain A1055), missing pXO1, has been
509 reported as atypical (Antonation et al 2016). For 37.6% (154/410) of plasmids, between 0%
510 and 90% of reference genes were detected in at least one isolate and only 2 of them (pCTC
511 and pMC8, originally purified from *Bt* isolates) had no genes present in our dataset. At least
512 one gene from 119 plasmids was found in all 190 genomes used in this study. These included
513 variants of the pXO2 *Ba* virulence plasmid, implying that genes from this plasmid are present
514 in the genome of the species, either as a result of: (i) homology with chromosomal core
515 genes, or chromosomally-integrated plasmid genes (Zheng et al 2015); or (ii) homology with
516 a widespread “plasmid core genome”. While more reference plasmid sequences are necessary
517 to describe the full diversity within our dataset our results are consistent with the wide
518 distribution of plasmids in the *B. cereus* group, potentially with every genome containing
519 plasmid genes. Additionally, there were a large number of plasmids initially attributed to *Bt*
520 that were detected in clade 2 Cry-positive isolates (**Figure 6, Table 1**). Interestingly, this did
521 not seem to be the case for Cry-positive isolates from clades 1, 3, 4 and 5, suggesting possible
522 clade specific virulence patterns.
523

524 In this analysis, we considered that the 53 closed genomes included in our dataset would
525 harbour fewer plasmid genes than draft genomes. As closing genomes requires the
526 experimental validation of the order of contigs using methods such as PCR, plasmid genes

not integrated in the chromosome would be present in another amplicon at the time of DNA isolation, and would be excluded from the finished closed sequence. In contrast, draft genomes are produced from the total genomic DNA of bacteria, without discrimination for plasmid or chromosomal origin. To assess the extent of missing information in genomes included in our analysis, we created a gene list including all unique genes from the 410 annotated plasmid genomes used above. A total of 7,248 genes were identified and their presence was recorded in 53 closed genomes (Table S1) and 136 draft genomes, including 23 sequenced as part of this study (**Figure S2**). The 23 genomes generated in this study contained significantly more putative plasmid genes than the 53 previously published closed genomes (**Figure S2**, Mann-Whitney test, $p<0.0001$) and 113 previously published draft genomes (**Figure S2**, Mann-Whitney test, $p=0.0001$) which suggests that our sampling captured a large proportion of plasmid-harbouring isolates. Interestingly, around 400 plasmid genes from our list were detected in the closed genomes (**Figure S2**), consistent with frequent chromosomal integration of plasmids or movement of mobile elements between plasmids and chromosomes within the *B. cereus* group.

Validation analysis was carried out on paired chromosomal and plasmid sequence from isolates where the plasmids had been purified and sequenced separately. A total of 10 plasmids were extracted from isolates present in the genomic dataset of this study (**Figure S3**). Distinct plasmid sequences were not obtained from all isolates. This can be explained by multiple factors, including plasmid chromosomal integration or technical difficulties when isolating very large plasmids from bacteria using methods designed principally for high-copy small plasmids. Indeed, despite methods available (Kado and Liu 1981), obtaining correctly closed genomes of large plasmids remains a methodological challenge (Smalla et al 2015). Nevertheless, our approach allowed plasmid and chromosomal sequence to be accurately discriminated. We observed that 7/10 plasmids were only detected in a single isolate, 4 of which from the isolate they were extracted from (**Figure S3**). This reflects strain specific plasmid acquisition. Two plasmids (pBt407 and pStrain62) were detected in additional single isolates, reflecting the possible, but limited spread of these plasmids in the *Bacillus cereus* group. One plasmid (pBGSC 4J4) was not detected in any isolate, reflecting the absence of the corresponding isolate in our genome dataset. Notably, 3 plasmids from closely related isolates (p71o, pBGSC 4D4 and pBGSC 4D1) were detected in more than 1 isolate, all from the *kurstaki* ST8 group of Clade 2 *Bt* isolates (**Figure S3**). This could reflect an increased

560 spread of these plasmids and related plasmids in this ecological group, consistent with the
561 above observations on a larger plasmid dataset.

562

563 Discussion

564

565 Isolate genomes within the *B. cereus* group show evidence of horizontal gene transfer (HGT),
566 consistent with previous work (Van der Auwera et al 2007, Vilas-Bôas et al 2008) (Didelot et
567 al 2009). Using the current phenotypic definition, *Bt* is recognized as being polyphyletic and
568 since the multiple clades containing *Bt* are comprised of both *Bt* and *Bc*, *Bt* is also
569 paraphyletic (Cardazzo et al 2008, Didelot et al 2009, Priest et al 2004, Raymond et al 2010b,
570 Raymond and Bonsall 2013, Tourasse et al 2011). Unsurprisingly, there are disagreements
571 about the distinctiveness of *Bc* and *Bt*, which are compounded by the practice of applying “*B.*
572 *cereus*” as a catch-all species term when other species-specific taxonomic data are missing.
573 Solutions to these taxonomic inconsistencies have been debated. One view is that the entire
574 *B. cereus* group containing *Bt*, *Bc*, *Ba*, *B. mycoides*, *B. weihenstephanensis* should be treated
575 as one species (Helgason et al 2000, Tourasse et al 2006). Our genomic analysis highlights
576 the inconsistency of *Bc*, *Ba* or *Bt* as species designations based upon phenotype comparisons,
577 particularly for *Bc* and *Bt* that can share aspects of their ecology and do not represent discrete
578 cohesive lineage clusters. However, all subsequent phylogenies of *B. cereus* group isolates,
579 including this work and previous MLST studies, have shown that there are several cohesive
580 genetically distinct clades in the *B. cereus* group (Cardazzo et al 2008, Didelot et al 2009,
581 Guinebretière et al 2008, Priest et al 2004, Raymond et al 2010b, Sorokin et al 2006,
582 Vassileva et al 2006, Vilas-Boas et al 2002). The three major clades originally defined by
583 MLST (*Ba* and relatives – Clade 1, *B. kurstaki* and *Bc* - Clade 2 and *B. weihenstephanensis* -
584 Clade 3) were recovered in this study, although the distribution of predicted insecticidal
585 genes and of isolates identified as *B. weihenstephanensis* and *B. mycoides* indicates that there
586 can be additional significant heterogeneity within these clades (**Figure 1**, **Figure S1**)(Zheng
587 et al 2017).

588

589 In addition, there is abundant evidence for substantial ecological differentiation between
590 clades, either in terms of their ability to colonize plants (Raymond et al 2010b, Vidal-Quist et
591 al 2013); their carriage of virulence factors such as enterotoxins (Cardazzo et al 2008); the
592 risks they pose to vertebrates (Cardazzo et al 2008, Guinebretière et al 2010, Raymond and
593 Bonsall 2013) or their metabolic and growth characteristics (Guinebretière et al 2008).

594 Moreover, analyses of the patterns of HGT indicate that most recombination occurs within,
595 rather than between clades, making these groups something akin to ‘biological species’
596 (Didelot et al 2009). The analysis of the distribution of *cry* genes in this study also suggests
597 real biological differences. Clade 2 is unique in terms of both the high proportion of genomes
598 carrying predicted insecticidal or nematicidal *cry* genes, the large number of insecticidal
599 toxins (Cry and Vip) encoded in each genome, and the presence of a substantial number of
600 isolates with complements of genes conferring virulence to Lepidoptera and Diptera species.

