Research article

From root to tips: sporulation evolution and specialization in *Bacillus subtilis* and the intestinal pathogen *Clostridioides difficile*

Paula Ramos-Silva[1*], Mónica Serrano[2], Adriano O. Henriques[2]

[1]Instituto Gulbenkian de Ciência, Oeiras, Portugal

[2]Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Oeiras, Portugal

***Corresponding author:**

Present address: Naturalis Biodiversity Center, Marine Biodiversity, Leiden, The Netherlands

Phone: +31 717519283

Email: paula.srs@gmail.com

Running title: Sporulation from root to tips

Keywords: sporulation, bacterial genome evolution, horizontal gene transfer, taxon-specific genes, *Bacillus subtilis*, *Clostridioides difficile*

Abstract

Bacteria of the Firmicutes phylum are able to enter a developmental pathway that culminates with the formation of a highly resistant, dormant spore. Spores allow environmental persistence, dissemination and for pathogens, are infection vehicles. In both the model *Bacillus subtilis*, an aerobic species, and in the intestinal pathogen *Clostridioides difficile*, an obligate anaerobe, sporulation mobilizes hundreds of genes. Their expression is coordinated between the forespore and the mother cell, the two cells that participate in the process, and is kept in close register with the course of morphogenesis. The evolutionary mechanisms by which sporulation emerged and evolved in these two species, and more broadly across Firmicutes, remain largely unknown. Here, we trace the origin and evolution of sporulation. Using the genes involved in the process in *B. subtilis* and *C. difficile*, and estimating their gain-loss dynamics in a comprehensive bacterial macro-evolutionary framework we show that sporulation evolution was driven by two major gene gain events, the first at the base of the Firmicutes and the second at the base of the *B. subtilis* group and within the Peptostreptococcaceae family, which includes *C. difficile*. We also show that early and late sporulation regulons have been co-evolving and that sporulation genes entail greater innovation in *B. subtilis* with many Bacilli-lineage restricted genes. In contrast, *C. difficile* more often recruits new sporulation genes by horizontal gene transfer, which reflects both its highly mobile genome, the complexity of the gut microbiota and an adjustment of sporulation to this particular ecosystem.

51 **Introduction**

52 Bacterial endospores (spores for simplicity) are among the most resilient cells
53 known to us. Because of their resilience to extremes of several physical and
54 chemical conditions (Setlow 2014), spores can spread and persist in the environment
55 for long periods of time and in a variety of ecosystems (Nicholson et al. 2000), from
56 the human gut (Hong et al. 2009; Browne et al. 2016) to the deep sea (Urios 2004;
57 Dick et al. 2006; Fang et al. 2017). Spores can also have a profound impact on
58 human health (Yano et al. 2015; Browne et al. 2016). Spore-formers can either be
59 symbiotic (Angert and Losick 1998; Flint et al. 2005), pathogenic (Abt et al. 2016) or
60 free-living (Ramos-Silva et al. 2015; Fang et al. 2017). Spores have a core
61 compartment that encloses the DNA delimited by a membrane and surrounded by a
62 thin layer of peptidoglycan that will serve as the cell wall of the new cell that emerges
63 from spore germination. A second layer of modified peptidoglycan essential for
64 dormancy, the cortex, is surrounded by several proteinaceous layers (Henriques and
65 Moran 2007). In *Bacillus subtilis* a multilayered protein coat forms the spore surface
66 (Lai et al. 2003; Kim et al. 2006); in other species, such as *Bacillus anthracis* and
67 *Clostridioides difficile* the spore is further enclosed in a glycosylated exosporium
68 (Redmond 2004; Permpoonpattana et al. 2011; Thompson et al. 2012).

69 Sporulation usually begins with the formation of a polar septum that leads to
70 an asymmetric cell division resulting in two cells, the forespore (or future spore) and
71 the mother cell, each with an identical copy of the chromosome but expressing
72 different genetic programs. At an intermediate stage, the forespore is engulfed by the
73 mother cell, and becomes isolated from the external medium. At later stages, while
74 gene expression in the forespore prepares this cell for dormancy, the mother cell
75 drives the assembly of the cortex, coat and exosporium, and eventually triggers its
76 own lysis to release the mature spore (Higgins and Dworkin 2012). Entry in
77 sporulation is triggered by the activation of the transcriptional regulator Spo0A (Fujita
78 and Losick 2005). Spo0A is conserved among spore-formers and its presence is

79 often used as an indicator of sporulation ability (Galperin 2013; Filippidou et al.
80 2016). The activity of Spo0A is required for the division that takes place at the onset
81 of sporulation. Following division, gene expression in the forespore and the terminal
82 mother cell is sequentially activated by cell type-specific RNA polymerase sigma
83 factors (Losick and Stragier 1992; Stragier and Losick 1996; Kroos and Yu 2000;
84 Rudner and Losick 2001): prior to engulfment completion, gene expression is
85 governed by σ^F in the forespore and by σ^E in the mother cell; following engulfment
86 completion, σ^F is replaced by σ^G and σ^E is replaced by σ^K. These sigma factors are
87 also conserved among spore-formers (de Hoon et al. 2010; Galperin et al. 2012;
88 Abecasis et al. 2013). The identification of the genes whose transcription is directly
89 controlled by σ^F, σ^E, σ^G and σ^K was initially made in *Bacillus subtilis*, an aerobic
90 organism, representative of the class Bacilli, which is still the best characterized
91 spore-former (Fawcett et al. 2000; Eichenberger et al. 2003; Wang et al. 2006;
92 Meisner et al. 2008; Mäder et al. 2012).

93 *Clostridium difficile*, later re-classified as *Clostridioides difficile*, (Lawson et al.
94 2016), is a nosocomial pathogen and the main cause of intestinal diseases linked to
95 antibiotic therapy worldwide (Bauer et al. 2011; Lessa et al. 2015; Abt et al. 2016).
96 As an obligate anaerobe, *C. difficile* relies on its sporulation ability to spread
97 horizontally between hosts via fecal and oral transmission (Lawley et al. 2009; Janoir
98 et al. 2013). This has motivated a number of studies on the control of sporulation and
99 *C. difficile* has emerged in recent years as the anaerobic model for the process,
100 being currently the best characterized species within the Clostridia (Paredes-Sabja et
101 al. 2014). Importantly, several studies have been directed at the identification of the
102 genes under the control of the cell type-specific sigma factors (Fimlaid et al. 2013;
103 Saujet et al. 2013), and several genes have been characterized, highlighting
104 conservancy of function of the main regulatory proteins, but also major differences in

105 the control of sporulation relative to the model *B. subtilis* (Fimlaid et al. 2015; Fimlaid
106 and Shen 2015; Pishdadian et al. 2015; Serrano et al. 2016).

107 Based on the extensive knowledge of sporulation in *B. subtilis*, previous
108 comparative genomics studies have led to the identification of a core (*i.e.* genes
109 conserved in all spore-formers), a signature (present in all spore-formers but absent
110 in non-spore-forming lineages) and Bacilli-specific sporulation genes (Onyenwoke et
111 al. 2004; Galperin et al. 2012; Abecasis et al. 2013) as well as to elucidate how
112 evolutionary patterns are correlated with gene functional roles (de Hoon et al. 2010).