601

602 While acquisition of Cry toxin genes enables bacteria to be pathogenic to invertebrates, it
603 imposes considerable metabolic costs on the cell both in terms of growth rate *in vivo*
604 (Raymond et al 2007, Raymond et al 2012) and the ability to grow or persist in soil (West et
605 al 1985, Yara et al 1997). This high metabolic burden could explain why specialized
606 insecticidal *cry* gene complements are largely restricted to a subset of lineages within Clade
607 2. Reduced allelic diversity in Cry positive lineages could be driven by directional selection
608 on specialized invertebrate pathogen genotypes, or the clonal expansion of successful
609 genotypes. High cost of Cry toxin production, and specialization to invertebrate hosts could
610 explain the excellent safety record of *Bt*-based biopesticides. Despite their close
611 phylogenetic relationship to *Bc* isolates capable of causing diarrhoea (Raymond et al 2010;
612 Raymond & Federici 2017) growth in the vertebrate gut and vegetative production of
613 enterotoxins are required for diarrhoeal food poisoning (Ceuppens et al 2012) and production
614 of Cry toxins is likely to hamper vegetative outgrowth considerably.

615

616 Bacterial ecology is clearly related to carriage of specific *cry* genes but a species definition
617 based on virulence genes, rather than phenotype, offers few advantages. This is partly due to
618 the uncertainties of gene expression but also because of the surprisingly widespread
619 distribution of *cry* genes with no known host affiliation. For example, the parasporins
620 *cry31Aa*, *41Aa*, *42Aa* *46Aa*, *64A*, *65A*, *66A*, which are cytotoxic to a range of cancer cells,
621 were found in 5% of the isolates in this study despite having no known function in infection
622 (Hayakawa et al 2007, van Frankenhuyzen 2009, van Frankenhuyzen 2013, Yamashita 2005).
623 In contrast, the *cry* toxin gene complements of genomes in clade 2 typically have readily
624 identifiable host ranges comprising a particular insect order or nematodes (**Figure 1B**), again
625 suggesting that isolates in this clade in particular are well adapted to exploiting invertebrate
626 hosts (Raymond et al 2010a, Raymond et al 2010b, Raymond and Bonsall 2013). Arguably,

any revision of the nomenclature would be most informative if it could reflect both phylogenetic affiliation and presence of Cry toxin inclusions.

629

Our analysis of plasmid distribution across the group revealed important patterns, illustrating the relationship between key plasmids and the genomes of specialized pathogens. Substantial sharing of near complete plasmids across genomes (**Figure 4, Table 1**) can indicate clonal expansions, sampling / sequencing bias of particular genotypes or horizontal transfer of plasmids between distinct lineages. The clonal expansion of *Ba* ST1 is well-established (Keim et al 2009, Zwick et al 2012); however the clonal expansion of the invertebrate pathogen *Bt* *subsp. kurstaki* (ST8), indicated by the central block of high plasmid sequence homology in clade 2 in Figure 4, is less well appreciated. This is the most frequently recorded genotype in the pubMLST database (Jolley et al 2004). It is also the most common genotype/serotype found on plants in a number of countries (Damgaard et al 1997, Maduell et al 2002, Ohba 1996, Raymond et al 2010b), possibly due to its ability to colonize plants from the soil (Raymond et al 2010b). Therefore, in terms of global abundance, the clonal expansion of *Bt. subsp. kurstaki* ST8 dwarfs that of *Ba*. The other abundant clone in our genomic dataset corresponds to ST26, or the ‘emetic cluster’ of cereulide producing *Bc* that are capable of causing lethal food-poisoning (Priest et al 2004, Vassileva et al 2007). In this case the strong representation of this cluster in the genomic database may be due to sampling bias.

647

If plasmid-bacteria associations are driven by co-evolution we predicted that particular plasmids should be associated with particular lineages. This was true for some plasmids (**Figure 4, Table 1**). The pXO2 plasmid of *Ba* was phylogenetically restricted to *Ba*; although plasmids with homology to pXO1 are widely distributed in clades 1 and clades 2 (Hu et al 2009a, Zheng et al 2013). A large number of plasmids, including the Cry-bearing plasmids which possess *orf156/157* minireplicons (Zheng et al 2013), were phylogenetically restricted to clade 2 (**Figure 4, Figure 6**), as has been found previously (Zheng et al 2017). Infectious cooperation, on the other hand, predicts that conjugative plasmids carrying social genes such as Cry toxins should be widely distributed across clades and show evidence of recent horizontal transfer. Several groups of plasmids that were widely distributed either within or between clades and which had conserved gene content were observed. However, several of these are small putatively parasitic plasmids such as the mobilizable 3kb plasmid sequenced from strains present in the ST26 emetic cluster (synonymous with pNC4), note

661 that this plasmid does not carry the cereulide toxin (Hattori et al 2012) (**Figure 4**). Within
662 clade 2 the widely distributed mobile elements with the highest levels of conserved gene
663 content are 60-80kb transposase-rich plasmids related to pKur6 and a class of \approx 8kb plamids
664 related to pKur 11, 12 and 13 (**Figure 6**). These are shared widely amongst *Bt* subspecies and
665 isolates infectious for Lepidoptera and Coleoptera (*kurstaki* ST8, *thuringiensis* ST10;
666 *morrisoni* ST23 *darmastadiensis* BGSC 4M3; *alesti* 4C3, T01-328, T0-40001) (**Figure 6**).
667 These plasmids are not associated with Cry toxin genes, and their association with particular
668 hosts could simply be the result of the increased opportunities for plasmid transfer between
669 strains that share an ecological niche in insect cadavers (Vilas-Bôas et al 2008).

670

671 For plasmids associated with the production of Cry toxins we also see that distantly related
672 lineages within clade 2 can share closely related plasmids, indicating recent horizontal
673 transfer. Plasmids closely related to pBtoxis, which carries multiple mosquitocidal Cry
674 proteins and was originally described in *Bt. israelensis* 4Q1 (Berry et al 2002) are found in
675 *Bt. morrisoni* PG14 and nematicidal *Bt. pakistani* ST17. A group of plasmids related to
676 *kurstaki* Cry toxin 300kb mega-plasmid (pKur2) are very widely distributed amongst nearly
677 all other *Bt* within clade 2. The 85kb plasmids carrying single Cry1A toxins (pHT73 and
678 pKur6) with *ori44* minireplicons are also shared by several distinct lineages (*kurstaki* ST8,
679 *thuringiensis* ST10; *darmastadiensis* BGSC 4M3); plasmids with these minireplicons have
680 been found widely across the *B. cereus* group (Zheng et al 2017). Not only are Cry toxins
681 plasmids present in distinct lineages but sister taxa, for example *kurstaki* HD73 and HD1;
682 *buibui* and BcRock42; *entomocidus* BSSC 4I4 and BcVD184, may or may not carry mega-
683 plasmids, a pattern also indicating recent loss or acquisition. This pattern of recent transfer is
684 consistent with infectious cooperation of Cry toxins, which are known public goods
685 (Raymond et al 2012). Nevertheless, gene content in these large plasmids is very unstable
686 indicating that costly social genes may be quickly lost in many lineages, perhaps in those not
687 fully adapted to a specialized pathogenic niche.