113 Here, we have manually compiled the sporulation genes of *B. subtilis* strain
114 168 together with the most recent set of sporulation genes of *C. difficile* strain 630
115 whose expression is dependent on the cell type-specific sigma factors σ^F, σ^E, σ^G and
116 σ^K. Using comparative genomics we show that the σ regulons are highly divergent
117 between the two species, which share only a small fraction of genes. We then
118 mapped the presence/absence of both sets of sporulation genes in a solid and
119 comprehensive bacterial macro-evolutionary framework in order to estimate the
120 origin and evolution of the sporulation gene families, from an ancient bacterial
121 ancestor to the contemporary *B. subtilis* and *C. difficile.* Our analyses show common
122 evolutionary patterns for the two species, with two major gene gain events having
123 expanded the cell type-specific regulons. Moreover, per each gain event, the cell
124 type-specific regulons appear to grow in the same size-proportions, suggesting that
125 sporulation stages have been co-evolving along the bacterial phylogeny, whereas
126 multiple independent gene losses emerge in some branches of the tree for both
127 Bacilli and Clostridia lineages. We also show that in the evolution of *B. subtilis*, gene
128 gains more often involve the emergence of taxonomically restricted genes (TRGs)
129 than in *C. difficile*, where more sporulation genes have orthologous in other bacteria
130 and were likely acquired by horizontal gene transfers (HGTs). Still, the two
131 mechanisms are expected to occur to a greater or lesser extent in both lineages and
132 are fully demonstrated here for the genes coding the exosporium protein CdeA in *C.*

133 *difficile* (Díaz-González et al. 2015) and SpoIIIAF, which is a component of the
134 SpoIIIA-SpoIIQ sporulation-essential secretion system conserved across spore
135 formers (Meisner et al. 2008; Fimlaid et al. 2015; Morlot and Rodrigues 2018).
136 Overall, our analysis supports the hypothesis that sporulation emerged at the base of
137 Firmicutes, 2.5 billion years ago, with a first major gene gain event in response to
138 oxygen appearance. This major evolutionary step was followed by a series of
139 intermediate gains that culminated with a second major gene gain event
140 corresponding to the specialization of this developmental program and its adjustment
141 to particular ecosystems.
142

143 **Results**

144 A small core and large diversity of sporulation genes between *Bacillus subtilis* and
145 *Clostridioides difficile*

146 The species *Bacillus subtilis* is by far the best characterized spore-former in
147 terms of sporulation genes and mechanisms (Fawcett et al. 2000; Eichenberger et al.
148 2003; Wang et al. 2006; Mäder et al. 2012). *Clostridioides difficile* is the second
149 most-studied species and the best characterized within the class Clostridia (Fimlaid
150 et al. 2013; Saujet et al. 2013; Paredes-Sabja et al. 2014; Pishdadian et al. 2015).
151 For this study we have manually compiled 734 sporulation genes for *B. subtilis* strain
152 168 (Table S1, BSU) and 308 sporulation genes from the *C. difficile* strain 630 (Table
153 S1, CD). Each gene is under the direct control of one or more sporulation-specific
154 sigma factors and thus part of at least one sporulation regulon. The regulons have
155 different sizes (Table 1): $\boldsymbol{\sigma}^E$ is the largest followed by $\boldsymbol{\sigma}^G$ and $\boldsymbol{\sigma}^K$ (of similar sizes) and
156 $\boldsymbol{\sigma}^F$. Genes involved in sporulation and whose expression depends on Spo0A have
157 only been thoroughly identified for *B. subtilis* (Fawcett et al. 2000).

158 Homology mapping between *B. subtilis* and *C. difficile* genomes was carried
159 out using a bi-directional BLASTP approach, followed by selection of sporulation

160 gene clusters by regulon and presence/absence analysis of orthologous genes in
161 both species. In Figure 1 we show that only a small fraction of sporulation genes
162 have orthologous in the two species. Some of these genes are present in the same
163 regulon (Figure 1, orange, Table S2): 2 genes in $\boldsymbol{\sigma}^F$, 29 in $\boldsymbol{\sigma}^E$, 11 in $\boldsymbol{\sigma}^G$, and 5 in $\boldsymbol{\sigma}^K$
164 while less than 1 % belongs to different regulons (yellow, Figure 1, Table S3). A
165 larger proportion of genes corresponds to those present in both genomes but only
166 reported to be involved in sporulation by one species (Table S4). These proportions
167 vary depending on the regulon, in *B. subtilis* from 10 to 19% and in *C. difficile* from
168 12 to 35 %. Finally, the majority of sporulation genes have no candidate orthologous
169 in the two species (Table S5) (> 70% for BSU, > 50% for CD), highlighting the great
170 genetic diversity of the sporulation process. In order to elucidate the emergence of
171 this diversity in a larger context of bacterial evolution we extended the analysis of
172 presence/absence of sporulation genes to 258 bacterial genomes (Figure S1, Table
173 S6).
174

175 <u>Distribution of sporulation genes across bacteria reveals a small core of conserved</u>
176 <u>genes and lineage-specificity</u>

177 Presence/absence of sporulation genes was mapped across 258 bacterial
178 genomes (Figure S1, Table S6) using an optimized orthology mapping approach
179 (one *vs.* all and all *vs.* one) applied twice: first using *B. subtilis* 168 as reference,
180 second using *C. difficile* 630 (Figure S2, Figure S3). The results were represented in
181 super-matrices of presence (1 or more) and absence (0) of genes and used to
182 calculate the distribution of sporulation genes across bacterial genomes (Figure 2).
183 Both distributions show that only a small proportion of sporulation genes are
184 conserved in the bacterial genomes (Figure 2A). A unimodal distribution is observed
185 for sporulation genes from *B. subtilis* (D = 0.024757, p-value = 0.3179) with a peak
186 around the 20%, whereas for *C. difficile* sporulation genes the distribution is bimodal

187 (D = 0.05604, p-value = 7.262e-06) with high densities around 20% and a secondary
188 high-density peak at 42% (Figure 2B). These results suggest that *B. subtilis* has
189 more taxonomically restricted sporulation genes than *C. difficile*, whose genes
190 appear more frequently in the bacterial genomic dataset. To better elucidate the
191 causes underlying these differences we conducted an evolutionary analysis of
192 sporulation gene families by posterior probabilities in a phylogenetic birth-death
193 model.
194
195 A bacterial evolutionary backbone
196 Prior to the evolutionary analysis of sporulation gene families we started by
197 building a solid and comprehensive bacterial phylogeny based on 70 marker genes
198 conserved across the genomic dataset (Figure S1). Overall, the resulting tree
199 topology (Figure 3, Figure S4) is similar to recent macroevolutionary bacterial trees
200 (Hug et al. 2016; Marin et al. 2017). The phyla Dictyoglomi, Thermotogae, Aquificae,
201 Synergistetes, and Deinococcus-Thermus form an early diverging clade. The second
202 earliest split in the tree corresponds to the divergence of the taxon coined as
203 hydrobacteria (which includes Proteobacteria, Bacteroidetes, Spirochaetes,
204 Fusobacteria) and terrabacteria (which includes Actinobacteria, Cyanobacteria and
205 the Firmicutes), estimated to have occurred between 2.83-3.54 billion years ago (Ga)
206 (Battistuzzi et al. 2004; Battistuzzi and Hedges 2009). Within terrabacteria,
207 Firmicutes (monoderms, low GC, mostly gram-positive) is an early diverging phylum
208 currently divided into seven classes: Bacilli, Clostridia, Negativicutes,
209 Erysipelotrichia, Limnochordia, Thermolitobacteria and Tissierellia (source: NCBI
210 taxonomy). In our study we included genomes from four of these classes. The major
211 split is between Bacilli (n=77) and Clostridia (n=95) but other groups elevated to the
212 class level emerge within Clostridia, namely, the Negativicutes (n= 11) coding for
213 diderm bacteria with a gram-negative-type cell envelope and Erysipelotrichia (n=1).
214 Other two spore-formers (Figure S1), *Symbiobacterium thermophilum* and

215 *Thermaerobacter marianensis* (gram-positive, GC > 60%) appear as a distinct
216 lineage from the Firmicutes and have been suggested to be an intermediate phylum
217 between the high-GC (Actinobacteria) and the low GC (Firmicutes) groups (Ueda et
218 al. 2004; Han et al. 2010). Here, however, they remain as Firmicutes.
219

220 Two major gene gain events drive sporulation evolution in *Bacillus subtilis* and
221 *Clostridioides difficile*

222 Super matrices of presence/absence of sporulation genes (0,1, >1) sized
223 726x258 and 307x258 from references *B. subtilis* 168 and *C. difficile* 630,
224 respectively, were used in an evolutionary analysis of sporulation genes across the
225 bacterial phylogeny (Tables S7-S8, Figure 3). We first focused on the gene gain
226 events, occurred from the root to the terminal branches of *B. subtilis* (Figure 3A) and
227 *C. difficile* (Figure 3B). Our approach estimates that throughout evolution of both
228 species, sporulation genes were acquired in two major evolutionary steps. The first
229 major gene gain event occurred at the base of the Firmicutes, comprising at least
230 166 genes (branch 2, Figure 3, Table S9), including those coding for Spo0A and the
231 four cell type-specific sigma factors, as well as other core sporulation genes,
232 conserved among members of the phylum (Tables S9-S10). The second major gain
233 event occurred at terminal branches, at the base of the *B. subtilis* group (Berkeley et
234 al. 2002) (Figure 3A, +90 genes, branch 7, Figure 3A) as well as within members of
235 the family Peptostreptococcaceae (Clostridium cluster XI) (Figure 3B, +73 genes,
236 branch 10), at the branch that predates the split between the genera Clostridioides,
237 Paeniclostridium, Paraclostridium and Terrisporobacter (Galperin et al. 2016).
238 Intermediate gains also occurred in the evolutionary path of *B. subtilis* (branches 1, 3
239 to 6, 8 and 9) and *C. difficile*, in this case involving two additional gain events on
240 branches 11 (+21 genes) and 12 (+39 genes) (Figure 3).
241

242 Extensive gene loss in spore-forming and asporogenous lineages

243 Besides gene gains, also significant losses have occurred in certain Bacilli
244 and Clostridia lineages. In Bacilli, extensive gene loss (Figure 3A, >=90, red
245 triangles) was predicted for the branch ancestral to the asporogenous groups *Listeria*
246 and Lactobacillales, *Solibacillus* and Staphylococcaceae, *Exiguobacterium* and
247 Erysipelotrichia. In each of these groups, succeeding gene losses have occurred for
248 Staphylococcaceae, *Exiguobaterium*, *Listeria* and Lactobacillales. In addition,
249 multiple gene losses arose independently in other groups, including spore-formers,
250 for *e.g.*, at the base of Alicyclobacillaceae, of the *Geobacillus* and *Anoxybacillus*
251 genera and in some other *Bacillus* species (Figure 3A, orange triangles). In
252 Clostridia, extensive gene losses have occurred independently for *Candidatus*
253 *Arthromitus* and at the branch preceding the split of the genus *Filifactor* and
254 *Proteocatella* (Figure 3B, red triangles). Other significant losses are predicted for the
255 branches preceding the lineages Peptococcaceae and Thermoanaerobacterales
256 groups III and IV, Thermoanaerobacterales, Lachnospiraceae, *Ruminiclostridium* and
257 *Mageeibacillus*, and *Ruminococcus* (Figure 3B, yellow triangles).
258 Significant gains and losses were also predicted for some terminal branches of Bacilli
259 and Clostridia species/strains (Table S7-S8). However, gain/loss expectations on
260 terminal branches are highly prone to error because false positives/negatives, or
261 lower quality genomes, have greater impact on the total estimates. Thus, only events
262 occurring on internal branches were considered for discussion and presented in
263 Figure 3.
264

265 Co-evolution of early and late sporulation stages

266 Sporulation genes belong to at least one cell type-specific regulon and thus it
267 is possible to explore the effect of gene gain events in the size of the regulons as a
268 function of the evolutionary steps. Here we have focused on the gain events
269 occurring throughout the evolutionary paths of *B. subtilis* and *C. difficile*, highlighted

270 in green in Figure 3 (branches numbered from 1 to 12). The general pattern for *B.*
271 *subtilis* (Figure 4A) and *C. difficile* (Figure 4B) is that regulons are growing in similar
272 proportions. Larger increases in the size of the regulons corresponded to the major
273 gene gain events highlighted in Figure 3: at the base of Firmicutes (branch 2), at the
274 base of the *B. subtilis* group (branch 7) and within the family Peptostreptococcaceae
275 (branch 10). In these branches we could also observe slightly larger increases for the
276 regulons controlling late sporulation stages, $\boldsymbol{\sigma}^G$ (branch 2 and 10) and $\boldsymbol{\sigma}^K$ (branch 7)
277 (Figure 4). Still these differences are small and do not affect the general pattern in
278 which sporulation regulons seem to be co-evolving with gene gain dynamics until
279 reaching their current size.
280
281 Sporulation genes: a landmark of Firmicutes
282 Evolutionary analyses of sporulation genes from *B. subtilis* and *C. difficile*
283 estimate that 28,2% and 23,8% of their machinery was already present at the last
284 bacterial ancestor (Figure 3, Tables S9-S10). These genes are conserved across
285 early diverging bacterial clades and may have been co-opted for sporulation.
286 Subsequent gene gains occurred later in the bacterial phylogeny, at different
287 taxonomic levels and included TRGs, HGTs and duplications as illustrated in Figure
288 5. In particular, the first major gain event of sporulation genes (Figure 3, branch 2),
289 which correlates with the emergence of the Firmicutes as a phylum, involved a gain
290 of at least 166 genes, from which the vast majority, 80-90%, are TRGs, *i.e.* only
291 found in Firmicutes members (Figure 5, Tables S9-S10). A much smaller amount of
292 gene gains are estimated to be derived by HGT, 10,4-14,3% (Tables S9-S10), and
293 having orthologous in other lineages outside the Firmicutes.
294
295
296

297 *B. subtilis versus C. difficile*: sporulation driven by rapid divergence and horizontal
298 gene transfer

299 Besides the base of the Firmicutes we also estimated the origin of gene gains
300 at lower taxonomic ranks. In particular we chose branches in the phylogenetic tree
301 for which we could strictly define descendants and outgroups in agreement with the
302 NCBI taxonomy (release from May 2018). These ranks, shown in Figure 5, included:
303 branch 3 (at the base of the class Bacilli), branch 5 (Bacilli families but excluding the
304 early diverging Alicyclobacillaceae and Paenibacillaceae), branch 7 (at the base of
305 the *B. subtilis* group), branch 10 (within the family Peptostreptococcaceae - a
306 subgroup containing the genera *Clostridioides*, *Paeniclostridium*, *Paraclostridium* and
307 *Terrisporobacter*), branch 11 (at the base of the genus *Clostridioides*) and branch 12
308 (at the base of the species *C. difficile*).

309 At the base and within Bacilli, the estimated gains of TRGs were dominant
310 over HGTs, representing 61,3% (branch 3), 89,8% (branch 5) and 56,3 % (branch 7)
311 of the acquired genes (Figure 5A, in green). In contrast, genes acquired along the
312 clostridial lineages were more often present in other clades (Figure 5B, in blue) and
313 likely acquired by HGT, representing 55,6% in branch 10, 77,8% in branch 11 and
314 64,1% of the total gains in branch 12. Family expansions were significantly lower
315 when compared to other mechanisms of gene acquisition with a maximum of four
316 duplicated genes in branches 10 and 12. List of HGTs, TRGs and expansions by
317 gene duplication are summarized in Tables S9 for Bacilli and S10 for Clostridia.
318 Beyond the differences in rates of HGTs *vs.* TRGs and family expansion, we can find
319 evidence of both mechanisms occurring in Bacilli and Clostridia (Figure 5).

320

321 HGT and rapid divergence in Clostridia

322 In Clostridia we find CdeA, a protein involved in the exosporium assembly in
323 *C. difficile* (Díaz-González et al. 2015; Calderón-Romero et al. 2018) coded by a
324 gene gained in branch 10 (Entrez GI: 126699990, Table S10) which has orthologous

325 outside the Peptostreptococcaceae family, notably in species of the intestinal tract
326 (Figure S5). The orthologous sequences share in common two blocks of a four
327 cysteine conserved domain with unknown function (InterPro Ac. Nr.: IPR011437) and
328 together provide a clear example of HGT, likely to have occurred in the gut
329 ecosystem.