688

689 Together, our analyses describe multiple groups of specialized pathogens (*Ba* and several *Bt*
690 lineages) that are associated with phylogenetically restricted virulence plasmids. This
691 stratification among mobile plasmids, and the conserved allelic content, suggests that
692 particular plasmid-chromosome combinations result in clonal expansion of successful
693 pathogens. The distribution of virulence plasmids in particular suggests an association that
694 emerges out of the ability of plasmids to rapidly change gene content and associate with new

chromosomes (Keim and Wagner 2009), and the subsequent proliferation of successful plasmid chromosome combinations. While plasmid/bacteria coevolution may not appear to be consistent with regular transfer of plasmids during infectious cooperation, we do in fact see evidence for repeated transfer and loss of plasmids carrying cooperative Cry genes at a different taxonomic scale, namely within clade 2 (Rankin et al 2010, Raymond et al 2012). More widespread evidence of recent horizontal transfer may not be present because either these plasmids are restricted to one clade or because of lack of ecological opportunities for transfer in lines that are more distantly related. Infectious cooperation, of course, may occur at the level of MGEs (integrons, transposons) within plasmids or result in chromosome/plasmid combinations that are highly unstable due to genetic conflict. The most striking finding from the plasmid distribution data set was the very rapid and dynamic change in plasmid gene content between closely related genomes. Coupled with the extremely open pangenomic structure seen in this study and the evidence of widespread exchange of genes between plasmids and chromosome in previous work (Zheng et al 2015), this level of variability suggests that plasmids could be gaining and shedding genes on ecological timescales- a process that could explain hitch-hiking to high frequencies (Bergstrom et al 2000) as well as a means of rapidly responding to selective bottlenecks imposed by host colonization.

713

714 **Acknowledgements**

715 This work was supported by Medical Research Council (MRC) grants MR/M501608/1 and
716 MR/L015080/1 awarded to SKS, and a NERC fellowship NE/E012671/1 and BBSRC
717 BB/L00819X/1 grant to BR. GM was supported by a NISCHR Health Research Fellowship
718 (HF-14-13). EM is supported by a University of Bath PhD studentship. Computational
719 calculations were performed with HPC Wales (UK) and MRC CLIMB cloud-based
720 computing servers.

721

722 **Data accessibility**

723 Raw reads and assembled contiguous sequences of bacterial and plasmid genomes generated
724 in this study are accessible and associated with NCBI BioProject PRJNA395643.

725

726 **Conflict of interest**

727 Authors declare no conflict of interest.

728

729 **Figure and table legends**

730

731 **Figure 1. Phylogeny of 190 genomes and *cry* toxicity in the *Bacillus cereus* species**
732 **complex.** (A) The phylogenetic tree was reconstructed using gene-by-gene concatenated
733 alignments of 2,274 core genes, and an approximation of the maximum-likelihood algorithm
734 implemented in RAxML. The scale represents the number of substitutions per site. Clades
735 previously defined by MLST are specified in bold. *cry* endotoxin genes were identified in the
736 genomes with BtToxin_Scanner software and are indicated as present (green) or absent
737 (white) for each genome. Isolates from the *B. anthracis* clade are shown in pink. Numbers
738 next to the tip of branches on the tree indicate sequence types (ST) from the *B. cereus*
739 pubMLST database (<https://pubmlst.org/bcereus/>). (B) Inferred invertebrate host range of *B.*
740 *cereus* group isolates based on known toxicity spectra of *cry* genes present in genomes. Host
741 range allocations are detailed in Table S1 and based on data in van Frankenhuyzen (2009)
742 and sources within the Cry nomenclature database (Crickmore et al 2016)).

743

744 **Figure 2. Detection of chromosomal and plasmid genes in *B. cereus* group isolates.** (A)
745 Number of detected genes from a pangenome reference list of 27,016 genes in 190 *B. cereus*
746 group clades and Cry-positive and Cry-negative groups (as defined in Figure 1). (B) Total
747 number of detected genes from an unfiltered list of genes present in 410 full plasmids. The
748 number of isolates within each group is indicated below each distribution plot. Significance
749 of the difference in distribution averages was calculated after a one-way ANOVA with
750 Sidak's multiple comparison tests, with significance summarised as follow: ***: $p<0.0001$,
751 **: $p<0.01$.

752

753 **Figure 3. Allelic diversity of Cry-positive and Cry-negative *B. cereus* and *B. anthracis***
754 **isolates.** Allelic diversity was compared by calculating the number of unique alleles per
755 isolate for 2,192 core genes shared by all isolates of the dataset. (A) Overall distribution
756 shown as boxplots (min. to max.), with statistical significance between the distribution
757 inferred using a Kruskal-Wallis test with Dunn's multiple comparisons test, with
758 significances summarised as follow: ***, $p<0.0001$. (B) Frequency distribution of core
759 allelic diversity in each group. (C) gene-by-gene comparison of allelic diversity/isolate
760 between Cry-positive and Cry-negative isolates for each of 2,192 core genes (circles). The
761 proportionality line of equal allelic diversity between the two groups is shown in red. (D)
762 Distribution shown as boxplots (min. to max.) for each clade (1 to 3, excluding *B. anthracis*)

763 from clade 1 isolates). Each group was statistically different from one another (Kruskal-
764 Wallis test with Dunn's multiple comparisons tests; $p<0.0001$; except clade 2 vs. clade 4
765 which were not; $p=0.1306$). (D) Frequency distribution of core allelic diversity in each clade.
766

767 **Figure 4. Prevalence of 410 plasmids in 190 *B. cereus* group isolates.** The presence of
768 44,759 plasmid genes from 410 plasmid reference sequences (rows) was examined in 190
769 genomes (columns), and the proportion of detected plasmid genes per plasmid reference
770 sequence was calculated for each isolate. On the heatmap, blue indicates 100% of genes from
771 that plasmid are in the genome with progressively lighter shades of purple indicating
772 decreasing prevalence to white (fewer than 30% of genes are detected). The source of
773 plasmid isolations (coloured row headers) and the "species" of the bacterial genome
774 examined (coloured column headers) are given for *B. anthracis* (pink), *B. thuringiensis* or
775 Cry-positive isolate (green), *B. cereus* or Cry-negative isolate (grey). Isolates are ordered by
776 the tree (Figure 1A) and plasmids are clustered based on gene prevalence patterns inferred by
777 WebGimm (Joshi et al 2011) using the Context Specific Infinite Mixture Model (Freudenberg
778 et al 2010). Names on the figure indicate known plasmid names of interest.
779

780 **Figure 5. Frequency of plasmid genes in isolates.** Genes from 410 plasmids in a reference
781 collection were identified in all isolates by BLAST. The ratio between the number of plasmid
782 genes detected in each isolate and the total number of genes for the corresponding full
783 plasmid sequence is shown as a frequency plot for each group examined (Cry-positive
784 isolates, n=76; Cry-negative isolates, n=85; and *B. anthracis*; n=17). A value of 1 means that
785 all genes from corresponding plasmids were detected (full plasmid detection) while a value of
786 0 means that no gene from corresponding plasmids were detected. Values in between denote
787 partial detection of plasmids.
788

789 **Figure 6. Prevalence of 116 selected plasmids in 71 Clade 2 *B. cereus* group isolates.**
790 Visualisation is complementary to and focuses on specific plasmids and isolates from Figure
791 4. Isolates are ordered by the phylogeny from Figure 1 and plasmids from which >90% of
792 genes were detected in at least 1 Clade 2 isolate (n=116) are clustered based on gene
793 prevalence patterns inferred by WebGimm (Joshi et al 2011) using the Context Specific
794 Infinite Mixture Model (Freudenberg et al 2010). The plasmid names in red indicate Cry-
795 harbouring plasmids as inferred from a BtToxin_Scanner analysis presented in Table S3.
796

797 **Figure S1. Phylogeny of 190 genomes in the *Bacillus cereus* species complex in relation**
798 **to 40 additional *B. mycoides* and *B. weihenstephanensis* reference genomes.** White circles
799 denote isolates from this study (n=190), with an additional 19 isolates identified as *B.*
800 *mycoides* in NCBI (blue circles) and 21 isolates identified as *B. weihenstephanensis* (red
801 circles). The phylogenetic tree was reconstructed using gene-by-gene concatenated
802 alignments of ribosomal MLST core genes, and a rapid neighbour-joining algorithm
803 implemented in RapidNJ. MLST designations are indicated and the scale represents the
804 number of substitutions per site.

805

806 **Figure S2. Detection of plasmid-genes in closed, draft and novel genome sequences from**
807 **this study.** The presence of each of the 7,248 unique genes from 410 annotated plasmid
808 sequences was established in 53 closed genomes, 113 previously published draft genomes
809 and 23 novel draft genomes sequenced as part of this study. Asterisks denote significance as
810 measured by the p-value of nonparametric Mann-Whitney tests as follows: **: p=0.0052,
811 ***: p=0.0001 and ****: p<0.0001. Horizontal bars represent the average value of the
812 corresponding distribution.

813

814 **Figure S3. Detail of prevalence of 10 plasmids sequenced in this study in 190 *B. cereus***
815 **group isolates.** Visualisation method is as in Figure 4. Clear presence patterns were
816 observed, for which >90% of genes from given plasmids were detected in single strains, or
817 for the case of 3 plasmids up to 4 strains (p71o, pBGSC 4D4 and pBGSC 4D1; right column).
818 Names of strains are written next to top prevalence hits, and bold indicates the strain from
819 which the corresponding plasmid has been isolated from.

820

821 **Figure S4. Comparison of functional categories (COG) prevalence between groups of**
822 **genes from our dataset.** The proportion of 21 functional categories between genes with a
823 COG match shared by >95% of all isolates from our dataset (black), accessory genes shared
824 in <95% of all isolates from our dataset (red) and unique genes from 410 annotated *Bacillus*
825 plasmid genomes (blue) were compared. The number specified in the legend is the number of
826 genes with a COG match after analysis. Results were sorted (top to bottom) according to the
827 proportions observed in the core genome group of genes. Tukey multiple comparisons tests
828 after a two-way ANOVA only identified COG class L (Replication, recombination and
829 repair) to show significantly different proportions between accessory and plasmid genes

830 (adjusted p=0.0062 and p=0.0021 respectively) and core genes. No significant differences
831 were observed between accessory and plasmid genes.

832

833 **Table 1. Toxin-harbouring and plasmid detection in different groups of isolates used in**
834 **this study.** Details for each isolate are available in Table S1.

835

836 **Table S1. Isolates and genomes used in this study, including results of BtToxinScanner**
837 **detection of virulence factors.**

838

839 **Table S2. Assembly statistics for bacterial and plasmid genomes sequenced in this**
840 **study.**

841

842 **Table S3. List of 420 *B. cereus* group plasmid sequences used in this study.** A total of 410
843 plasmid sequences were obtained from NCBI, whereas 10 additional plasmid were sequenced
844 as part of this study

845

846 **Table S4. Detailed results of BtToxin_Scanner and filtering of results using BLAST on**
847 **the nr protein database on NCBI.**

848

849 **Table S5. List of genes detected in more than 10 Cry-positive isolates and absent in all**
850 **Cry-negative isolates.** The gene name prefix "id_" denotes genes detected in novel genomes
851 that do not correlate to any reference sequence in our analysis.

852

853 **Table S6. List of clade-specific genes present in all isolates from given clades.** No gene
854 was detected exclusively in Clade 1 excluding *B. anthracis* nor in Clade 5. The gene name
855 prefix "id_" denotes genes detected in novel genomes that do not correlate to any reference
856 sequence in our analysis.

857

858 **References**

- 859
- 860 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search
861 tool. *Journal of molecular biology* **215**: 403-410.
- 862 Antonation KS, Grutzmacher K, Dupke S, Mabon P, Zimmermann F, Lankester F et al
863 (2016). *Bacillus cereus* Biovar Anthracis Causing Anthrax in Sub-Saharan Africa-
864 Chromosomal Monophyly and Broad Geographic Distribution. *PLoS neglected*
865 *tropical diseases* **10**: e0004923.
- 866 Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA et al (2008). The RAST
867 Server: rapid annotations using subsystems technology **9**: 75.
- 868 Bergstrom CT, Lipsitch M, Levin BR (2000). Natural selection, infectious transfer and the
869 existence conditions for bacterial plasmids. *Genetics* **155**: 1505-1519.
- 870 Berry C, O'Neil S, Ben-Dov E, Jones AF, Murphy L, Quail MA et al (2002). Complete
871 sequence and organization of pBtoxis the toxin-coding plasmid of *Bacillus*
872 *thuringiensis* subsp. *israelensis*. *Appl Env Microbiol* **68**: 5082-5095.
- 873 Bologa M, Kamtchoua T, Hopfer R, Sheng X, Hicks B, Bixler G et al (2012). Safety and
874 immunogenicity of pneumococcal protein vaccine candidates: monovalent choline-
875 binding protein A (PcpA) vaccine and bivalent PcpA-pneumococcal histidine triad
876 protein D vaccine. *Vaccine* **30**: 7461-7468.
- 877 Bravo A, Likityvatanavong S, Gill SS, Soberon M (2011). *Bacillus thuringiensis*: A story of a
878 successful bioinsecticide. *Insect Biochem Mol Biol* **41**: 423-431.
- 879 Cardazzo B, Negrisolo E, Carraro L, Alberghini L, Patarnello T, Giaccone V (2008). Multiple-
880 locus sequence typing and analysis of toxin genes in *Bacillus cereus* food-borne
881 isolates **74**: 850-860.
- 882 Ceuppens S, Uyttendaele M, Drieskens K, Heyndrickx M, Rajkovic A, Boon N et al (2012).
883 Survival and germination of *Bacillus cereus* spores during in vitro simulation of
884 gastrointestinal transit occurred without outgrowth and enterotoxin production. *Appl*
885 *Environ Microbiol* **78**: AEM.02142-02112-07705.
- 886 Crickmore N, Baum J, Bravo A, Lereclus D, Narva K, Sampson K et al (2016). *Bacillus*
887 *thuringiensis* toxin nomenclature.
- 888 Damgaard PH, Hansen BM, Pedersen JC, Eilenberg J (1997). Natural occurrence of *Bacillus*
889 *thuringiensis* on cabbage foliage and in insects associated with cabbage crops. *J*
890 *Appl Microbiol* **82**: 253-258.
- 891 de Maagd RA, Bravo A, Berry C, Crickmore N, Schnepf HE (2003). Structure, diversity, and
892 evolution of protein toxins from spore-forming entomopathogenic bacteria. *Annu Rev*
893 *Genet* **37**: 409-433.
- 894 Deng C, Slamti L, Ben R, Liu G, Lemy C, Gominet M et al (2015). Division of labour and
895 terminal differentiation in a novel *Bacillus thuringiensis* strain. *ISME J* **9**: 286-296.
- 896 Didelot X, Barker M, Falush D, Priest FG (2009). Evolution of pathogenicity in the *Bacillus*
897 *cereus* group. *Syst Appl Microbiol* **32**: 81-90.
- 898 Dimitriu T, Lotton C, Bénard-Capelle J, Misevic D, Brown SP, Lindner AB et al (2014).
899 Genetic information transfer promotes cooperation in bacteria. *Proc Natl Acad Sci*
900 *USA* **111**: 11103-11108.
- 901 Eberhard WG (1990). Evolution in bacterial plasmids and levels of selection. *Q Rev Biol* **65**:
902 3-22.
- 903 EFSA (2016). Risks for public health related to the presence of *Bacillus cereus* and other
904 *Bacillus* spp. including *Bacillus thuringiensis* in foodstuffs. *EFSA Journal* **14**: 99.
- 905 Federici BA, Siegel JP (2007). Assessment of safety of *Bacillus thuringiensis* and Bt crops
906 used for insect control. In: Hammond BG (ed). *Safety of Food Proteins in Agricultural*
907 *Crops*. Taylor and Francis: London. pp 46-101.
- 908 Freudenberg JM, Sivaganesan S, Wagner M, Medvedovic M (2010). A semi-parametric
909 Bayesian model for unsupervised differential co-expression analysis. *BMC*
910 *bioinformatics* **11**: 234.