330 In turn *spoIIIAF* (Entrez GI: 126698793) has diverged to the point that no
331 longer clusters with other genes that code for Stage III sporulation protein AF,
332 forming two distinct orthologous clusters by our orthology mapping approach (Figure
333 6A) as well as in orthology databases EggNOG 4.5.1 and OMA: ENOG4105X2R and
334 TLCVVIV for *spoIIIAF*, ENOG410644B and IWEIGSE for *spoIIIAF* (Huerta-Cepas,
335 Szklarczyk, et al. 2016; Altenhoff et al. 2018). The gene is part of the *spoIIIA* locus,
336 conserved in all spore formers (Galperin et al. 2012; Morlot and Rodrigues 2018),
337 which codes for eight proteins, SpoIIIA to H, the components of a channel also
338 known as the "feeding tube" and whose main role is to connect the mother cell to the
339 forespore and transfer substrates required for σ^G activity in the forespore (Meisner et
340 al. 2008; Camp and Losick 2009). Because *spoIIIAF* (Figure 6B) is poorly conserved,
341 it is predicted here as an innovation 'gained' in branch 10, within the
342 Peptostreptococcaceae family (Table S10). The *spoIIIAF* cluster includes
343 orthologous from all strains of *C. difficile*, *C. magenotii, Paraclostridium bifermentans,*
344 *Paeniclostridium sordellii* and *Terrisporobacter glycolicus* (Figure 6A). An alignment
345 of SpoIIIAF orthologs from both clusters on Figure 6B shows two-conserved TM
346 domains (residues 1 - 57) and divergence in the remaining sequence, except for the
347 hydrophobic residues highlighted in blue. These residues are part of the hydrophobic
348 core of the protein which are associated with a conserved ring-building motif (RBM)
349 fold (Zeytuni et al. 2018). Since a structure for SpoIIIAF is available (*B. subtilis*, PDB:
350 6dcs) (Zeytuni et al. 2018), we simulated the 3D structure of SpoIIIAF from *C.*
351 *difficile* using UCSF Chimera (Pettersen et al. 2004) and Modeller (Eswar et al. 2006)
352 with 6dcs chain A as template. Simulations generated five reliable models (Figure

353 S6), which are structurally similar to SpoIIIAF and adopt $\alpha_1\beta_1\beta_2\alpha_2\beta_4$ topology (Figure
354 6C). Still, the high divergence rates between the two orthologs is reflected on the
355 distribution of charged and hydrophobic residues on the protein surface (Figure 6C,
356 Figure S6 B) and suggests that they have been under different evolutionary
357 constraints.
358

359 **Discussion**

360 In this work we have extended previous comparative genomics analyses
361 based on *B. subtilis* sporulation genes (Paredes et al. 2005; de Hoon et al. 2010;
362 Galperin et al. 2012; Abecasis et al. 2013; Galperin 2013) by including the
363 sporulation genes identified in the human pathogen *Clostridioides difficile*. Recent
364 advances in *C. difficile* genetics and the ability to construct sporulation mutants in
365 this pathogenic anaerobe have enabled the identification of over 300 genes, directly
366 regulated by the cell type-specific sigma factors of sporulation (Fimlaid et al. 2013;
367 Saujet et al. 2013; Pishdadian et al. 2015) (Table S1). By comparing the sporulation-
368 specific regulons between the two species we identified four main groups of genes
369 based on their orthology/paralogy relationships. The largest group, representing over
370 50% of each regulon in both species, consist of genes only present in one species
371 (Figure 1, Table S5). These genes highlight the existing diversity and complex origins
372 of the sporulation machinery, genes coding for the regulatory complex
373 SpoIVFA/SpoIVFB/BofA present in the forespore membrane of *B. subtilis* and
374 involved in the activation of σ^K (Rudner 2002), are absent in the *C. difficile* genome
375 (Pereira et al. 2013), as well as in other Clostridia species (Galperin et al. 2012). In
376 some cases, genes having the same function in both species (*i.e.* functional
377 equivalents) are annotated with the same name but have distinct evolutionary paths,
378 with different ancestors, as is the case of the forespore proteins SpoIIQ. These
379 proteins form a complex with SpoIIIAH through which the mother cell is thought to

380 convey to the forespore metabolites essential for maintaining its metabolic potential
381 and regulating engulfment in *B. subtilis* (Camp and Losick 2008; Meisner et al. 2008)
382 and *C. difficile* (Fimlaid et al. 2015; Serrano et al. 2016). Despite having the same
383 name and similar roles in both species, they are non-orthologous. Both SpoIIQ have
384 a catalytic peptidase M23 motif (InterPro Ac. Nr.: IPR016047) and a transmembrane
385 domain but distinct N-terminal sides, being a recognized case of non-orthologous
386 gene displacement (Galperin et al. 2012). The second largest group of genes
387 corresponds to those present in the two genomes but only described as part of the
388 sporulation regulons in one species (Figure 1, Table S4). These genes are involved
389 in other processes and may have been co-opted for roles in sporulation. For instance
390 the genes *dapA* and *dapB*, which are present in many bacterial phyla and involved in
391 the lysine biosynthesis pathway (Mukherjee et al. 2017), have been recruited to the
392 sporulation process in *B. subtilis*, with *dapA* being part of the spoVF operon,
393 expressed in the mother cell and involved in the synthesis of dipicolinic acid (DPA)
394 (Chen et al. 1993; Daniel and Errington 1993; Steil et al. 2005), a compound that is
395 characteristic of spores (Slieman and Nicholson 2001; Moeller et al. 2014). The
396 same genes *dapA* and *dapB* in *C. difficile* are located in the *lysA-asd-dapA-dapB*
397 operon (Mukherjee et al. 2017) and were shown to be dispensable for spore
398 formation (Donnelly et al. 2016) whereas a study on *C. perfringens* suggests a more
399 ancestral mechanism for DPA synthesis, involving the electron transfer flavoprotein
400 Etfa (Orsburn et al. 2010). In the same group, the gene *uppS* (undecaprenyl
401 pyrophosphate synthetase) is involved in polysaccharide biosynthesis and part of σ^G
402 regulon in *C. difficile* (Saujet et al. 2013) but a role in *B. subtilis* sporulation has not
403 been determined. Finally, there are two groups of orthologous genes in *B. subtilis*
404 and *C. difficile*. One group corresponds to genes present in the same regulon in both
405 species: 2 in σ^F, 29 in σ^E, 11 σ^G and 5 in σ^K (Table S2). For example the genes
406 *spmA* and *spmB*, under the control of σ^E and involved in spore core dehydration
407 (Popham et al. 1995; Steil et al. 2005; Pereira et al. 2013; Saujet et al. 2013). Other

408 examples include the genes of the *spoIIIA* locus, also controlled by σ^E whose
409 products are part of the SpoIIQ-SpoIIIAH channel (Illing and Errington 1991; Camp
410 and Losick 2008; Fimlaid et al. 2015), and *spoIID* (σ^E) required for septal
411 peptidoglycan hydrolysis at the onset of the engulfment process and for the proper
412 localization of SpoIIQ (Rodrigues et al. 2013; Serrano et al. 2016). In the σ^G regulon,
413 the *pdaA* gene*,* coding for a N-acetylmuramic acid deacetylase, is involved in spore
414 cortex peptidoglycan biosynthesis in both species (Fukushima et al. 2002; Saujet et
415 al. 2013; Diaz et al. 2018). Still, despite being orthologs and regulated by the same
416 sigma factors, these genes may show different modes of action in the two species
417 through the link to non-homologous intermediates (Fimlaid et al. 2015; Serrano et al.
418 2016). At last, there are also orthologous pairs that belong to different regulons in the
419 two organisms (Table S3). For example, the *seaA*, involved in the spore envelope
420 assembly (Steil et al. 2005), is expressed early in the forespore, under the control of
421 σ^F in *B. subtilis*, but is only activated following engulfment completion, under the
422 direction of σ^G in *C. difficile* (Saujet et al. 2013). Also, *yraF,* coding for an
423 uncharacterized protein similar to a spore coat protein, is transcribed in the forespore
424 under the control of σ^G in *B. subtilis* (Wang et al. 2006), while in *C. difficile* is
425 transcribed in the mother cell under the control of σ^E (Saujet et al. 2013).