- 911 Gonzalez JM, Brown BJ, Carlton BC (1982). Transfer of *Bacillus thuringiensis* plasmids
912 coding for delta-endotoxin among strains of *B. thuringiensis* and *B. cereus*. *Proc Natl
913 Acad Sci USA* **79**: 6951-6955.
- 914 Guinebretière M-H, Thompson FL, Sorokin A, Normand P, Dawyndt P, Ehling-Schulz M et al
915 (2008). Ecological diversification in the *Bacillus cereus* Group. *Environ Microbiol* **10**:
916 851-865.
- 917 Guinebretière M-H, Velge P, Couvert O, Carlin F, Debuyser ML, Nguyen-The C (2010).
918 Ability of *Bacillus cereus* Group strains to cause food poisoning varies according to
919 phylogenetic affiliation (Groups I to VII) rather than species affiliation. *J Clin Microbiol*
920 **48**: 3388-3391.
- 921 Hacker J, Carniel E (2001). Ecological fitness, genomic islands and bacterial pathogenicity.
922 *EMBO reports* **2**: 376-381.
- 923 Hattori M, Yamashita A, Toh H, Oshima K, Shiba T (2012). Complete genome sequence of
924 *Bacillus cereus* NC7401, which produces high levels of the emetic toxin cereulide. *J
925 Bacteriol* **194**: 4767-4768.
- 926 Hayakawa T, Kanagawa R, Kotani Y, Kimura M, Yamagiwa M, Yamane Y et al (2007).
927 Parasporin-2Ab, a Newly Isolated Cytotoxic Crystal Protein from *Bacillus
928 thuringiensis*. *Curr Microbiol* **55**: 278-283.
- 929 Helgason E, Okstad OA, Caugant DA, Johansen HA, Fouet A, Mock M et al (2000). *Bacillus
930 anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis* - one species on the basis of
931 genetic evidence. *Appl Environ Microbiol* **66**: 2627-2630.
- 932 Hu X, Swiecicka I, Timmery S, Mahillon J (2009a). Sympatric soil communities of *Bacillus*
933 *cereus* sensu lato: population structure and potential plasmid dynamics of pXO1- and
934 pXO2-like elements. *FEMS Microbiol Ecol* **70**: 344-355.
- 935 Hu X, Van der Auwera G, Timmery S, Zhu L, Mahillon J (2009b). Distribution, diversity, and
936 potential mobility of extrachromosomal elements related to the *Bacillus anthracis*
937 pXO1 and pXO2 virulence plasmids. *Appl Env Microbiol* **75**: 3016-3028.
- 938 Huang TW, Chen TL, Chen YT, Lauderdale TL, Liao TL, Lee YT et al (2013). Copy Number
939 Change of the NDM-1 sequence in a multidrug-resistant *Klebsiella pneumoniae*
940 clinical isolate. *PloS one* **8**: e62774.
- 941 Hugh-Jones M, Blackburn J (2009). The ecology of *Bacillus anthracis*. *Mol Aspects Med*.
- 942 Ivanova N, Sorokin A, Anderson I, Galleron N, Candelon B, Kapatral V et al (2003). Genome
943 sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. *Nature*
944 **423**: 87-91.
- 945 Jolley KA, Maiden MC (2010). BIGSdb: Scalable analysis of bacterial genome variation at
946 the population level **11**: 595.
- 947 Jolley KS, Chan MS, Maiden MC (2004). mlstNet - distributed multi-locus sequence typing
948 (MLST) databases. *BMC Bioinf* **5**: 86.
- 949 Joshi VK, Freudenberg JM, Hu Z, Medvedovic M (2011). WebGimm: An integrated web-
950 based platform for cluster analysis, functional analysis, and interactive visualization
951 of results. *Source code for biology and medicine* **6**: 3.
- 952 Kado CI, Liu ST (1981). Rapid procedure for detection and isolation of large and small
953 plasmids. *J Bacteriol* **145**: 1365-1373.
- 954 Katoh K, Standley DM (2013). MAFFT multiple sequence alignment software version 7:
955 improvements in performance and usability. *Mol Biol Evol* **30**: 772-780.
- 956 Keim P, Gruendike JM, Klevytska AM, Schupp JM, Challacombe J, Okinaka R (2009). The
957 genome and variation of *Bacillus anthracis*. *Mol Aspects Med* **30**: 397-405.
- 958 Keim PS, Wagner DM (2009). Humans and evolutionary and ecological forces shaped the
959 phylogeography of recently emerged diseases. *Nat Rev Micro* **7**: 813-821.
- 960 Lawrence JG (2005). Common themes in the genome strategies of pathogens **15**: 584-588.
- 961 Luo R, Mann B, Lewis WS, Rowe A, Heath R, Stewart ML et al (2005). Solution structure of
962 choline binding protein A, the major adhesin of *Streptococcus pneumoniae*. *The
963 EMBO journal* **24**: 34-43.