426 We also find that some sporulation genes are conserved across several
427 bacterial phyla, or even extended to other domains of life. For *e.g.* spoVR
428 (orthologous group Nr.: COG2719) (Huerta-Cepas, Szklarczyk, et al. 2016) is
429 present in Proteobacteria and Archaea. The gene *cotR*, with a conserved hydrolase
430 domain (InterPro Ac. Nr.: IPR016035) and structure, has a broader taxonomic profile,
431 which includes eukaryotes (orthologous group Nr.: COG3621) (Huerta-Cepas,
432 Szklarczyk, et al. 2016). Still, genes of broader taxonomic range constitute less than
433 one third of the sporulation machinery (Figure 3, Tables S9-S10). The great majority
434 of sporulation genes were acquired in the first major gain event, coinciding with the
435 origin of the Firmicutes (Figure 3) and including the genes of the so-called

436 sporulation signatures based on *B. subtilis* genes (Abecasis et al. 2013) and spore-
437 formers from the gut microbiota (Browne et al. 2016). Since the divergence of
438 Firmicutes is estimated to have occurred between 2.5 to 3.0 billion years ago (Ga),
439 upon the emergence of terrestrial ecosystems (2.6 Ga) and after the Great Oxidation
440 Event (2.3 Ga) (Battistuzzi et al. 2004; Battistuzzi and Hedges 2009; Marin et al.
441 2017) it has been suggested that sporulation is part of an adaptive process to cope
442 with O_2 exposure (de Hoon et al. 2010; Abecasis et al. 2013). Gain events occurring
443 after the origin of sporulation, and highlighted in Figure 3, correspond to lineage
444 specializations of the program. The fact that we only find two larger gain events (≥90)
445 along the evolutionary paths of *B. subtilis* and *C. difficile*, suggests that sporulation
446 has evolved in punctual moments for these two species, only tied to certain lineages.
447 Moreover, we find no evidence of these gains having greater influence in the size of
448 a specific regulon since they have all co-evolved proportionally (Figure 4). This and
449 the fact that the four sporulation-specific sigma factors were present in the last
450 common ancestor of Firmicutes (Table S9, Table S10) suggest that both early and
451 late sporulation stages were defined upon the emergence of the process.

452 Though the common macroevolutionary patterns in both lineages, our
453 analysis to detect possible origins of gene gains, suggests different modes of gene
454 acquisition. While in *B. subtilis* most of the genes gained in branches 3, 5 and 7 are
455 taxon-specific (Figure 5A, Table S9), in the evolutionary path of *C. difficile*, genes
456 were mostly acquired by horizontal gene transfer events, with a consistently smaller
457 proportion of taxonomically restricted genes (Figure 5B). This finding is in line with
458 the highly mobile genome of *C. difficile* 630, where 11% of mobile genetic elements
459 (mainly transposons) are the main cause for acquisition of foreign genes from other
460 species in the gut microbiota (Sebaihia et al. 2006; He et al. 2010). Genes acquired
461 by HGT are involved in antibiotic resistance, virulence and sporulation, conferring *C.*
462 *difficile* adaptive advantages in the gut ecosystem. A case-study emerging from our
463 analysis is the *cdeA* gene, coding for an exosporium protein (Díaz-González et al.

464 2015) only found in Clostridia, especially in species inhabiting gut ecosystems
465 (clusters ENOG410XVF3 and HESKKIA from EggNOG 4.5.1 and OMA databases).
466 A phylogenetic tree based on the alignment of 35 CdeA sequences (Figure S5) does
467 not reflect the species phylogeny suggesting multiple HGT events within the gut
468 environment and, to a lesser extent, from the interaction of clostridia species of
469 different niches, *i.e.* sludge/sediments and the gut.

470 In turn *B. subtilis* shows a stronger lineage-specific specialization of the
471 program, with genes that are only found in closely related species (Figure 5A, Table
472 S9). The appearance of a higher number of such genes in members of the *Bacillus*
473 *subtilis* group, may be due to the higher homologous recombination rates observed
474 for Bacilli species when compared to *C. difficile* (Vos and Didelot 2009; He et al.
475 2010). A key feature of *B. subtilis* and close relatives is their natural competence, *i.e.*
476 the ability to enter a physiological state in which cells are able to take up exogenous
477 DNA fragments (Dubnau 1991; Haijema et al. 2001; Smits et al. 2005; Chen et al.
478 2007). Although the consequences of natural competence on the genomic
479 composition and species diversity remain to be fully elucidated, this mechanism of
480 DNA uptake as been proposed as a main driver of genome diversity within the
481 *Bacillus subtilis* group (Brito et al. 2018) and to facilitate homologous recombination
482 (Vos 2009; Mell and Redfield 2014).

483 The other mechanism giving rise to TRGs are high mutation rates that
484 saturate the signal of homology. Rapid divergence between orthologs may occur
485 driven by exposure to particular lineage-specific environment/niches and to different
486 protein interactions/binding partners. This latter is illustrated here for the counterparts
487 of SpoIIIAF from members of the Peptostreptococcaceae family (SpoiIIIAF) (Figure
488 6A). This protein is a component of the feeding tube, conserved in spore-formers,
489 connecting the mother cell to the forespore, which together with other proteins of the
490 spoIIIA operon: SpoIIIAB, SpoIIIAD, SpoIIIAE, SpoIIIAG form a multimeric membrane
491 complex with a ring shape located in the mother cell membrane and extending to the

492 intermembrane space between mother cell and forespore (Doan et al. 2009).
493 Although the role of SpoIIIAF in the complex has not yet been fully elucidated, the
494 protein contains a Ring-Building motif that may support ring formation in a multimeric
495 form, similar to SpoIIIAG and SpoIIIAH (Levdikov et al. 2012; Zeytuni et al. 2017;
496 Zeytuni et al. 2018). Like in *B. subtilis* and other spore-formers, SpoiIIIAF in *C.*
497 *difficile* preserves the N-terminal transmembrane region and the RBM hydrophobic
498 motif (Figure 6B). Still, the sequence as diverged considerably, only sharing 16% of
499 sequence identity with its counterpart in *B. subtilis* (Serrano et al. 2016). This
500 divergence is reflected in changes in the overall electrostatic potential and
501 hydrophobicity on the proteins surface (Figure 6C, Figure S6). Since SpoIIIAF is
502 predicted to be located in the inner mother cell membrane towards the cytoplasm
503 (Doan et al. 2009; Zeytuni et al. 2017; Morlot and Rodrigues 2018), the high
504 sequence divergence between SpoiIIIAF and SpoIIIAF suggests links with distinct
505 proteins or binding partners in the mother cell cytoplasm that need to be further
506 explored for Bacilli as well as for the members of the Peptostreptococcaceae family.