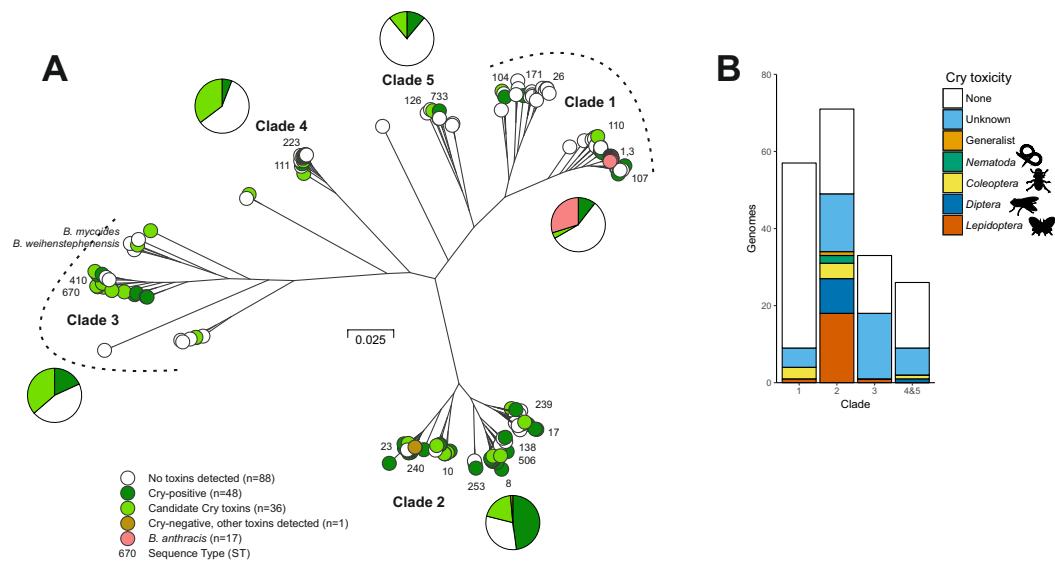
- 964 Maduell P, Callejas R, Cabrera KR, Armengol G, Orduz S (2002). Distribution and
965 characterization of *Bacillus thuringiensis* on the phylloplane of species of *Piper*
966 (Piperaceae) in three altitudinal levels. *Microb Ecol* **44**: 144-153.
- 967 Mahillon J, Rezsöhazy R, Hallet B, Delcour J (1994). IS231 and other *Bacillus thuringiensis*
968 transposable elements: a review **93**: 13-26.
- 969 Maiden MC, van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA et al (2013). MLST
970 revisited: the gene-by-gene approach to bacterial genomics **11**: 728-736.
- 971 Meric G, Yahara K, Mageiros L, Pascoe B, Maiden MC, Jolley KA et al (2014). A reference
972 pan-genome approach to comparative bacterial genomics: identification of novel
973 epidemiological markers in pathogenic *Campylobacter*. *PLoS one* **9**: e92798.
- 974 Meric G, Miragaia M, de Been M, Yahara K, Pascoe B, Mageiros L et al (2015). Ecological
975 Overlap and Horizontal Gene Transfer in *Staphylococcus aureus* and
976 *Staphylococcus epidermidis*. *Genome biology and evolution* **7**: 1313-1328.
- 977 Meric G, Hitchings MD, Pascoe B, Sheppard SK (2016). From Escherich to the Escherichia
978 coli genome. *The Lancet Infectious diseases* **16**: 634-636.
- 979 Monteil CL, Yahara K, Studholme DJ, Mageiros L, Meric G, Swingle B et al (2016).
980 Population-genomic insights into emergence, crop-adaptation, and dissemination of
981 *Pseudomonas syringae* pathogens **2**.
- 982 Morley L, McNally A, Paszkiewicz K, Corander J, Meric G, Sheppard SK et al (2015). Gene
983 Loss and Lineage-Specific Restriction-Modification Systems Associated with Niche
984 Differentiation in the *Campylobacter jejuni* Sequence Type 403 Clonal Complex. *Appl*
985 *Environ Microb* **81**: 3641-3647.
- 986 Murray S, Pascoe B, Meric G, Mageiros L, Yahara K, Hitchings MD et al (2017).
987 Recombination-Mediated Host Adaptation by Avian *Staphylococcus aureus*. *Genome*
988 *biology and evolution* **9**: 830-842.
- 989 Nogueira T, Rankin DJ, Touchon M, Taddei F, Brown SP, Rocha EPC (2009). Horizontal
990 gene transfer of the secretome drives the evolution of bacterial cooperation and
991 virulence. *Curr Biol* **19**: 1683-1691.
- 992 Noguera PA, Ibarra JE (2010). Detection of new cry genes of *Bacillus thuringiensis* by use of
993 a novel PCR primer system. *Appl Environ Microb* **76**: 6150-6155.
- 994 Ohba M (1996). *Bacillus thuringiensis* populations naturally occurring on mulberry leaves: A
995 possible source of the populations associated with silkworm-rearing insectaries. *J*
996 *Appl Bacteriol* **80**: 56-64.
- 997 Okinaka RT, Cloud K, Hampton O, Hoffmaster AR, Hill KK, Keim P et al (1999). Sequence
998 and organization of pXO1, the large *Bacillus anthracis* plasmid harboring the anthrax
999 toxin genes. *J Bacteriol* **181**: 6509-6515.
- 1000 Okinaka RT, Price EP, Wolken SR, Gruendike JM, Chung WK, Pearson T et al (2011). An
1001 attenuated strain of *Bacillus anthracis* (CDC 684) has a large chromosomal inversion
1002 and altered growth kinetics. *Bmc Genomics* **12**: 477.
- 1003 Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T et al (2014). The SEED and
1004 the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).
1005 *Nucleic acids research* **42**: D206-214.
- 1006 Priest FG, Barker M, Baillie LWJ, Holmes EC, Maiden MCJ (2004). Population structure and
1007 evolution of the *Bacillus cereus* group. *J Bacteriol* **186**: 7959-7970.
- 1008 Rankin DJ, Rocha EPC, Brown SP (2010). What traits are carried on mobile genetic
1009 elements, and why? *Heredity* **106**: 1-10.
- 1010 Raymond B, Davis D, Bonsall MB (2007). Competition and reproduction in mixed infections
1011 of pathogenic and non-pathogenic *Bacillus* spp. *J Invertebr Pathol* **96**: 151-155.
- 1012 Raymond B, Johnston PR, Nielsen-Leroux C, Lereclus D, Crickmore N (2010a). *Bacillus*
1013 *thuringiensis*: an impotent pathogen? *Trends Microbiol* **18**: 189-194.
- 1014 Raymond B, Wyres KL, Sheppard SK, Ellis RJ, Bonsall MB (2010b). Environmental factors
1015 determining the epidemiology and population genetic structure of the *Bacillus cereus*
1016 group in the field **6**: e1000905.
- 1017 Raymond B, West SA, Griffin AS, Bonsall MB (2012). The dynamics of cooperative bacterial
1018 virulence in the field. *Science* **337**: 85-88.

- 1019 Raymond B, Bonsall MB (2013). Cooperation and the evolutionary ecology of bacterial
1020 virulence: the *Bacillus cereus* group as a novel study system. *Bioessays* **35**: 706-
1021 716.
- 1022 Raymond B, Federici B (2017). In defense of *Bacillus thuringiensis*, the safest and most
1023 successful microbial insecticide available to humanity – a response to EFSA **In
1024 press**.
- 1025 Read TD, Peterson SN, Tourasse N, Baillie LW, Paulsen IT, Nelson KE *et al* (2003). The
1026 genome sequence of *Bacillus anthracis* Ames and comparison to closely related
1027 bacteria. *Nature* **423**: 81-86.
- 1028 Reyes-Ramirez A, Ibarra JE (2008). Plasmid patterns of *Bacillus thuringiensis* type strains.
1029 *Appl Environ Microbiol* **74**: 125-129.
- 1030 Ruan L, Crickmore N, Peng D, Sun M (2015). Are nematodes a missing link in the
1031 confounded ecology of the entomopathogen *Bacillus thuringiensis*? *Trends Microbiol*
1032 **23**: 341-346.
- 1033 San Millan A, Santos-Lopez A, Ortega-Huedo R, Bernabe-Balas C, Kennedy SP, Gonzalez-
1034 Zorn B (2015). Small-plasmid-mediated antibiotic resistance is enhanced by
1035 increases in plasmid copy number and bacterial fitness. *Antimicrob Agents Ch* **59**:
1036 3335-3341.
- 1037 Sansonetti PJ, Kopecko DJ, Formal SB (1981). *Shigella sonnei* plasmids: evidence that a
1038 large plasmid is necessary for virulence. *Infect Immun* **34**: 75-83.
- 1039 Schnepp E, Crickmore N, Van Rie J, Lereclus D, Baum J, Feitelson J *et al* (1998). *Bacillus
1040 thuringiensis* and its pesticidal crystal proteins. *Microbiol Mol Biol Rev* **62**: 775-806.
- 1041 Sheppard SK, Jolley KA, Maiden MCJ (2012). A Gene-By-Gene Approach to Bacterial
1042 Population Genomics: Whole Genome MLST of *Campylobacter* **3**: 261-277.
- 1043 Siegel JP (2001). The mammalian safety of *Bacillus thuringiensis* based insecticides. *J
1044 Invert Path* **77**: 13-21.
- 1045 Smalla K, Jechalke S, Top EM (2015). Plasmid Detection, Characterization, and Ecology.
1046 *Microbiology spectrum* **3**: PLAS-0038-2014.
- 1047 Smith J (2001). The social evolution of bacterial pathogenesis. *Proc R Soc Lond B* **268**: 61-
1048 69.
- 1049 Sorokin A, Candelon B, Guilloux K, Galleron N, Wackerow-Kouzova N, Ehrlich SD *et al*
1050 (2006). Multiple-locus sequence typing analysis of *Bacillus cereus* and *Bacillus
1051 thuringiensis* reveals separate clustering and a distinct population structure of
1052 psychrotrophic strains. *Appl Environ Microbiol* **72**: 1569-1578.
- 1053 Stamatakis A (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of
1054 large phylogenies **30**: 1312-1313.
- 1055 Tourasse NJ, Helgason E, Økstad OA, Hegna IK, Kolstø A-B (2006). The *Bacillus cereus*
1056 group: novel aspects of population structure and genome dynamics. *J Appl Microbiol*
1057 **101**: 579-593.
- 1058 Tourasse NJ, Helgason E, Klevan A, Sylvestre P, Moya M, Haustant M *et al* (2011).
1059 Extended and global phylogenetic view of the *Bacillus cereus* group population by
1060 combination of MLST, AFLP, and MLEE genotyping data **28**: 236-244.
- 1061 Turnbull PCB (2002). Introduction: Anthrax history, disease and ecology. In: Koehler TM
1062 (ed). *Anthrax*. Springer., pp 1-19.
- 1063 Van der Auwera G, Mahillon J (2008). Transcriptional analysis of the conjugative plasmid
1064 pAW63 from *Bacillus thuringiensis* **60**: 190-199.
- 1065 Van der Auwera GA, Timmery S, Hoton F, Mahillon J (2007). Plasmid exchanges among
1066 members of the *Bacillus cereus* group in foodstuffs. *Int J Food Microbiol* **113**: 164-
1067 172.
- 1068 Van Ert MN, Easterday WR, Huynh LY, Okinaka RT, Hugh-Jones ME, Ravel J *et al* (2007).
1069 Global genetic population structure of *Bacillus anthracis*. *PloS one* **2**: e461.
- 1070 van Frankenhuyzen K (2009). Insecticidal activity of *Bacillus thuringiensis* crystal proteins. *J
1071 Invert Pathol* **101**: 1-16.
- 1072 van Frankenhuyzen K (2013). Cross-order and cross-phylum activity of *Bacillus thuringiensis*
1073 pesticidal proteins. *J Invert Path* **114**: 76-85.

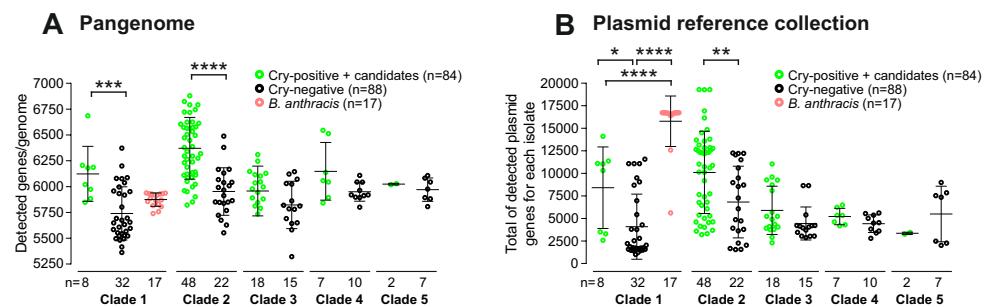
- 1074 Vassileva M, Torii K, Oshimoto M, Okamoto A, Agata N, Yamada K et al (2006).
1075 Phylogenetic analysis of *Bacillus cereus* isolates from severe systemic infections
1076 using multilocus sequence typing scheme. *Microbiol Immunol* **50**: 743-749.
1077 Vassileva M, Torii K, Oshimoto M, Okamoto A, Agata N, Yamada K et al (2007). A new
1078 phylogenetic cluster of cereulide-producing *Bacillus cereus* strains. *J Clin Microbiol*
1079 **45**: 1274-1277.
1080 Vidal-Quist JC, Rogers HJ, Mahenthiralingam E, Berry C (2013). *Bacillus thuringiensis*
1081 colonises plant roots in a phylogeny-dependent manner. *FEMS Microbiol Ecol* **86**:
1082 474-489.
1083 Vilas-Boas G, Sanchis V, Lereclus D, Lemos MVF, Bourguet D (2002). Genetic
1084 differentiation between sympatric populations of *Bacillus cereus* and *Bacillus*
1085 *thuringiensis*. *Appl Environ Microbiol* **68**: 1414-1424.
1086 Vilas-Bôas G, Vilas-Boas LA, Lereclus D, Arantes OMN (2008). *Bacillus thuringiensis*
1087 conjugation under environmental conditions. *FEMS Microbiol Ecol* **24**: 369-374.
1088 West AW, Burges HD, Dixon TJ, Wyborn CH (1985). Survival of *Bacillus thuringiensis* and
1089 *Bacillus cereus* spore inocula in soil - effects of pH, moisture, nutrient availability and
1090 indigenous microorganisms. *Soil Biol Biochem* **17**: 657-665.
1091 West S, Diggle S, Buckling A, Gardner A, Griffin A (2007). The social lives of microbes.
1092 *Annu Rev Ecol Syst* **38**: 53-77.
1093 Wilson MK, Vergis JM, Alem F, Palmer JR, Keane-Myers AM, Brahmbhatt TN et al (2011).
1094 *Bacillus cereus* G9241 makes anthrax toxin and capsule like highly virulent B.
1095 anthracis Ames but behaves like attenuated toxigenic nonencapsulated B. anthracis
1096 Sterne in rabbits and mice. *Infect Immun* **79**: 3012-3019.
1097 Xu D, Cote JC (2006). Sequence diversity of the *Bacillus thuringiensis* and *B. cereus* sensu
1098 *lato* flagellin (H Antigen) protein: comparison with H Serotype Diversity. *Appl Environ*
1099 *Microbiol* **72**: 4653-4662.
1100 Yahara K, Meric G, Taylor AJ, de Vries SP, Murray S, Pascoe B et al (2017). Genome-wide
1101 association of functional traits linked with *Campylobacter jejuni* survival from farm to
1102 fork. *Environ Microbiol* **19**: 361-380.
1103 Yamashita S (2005). Typical three-domain Cry proteins of *Bacillus thuringiensis* strain A1462
1104 exhibit cytoidal activity on limited human cancer cells. *J Biochem* **138**: 663-672.
1105 Yang F (2005). Genome dynamics and diversity of *Shigella* species, the etiologic agents of
1106 bacillary dysentery. *Nucleic Acids Res* **33**: 6445-6458.
1107 Yara K, Kunimi Y, Iwahana H (1997). Comparative studies of growth characteristic and
1108 competitive ability in *Bacillus thuringiensis* and *Bacillus cereus* in soil. *Appl Entomol*
1109 *Zool* **32**: 625-634.
1110 Ye W, Zhu L, Liu Y, Crickmore N, Peng D, Ruan L et al (2012). Mining new crystal protein
1111 genes from *Bacillus thuringiensis* on the basis of mixed plasmid-enriched genome
1112 sequencing and a computational pipeline. *Appl Environ Microbiol* **78**: 4795-4801.
1113 Zerbino DR, Birney E (2008). Velvet: algorithms for de novo short read assembly using de
1114 Bruijn graphs. *Genome Res* **18**: 821-829.
1115 Zheng J, Peng D, Ruan L, Sun M (2013). Evolution and dynamics of megaplasmids with
1116 genome sizes larger than 100 kb in the *Bacillus cereus* group. *BMC Evol Biol* **13**:
1117 262.
1118 Zheng J, Guan Z, Cao S, Peng D, Ruan L, Jiang D et al (2015). Plasmids are vectors for
1119 redundant chromosomal genes in the *Bacillus cereus* group. *Bmc Genomics* **16**: 6-6.
1120 Zheng J, Gao Q, Liu L, Liu H, Wang Y, Peng D et al (2017). Comparative genomics of
1121 *Bacillus thuringiensis* reveals a path to specialized exploitation of multiple
1122 invertebrate hosts. *mBio* **8**: e00822-00817.
1123 Zhou L, Slamti L, Nielsen-Leroux C, Lereclus D, Raymond B (2014). The social biology of
1124 quorum-sensing in a naturalistic host pathogen system. *Curr Biol* **24**: 2417-2422.
1125 Zhu Y, Shang H, Zhu Q, Ji F, Wang P, Fu J et al (2011). Complete genome sequence of
1126 *Bacillus thuringiensis* serovar *finitimus* strain YBT-020. *J Bacteriol* **193**: 2379-2380.