507 Besides gene-gains there were also extensive gene losses occurring
508 independently throughout the bacterial phylogeny (Figure 3). These losses have
509 resulted in the emergence of asporogenous lineages (*i.e.* non-spore-forming
510 Firmicutes, (Onyenwoke et al. 2004)) among spore-formers at different taxonomic
511 levels. Within Bacilli, we find the group including *Listeria*, Lactobacillales,
512 Staphylococcaceae, *Solibacillus silvestris* StLB046 and *Exiguobacterium*; in
513 Clostridia, the groups including the species *Proteocatella sphenisci* and *Filifactor*
514 *alocis*, members of the family Lachnospiraceae and the genus Caldicellulosiruptor
515 (within Thermoanaerobacterales group III) (Figure S1). Although the ability to form
516 endospores is often seen has an evolutionary advantage, the process is energetically
517 costly, involving the expression of hundreds of genes and taking up several hours to
518 complete (Fujita and Losick 2005; Paredes et al. 2005; Fimlaid et al. 2013). For
519 example, in nutrient-rich environments after 6,000 generations of *B. subtilis*, neutral

520 processes but also selection were found to facilitate loss of sporulation (Maughan et
521 al. 2007), occurring by indels or single-nucleotide substitutions (Maughan et al.
522 2009). In nature though, sporulation loss is mostly driven by consecutive gene losses
523 as suggested earlier (Galperin et al. 2012; Galperin 2013) and demonstrated in
524 Figure 3. Under relaxed environments for sporulation, emerging asporogenous
525 bacteria will have less genes, be more energy efficient and grow faster (Maughan
526 and Nicholson 2011; Nicholson 2012; Maitra and Dill 2015). Asporogenous
527 Firmicutes have developed alternative ways to persist in diverse (Sauders et al.
528 2012; Linke et al. 2014) and sometimes extreme environments, as is the case of the
529 extremophiles *Exiguobacterium* (Vishnivetskaya and Kathariou 2005; Vishnivetskaya
530 et al. 2009; Moreno-Letelier et al. 2012) and *Caldicellulosiruptor* (Van De Werken et
531 al. 2008; Hamilton-Brehm et al. 2010). The two *Exiguobacterium* species used here,
532 *E. antarcticum* and *E. sibiricum,* have only 21% of the sporulation signature
533 (Abecasis et al. 2013) while *Caldicellulosiruptor* species have between 62.5 and
534 73%, so there is still some speculation as to whether the latter form spores or not
535 (Abecasis et al. 2013). The same question has persisted for sometime for
536 *Ruminococcus* species (Galperin et al. 2012; Abecasis et al. 2013). *Ruminococcus*
537 sp. have lost some of the conserved sporulation genes (Figure 3B) and have
538 between 48 and 60% of the sporulation signature (Abecasis et al. 2013) (Figure S1,
539 Figure S3), in what appears to be an intermediary stage into becoming
540 asporogenous. Still, they were recently shown to sporulate (Browne et al. 2016;
541 Mukhopadhya et al. 2018), and are likely to spread as spore-like structures between
542 hosts (Schloss et al. 2014; Mukhopadhya et al. 2018), which plays as a counter
543 evolutionary force to the emergence of asporogenous phenotypes. In other cases,
544 notably *Epulopiscium* and *Candidatus Arthromitus*, sporulation has evolved as a
545 reproductive mechanism, involving some gene loss (Figure 3B) and eventually the
546 gain of new genes (Flint et al. 2005; Galperin 2013), under distinct evolutionary
547 forces.

548

549 ## Conclusions

550 In this study we were able to combine the comprehensive knowledge of
551 sporulation genes in *B. subtilis* with the emerging model *C. difficile*. Our analysis
552 validates earlier studies in that sporulation emerged once, at the base of Firmicutes.
553 From this first major gene gain event, we have identified intermediate evolutionary
554 steps combined with a second major lineage-specific gain event that underlies the
555 current specialization and diversity observed among *B. subtilis* and *C. difficile*
556 sporulation programs. Furthermore, predominant mechanisms of gene acquisition
557 and innovation differ in the lineages of the two species: *B. subtilis* entails greater
558 innovation with many TGRs while *C. difficile* has acquired new genes mostly through
559 HGTs. We also show that extensive gene losses underlie the emergence of
560 asporogenous lineages and the co-option of sporulation as a reproductive
561 mechanism. Overall our comparative genomics approach provides a framework with
562 which to elucidate the evolution and diversity observed in the sporulation program
563 across Firmicutes and unveils new evolutionary case studies, such as those
564 described for the genes *cdeA* and *spoIIIAF*.

565

566 ## Methods

567 <u>Datasets construction</u>

568 All known sporulation genes from *Bacillus subtilis* strain 168 and
569 *Clostridioides difficile* strain 630, the two species that are better characterized in
570 terms of their sporulation machinery, were collected from the literature (Fimlaid et al.
571 2013; Saujet et al. 2013; Pishdadian et al. 2015; Meeske et al. 2016) and SubtiWiki
572 2.0 (Mäder et al. 2012). These included the genes coding for the main regulators of
573 sporulation: spo0A - controlling initiation; sigE and sigK - coding for the mother cell

574 early and late specific sigma-factors; sigF and sigG - coding for the early and late
575 forespore-specific sigma factors; and the genes under their control (Table S1).
576 Bacterial genomes (n=258) were collected from the NCBI Reference Sequence
577 Database (release from April 2017) including species from the classes Actinobacteria
578 (15), Aquificae (2), Bacteroidetes (3), Cyanobacteria (3), Dictyogiomi (2), Firmicutes
579 (186), Fusobacteria (4), Proteobacteria (30), Spirochaetes (6), Synergistetes (1),
580 Thermotogae (5), Deinococcus-Thermus (1) (Figure S1, Table S6).

581

582 Orthology/paralogy mapping and selection of homologous clusters

583 The predicted proteins from each genome were categorized into homologous
584 clusters using the bidirectional best hits (BDBH) algorithm implemented in
585 GET_HOMOLOGUES (Contreras-Moreira and Vinuesa 2013) (Figure S2). The
586 software is built on top of BLAST+ (Camacho et al. 2009). First, inparalogues,
587 defined as intraspecific BDBHs, are identified in each genome. Second, new
588 genomes are incrementally compared to the reference genome and their BDBHs
589 annotated. The main script ran with BLASTP default options and a minimum
590 coverage fixed at 65% for the pairwise alignments (-C), option -s to save memory,
591 and reporting at least clusters with two sequences (-t set at 2). The choice of the
592 BDBH algorithm and corresponding cutoffs resulted from prior validation with sets
593 and subsets of the bacterial genomes used in this study (Figure S3). The method
594 was applied twice, the first time with *Bacillus subtilis* strain 168 (BSU) genome fixed
595 as reference (-r) and the second time with *Clostridioides difficile* strain 630 (CD)
596 (Figure S2). Homology mapping resulted in over 3500 clusters of homologous genes
597 for references BSU and CD. Clusters corresponding to sporulation genes were
598 selected making a total of 726 and 307 clusters for BSU and CD, respectively. The
599 results were represented in two large-scale profile matrices of presence/absence of
600 genes (0 for absence, 1 presence or >1 for multiple presence). The two reference

601 genomes were also compared between each other using the orthoMCL algorithm
602 implemented in GET_HOMOLOGUES. The main script ran with BLASTP default
603 options, -C set at 65 % and reporting all clusters (-t 0).

604

605 Bacterial phylogenies

606 Seventy gene clusters of single-copy orthologous conserved across the
607 genomic dataset (adk, alaS, aspS, coaD, cysS, efp, engA, fabG, fmt, frr, gcp, gltX,
608 guaA, infA, infC, ksgA, miaA, mraA, mraW, mraY, murC, murD, murE, murG, obgE,
609 pheS, pheT, prfA, pth, pyrH, rluD, rpe, rplA-F, rplI-L, rplN-P, rplR-X, rpmA, rpoA,
610 rpsB-E, rpsH, rpsJ, rpsK, rpsM, rpsQ, rpsS, ruvB, uppS, uvrB, yabC, yabD, ylbH,
611 yyaF) were assembled with good occupancy; allowing for a maximum of 5
612 sequences missing per cluster. Clusters were aligned independently using MAFFT
613 (Katoh and Standley 2013) or MUSCLE (Edgar 2004) under the default options. The
614 best alignment per cluster was selected and trimmed using trimAl (Capella-Gutiérrez
615 et al. 2009) with the -automated1 option. The 70 alignments were concatenated into
616 a single super-alignment comprising 258 species/strains and 14401 amino-acid
617 positions. The alignment was used to construct phylogenetic trees by maximum
618 likelihood and Bayesian methods (Figure S4). Calculations and tree construction
619 were performed with the RaxML 8 (Stamatakis 2014) (-N 100) and Mr. Bayes 3.2.5
620 (Ronquist et al. 2012) using the best protein evolution model (LG+G) selected with
621 ProtTest version 3 [9]. Tree topologies generated with both methods were compared
622 using ete-compare (Huerta-Cepas, Serra, et al. 2016) based on the normalized
623 Robinson-Foulds distance (nRF). The RF scores are the total number of leaf
624 bipartitions occurring in one but not in both trees. Since non-normalized measures of
625 phylogenetic distance typically increase with the number of leaves in the tree, we
626 used the normalized version, nRF, which divides RF by the maximum possible RF

627 value for *n* leaves (maxRF). For topologically identical trees nRF will be 0. Our nRF
628 of 0.03 (RF/maxRF = 16.0/512.0) indicates similar tree topologies.