1127 Zwick ME, Joseph SJ, Didelot X, Chen PE, Bishop-Lilly KA, Stewart AC *et al* (2012).
1128 Genomic characterization of the *Bacillus cereus* *sensu lato* species: backdrop to the
1129 evolution of *Bacillus anthracis*. *Genome Res* **22**: 1512-1524.
1130
1131

1132 Figure 1

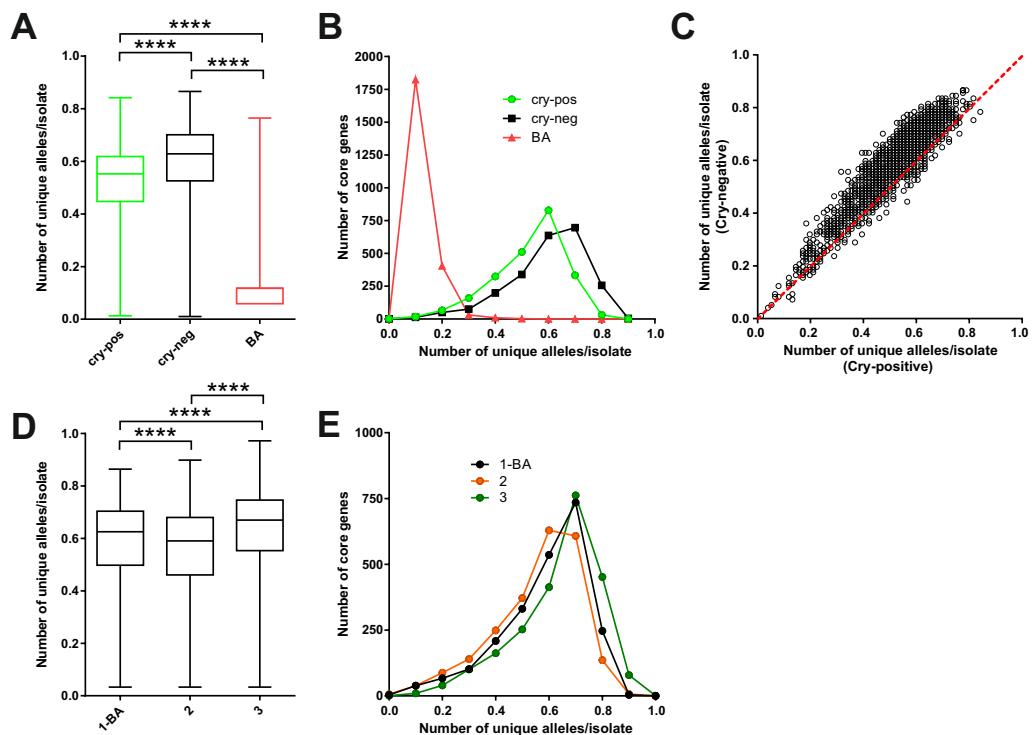


1133
1134 Figure 2



1135

1136 Figure 3

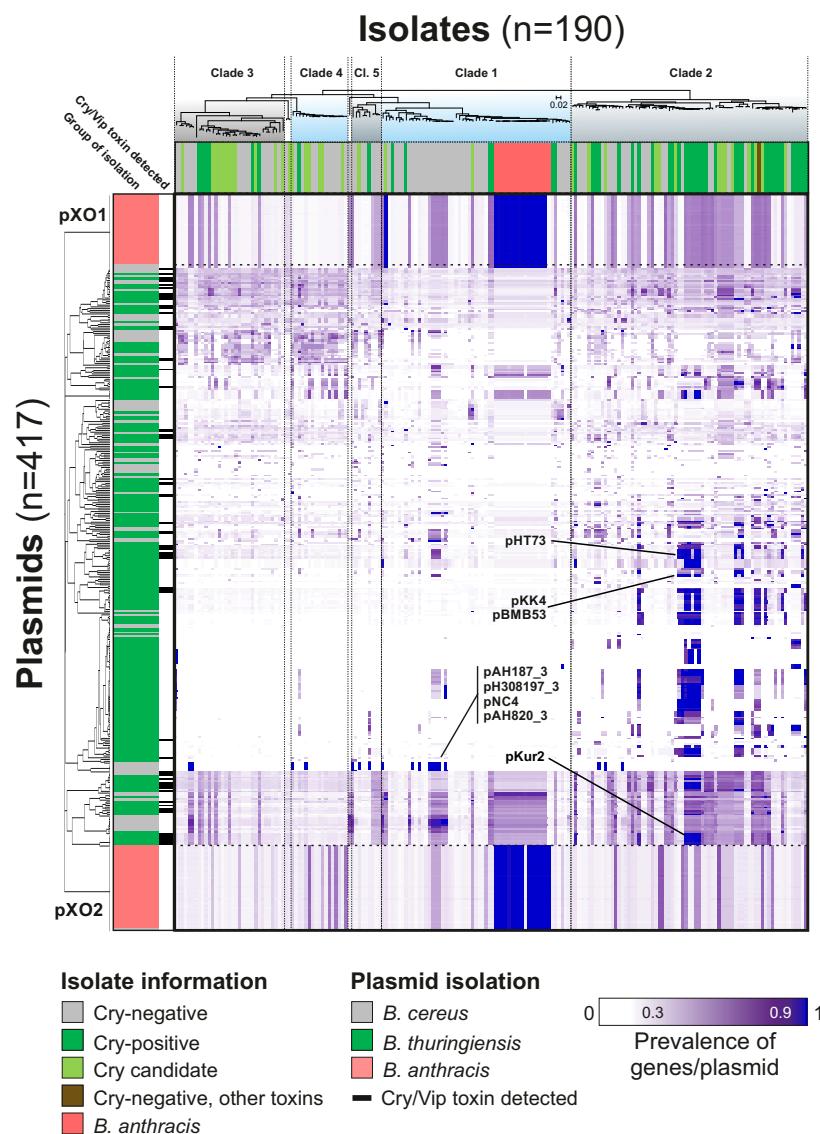


1137

1138

1139

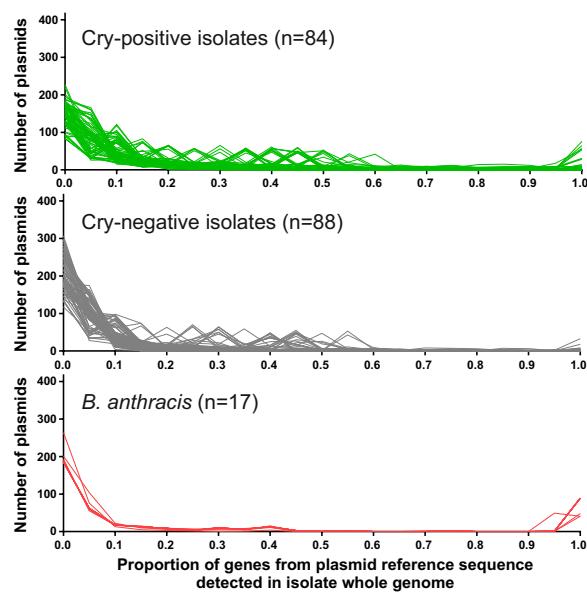
1140 Figure 4



1141

1142

1143 Figure 5



1144

1145