629

630 Phylogenetic profiles and evolutionary analysis of sporulation genes

631 Sporulation gene clusters were represented in two large-scale profile
632 matrices of presence/absence, sized 726x258 and 307x258, for references BSU and
633 CD. The profiles were combined with the bacterial phylogeny to estimate gain (0 ->
634 1), loss (1 -> 0) and duplication events (1 -> 2 or more) in every tree branch with the
635 software Count v 10.04 (Csurös 2010). First, rates were optimized using the gain–
636 loss–duplication model implemented in Count, allowing different gain–loss and
637 duplication–loss rates in all lineages and rate variation across gene families (1 to 4
638 gamma categories). One hundred rounds of optimization were computed. Next, a
639 gene family history analysis was computed by posterior probabilities based on a
640 phylogenetic birth-and-death model. Tables with the posterior probabilities based on
641 the two profile matrices (Tables S7-S8) were extracted for further analysis with
642 custom Python scripts to estimate total number of gene gains, losses and
643 duplications (per branch), and gene presences (per node).

644

645 Estimating origins of gene gains

646 To estimate the source of gene gains only events with a posterior probability
647 of ≥ 0.5 by Count (Csurös 2010) were considered. Gene gains were classified as
648 'taxon-specific' (TRGs) when present at a given branch, present in its descendants
649 but not present in other clades. Alternatively, they were considered 'non-specific'
650 when gained at a certain branch, present in its descendants but also present in other
651 clades of the phylogeny. The later included genes gained by HGT in both directions
652 (from and to the branch/clade), since it is often challenging to identify the donor
653 (Koonin et al. 2001), specially when dealing with large genomic sets. At last, a small

654 fraction of sporulation genes was also gained by gene duplication and classified as
655 'family expansions'.
656 Since our ancestral reconstruction was limited to 258 bacterial genomes we
657 needed to validate genes classified as 'taxon-specific' in a broader taxonomic
658 context. To do so, we performed searches against the NCBI NR database (released
659 in January 2018, Cut-offs: E-value 10E-10, 75% query coverage with 50% identity)
660 using BLASTP and checked for their best hits outside the clade and corresponding
661 taxonomic distribution. While using this approach we were confronted with several
662 false positives in the NCBI nr database that were removed from the BLASTP reports
663 either because of misidentified species, for *e.g.* a genome assigned to the species
664 *Varibaculum timonense*, from Actinobacteria (Guilhot et al. 2017), that is in fact a
665 Clostridia, or because of recurring genomic contaminations of spore-formers in
666 genome assemblies of other species as was shown for *Mycobacterium* (Ghosh et al.
667 2009; Traag et al. 2010). Genes gained in the braches 2, 3, 5, 7, 10, 11 and 12 and
668 classified as 'taxon-specific', 'non-specific' or expansions are listed in Tables S9 and
669 S10 with corresponding cluster names and accession numbers.
670

671 **Acknowledgements**

672 We thank José Pereira Leal for advice with the bioinformatics methods and analyses.
673 This work was supported by FCT, "Fundação para a Ciência e a Tecnologia",
674 Portugal through grants SFRH/BPD/103171/2014 to PRS, PEst-
675 OE/EQB/LA0004/2011 to AOH and by program IF (IF/00268/2013/CP1173/CT0006)
676 to MS. This work was also financially supported by project LISBOA-01-0145-FEDER-
677 007660 ("Microbiologia Molecular, Estrutural e Celular") funded by FEDER funds
678 through COMPETE2020 – "Programa Operacional Competitividade e
679 Internacionalização".
680

681 **Authors' Contributions**

682 PRS designed the study, collected datasets, implemented the bioinformatics
683 methods, performed analyses, interpreted the data and wrote the manuscript. MS
684 provided the manually curated sets of sporulation genes and contributed with
685 scientific input for the manuscript. AOH designed the study, interpreted the data and
686 co-wrote the manuscript.

687

688 **Tables**

689 Table 1 – Sporulation genes collected in the literature by regulon.

Sporulation stage	Expression	Regulon	# Genes *B. subtilis* 168	# Genes *C. dificille* 630
Initiation	Mother cell	Spo0A	65	-
Engulfment	Forespore	σ^F	81	39
Engulfment	Mother cell	σ^E	335	204
Maturation	Forespore	σ^G	156	60
Maturation	Mother cell	σ^K	151	56

690

691

692 **Figure Legends**

693 **Fig.1.** Orthology mapping between *B. subtilis* and *C. difficile* for genes under the
694 control of sporulation sigma factors. The fraction of genes with orthologous in the
695 same regulon is marked in orange. The fraction of genes with orthologous in different
696 sporulation regulona is in yellow. The fraction of genes without orthologous in the
697 regulons but with orthologous in the genome is in light purple. Fraction of genes
698 without orthologous is in dark purple.

699

700 **Fig.2.** Distribution of sporulation genes from *B. subtilis* and *C. difficile* across the 258
701 bacterial genomes sampled from NCBI. Distributions are represented in (A) Absolute
702 number of genes and (B) Relative percentages. Distributions were tested for
703 unimodality using the Hartigan's Dip test in R (D = 0.024757, p-value = 0.3179 for *B.*
704 *subtilis*; D = 0.05604, p-value = 7.262e-06 for *C. difficile*).

705

706 **Fig.3.** Sporulation gene gain and loss events across the bacterial phylogeny for (A)
707 Bacilli and (B) Clostridia. The total number of sporulation genes present at the root
708 and at the referential tips is highlighted in grey. Gain events are numbered from 1 to
709 12. Multigene maximum likelihood (RAxML) tree inferred from an alignment of 70
710 orthologs and corresponding bootstraps measures in %. The PROTGAMMALG
711 evolutionary model was used to infer the tree with branch support estimated with 100
712 bootstrap replicates.

713

714 **Fig.4.** Sporulation regulons increase in size represented from root to tips for (A) *B.*
715 *subtilis* and (B) *C. difficile*. Sporulation regulons have diverse sizes being σ^E the
716 largest regulon and also the one acquiring the largest number of genes in every
717 evolutionary step, to suppress this effect the gains per regulon were normalized by
718 size.

719
720 **Fig.5.** Predicted categories of sporulation gene gains in four branches for (A) *B.*
721 *subtilis* and (B) *C. difficile.* Gene gains were considered to be taxon-specific (green)
722 when gained at the branch, present in its descendants but not present or gained in
723 other clades. Alternatively, gene gains were considered non-taxon specific (dark
724 blue) when gained along a certain branch but also present in other non-descendent
725 clades. The later included genes gained by HGT in both directions (from and to the
726 branch/clade). Family expansions (light blue) were originated by duplication events.

727

728 **Fig.6.** Comparative analysis between stage III sporulation proteins AF: SpoIIIAF from
729 *B. subtilis* and SpoiIIIAF from *C. difficile*. (A) Orthologous clusters of SpoIIIAF and
730 SpoiIIIAF (n=64) were merged and aligned using MAFFT (Katoh and Standley 2013)
731 followed by a principal component analysis as implemented in Jalview (Waterhouse
732 et al. 2009) to access sequence divergence between clusters. (B) Selected
733 sequences clustering with SpoIIIAF (black square) and SpoiIIIAF (purple square)
734 were aligned with T-coffee (Notredame et al. 2000). (C) Aligned structures of the
735 monomeric forms of $SpoIIIAF_{85-206}$ (grey) and $SpoiIIIAF_{84-205}$ (orange) were
736 determined using Modeller homology-based modeling (Eswar et al. 2006) with
737 PDB:6dcs as template and MUSCLE (Edgar 2004) as sequence aligner. Structures
738 and electrostatic charge distributions were visualized with Chimera (Pettersen et al.
739 2004).

740

741 **References**

742 Abecasis AB, Serrano M, Alves R, Quintais L, Pereira-Leal JB, Henriques AO. 2013.
743 A Genomic Signature and the Identification of New Sporulation Genes. J.
744 Bacteriol. 195:2101–2115.

745 Abt MC, McKenney PT, Pamer EG. 2016. Clostridium difficile colitis: pathogenesis
746 and host defence. Nat. Rev. Microbiol. 14:609–620.

747 Altenhoff AM, Glover NM, Train C-M, Kaleb K, Warwick Vesztrocy A, Dylus D, de
748 Farias TM, Zile K, Stevenson C, Long J, et al. 2018. The OMA orthology
749 database in 2018: retrieving evolutionary relationships among all domains of life
750 through richer web and programmatic interfaces. Nucleic Acids Res. 46:D477–
751 D485.

752 Angert ER, Losick RM. 1998. Propagation by sporulation in the guinea pig symbiont
753 Metabacterium polyspora. Proc. Natl. Acad. Sci. U. S. A. 95:10218–10223.

754 Battistuzzi FU, Feijao A, Hedges SB. 2004. A genomic timescale of prokaryote

755 evolution: insights into the origin of methanogenesis, phototrophy, and the
756 colonization of land. BMC Evol. Biol. 4:44.
757 Battistuzzi FU, Hedges SB. 2009. A major clade of prokaryotes with ancient
758 adaptations to life on land. Mol. Biol. Evol. 26:335–343.
759 Bauer MP, Notermans DW, van Benthem BH, Brazier JS, Wilcox MH, Rupnik M,
760 Monnet DL, van Dissel JT, Kuijper EJ. 2011. Clostridium difficile infection in
761 Europe: a hospital-based survey. Lancet 377:63–73.
762 Berkeley R, Heyndrickx M, Logan N, De Vos P. 2002. Applications and Systematics
763 of Bacillus and Relatives. (Berkeley R, Heyndrickx M, Logan N, De Vos P,
764 editors.). Oxford, UK: Blackwell Science Ltd.
765 Brito PH, Chevreux B, Serra CR, Schyns G, Henriques AO, Pereira-Leal JB. 2018.
766 Genetic Competence Drives Genome Diversity in Bacillus subtilis. Genome Biol.
767 Evol. 10:108–124.
768 Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, Goulding D,
769 Lawley TD. 2016. Culturing of 'unculturable' human microbiota reveals novel
770 taxa and extensive sporulation. Nature 533:543–546.
771 Calderón-Romero P, Castro-Córdova P, Reyes-Ramírez R, Milano-Céspedes M,
772 Guerrero-Araya E, Pizarro-Guajardo M, Olguín-Araneda V, Gil F, Paredes-
773 Sabja D. 2018. Clostridium difficile exosporium cysteine-rich proteins are
774 essential for the morphogenesis of the exosporium layer, spore resistance, and
775 affect C. difficile pathogenesis. PLoS Pathog. 14:e1007199.
776 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.
777 2009. BLAST plus: architecture and applications. BMC Bioinformatics 10:1.
778 Camp AH, Losick R. 2008. A novel pathway of intercellular signalling in Bacillus
779 subtilis involves a protein with similarity to a component of type III secretion
780 channels. Mol. Microbiol. 69:402–417.
781 Camp AH, Losick R. 2009. A feeding tube model for activation of a cell-specific
782 transcription factor during sporulation in Bacillus subtilis. Genes Dev. 23:1014–

783 1024.
784 Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: A tool for
785 automated alignment trimming in large-scale phylogenetic analyses.
786 Bioinformatics 25:1972–1973.
787 Chen NY, Jiang SQ, Klein DA, Paulus H. 1993. Organization and nucleotide
788 sequence of the Bacillus subtilis diaminopimelate operon, a cluster of genes
789 encoding the first three enzymes of diaminopimelate synthesis and dipicolinate
790 synthase. J. Biol. Chem. 268:9448–9465.
791 Chen XH, Koumoutsi A, Scholz R, Eisenreich A, Schneider K, Heinemeyer I,
792 Morgenstern B, Voss B, Hess WR, Reva O, et al. 2007. Comparative analysis of
793 the complete genome sequence of the plant growth-promoting bacterium
794 Bacillus amyloliquefaciens FZB42. Nat. Biotechnol. 25:1007–1014.
795 Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software
796 package for scalable and robust microbial pangenome analysis. Appl. Environ.
797 Microbiol. 79:7696–7701.
798 Csurös M. 2010. Count: Evolutionary analysis of phylogenetic profiles with parsimony
799 and likelihood. Bioinformatics 26:1910–1912.
800 Daniel RA, Errington J. 1993. Cloning, DNA sequence, functional analysis and
801 transcriptional regulation of the genes encoding dipicolinic acid synthetase
802 required for sporulation in Bacillus subtilis. J. Mol. Biol. 232:468–483.
803 Díaz-González F, Milano M, Olguin-Araneda V, Pizarro-Cerda J, Castro-Córdova P,
804 Tzeng SC, Maier CS, Sarker MR, Paredes-Sabja D. 2015. Protein composition
805 of the outermost exosporium-like layer of Clostridium difficile 630 spores. J.
806 Proteomics 123:1–13.
807 Diaz OR, Sayer C V., Popham DL, Shen A. 2018. Clostridium difficile Lipoprotein
808 GerS Is Required for Cortex Modification and Thus Spore Germination.
809 mSphere 3:e00205-18.
810 Dick GJ, Lee YE, Tebo BM. 2006. Manganese(II)-oxidizing Bacillus spores in

811 Guaymas Basin hydrothermal sediments and plumes. Appl. Environ. Microbiol.
812 72:3184–3190.
813 Doan T, Morlot C, Meisner J, Serrano M, Henriques AO, Moran CP, Rudner DZ.
814 2009. Novel Secretion Apparatus Maintains Spore Integrity and Developmental
815 Gene Expression in Bacillus subtilis. PLoS Genet. 5:e1000566.
816 Donnelly ML, Fimlaid KA, Shen A. 2016. Characterization of Clostridium difficile
817 Spores Lacking Either SpoVAC or Dipicolinic Acid Synthetase. J. Bacteriol.
818 198:1694–1707.
819 Dubnau D. 1991. Genetic competence in Bacillus subtilis. Microbiol. Rev. 55:395–
820 424.
821 Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and
822 high throughput. Nucleic Acids Res. 32:1792–1797.
823 Eichenberger P, Jensen ST, Conlon EM, van Ooij C, Silvaggi J, González-Pastor J-
824 E, Fujita M, Ben-Yehuda S, Stragier P, Liu JS, et al. 2003. The σE Regulon and
825 the Identification of Additional Sporulation Genes in Bacillus subtilis. J. Mol. Biol.
826 327:945–972.
827 Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen M-Y,
828 Pieper U, Sali A. 2006. Comparative protein structure modeling using Modeller.
829 Curr. Protoc. Bioinforma. Chapter 5:Unit-5.6.
830 Fang J, Kato C, Runko GM, Nogi Y, Hori T, Li J, Morono Y, Inagaki F. 2017.
831 Predominance of Viable Spore-Forming Piezophilic Bacteria in High-Pressure
832 Enrichment Cultures from ~1.5 to 2.4 km-Deep Coal-Bearing Sediments below
833 the Ocean Floor. Front. Microbiol. 8.
834 Fawcett P, Eichenberger P, Losick R, Youngman P. 2000. The transcriptional profile
835 of early to middle sporulation in Bacillus subtilis. Proc. Natl. Acad. Sci. 97:8063–
836 8068.
837 Filippidou S, Wunderlin T, Junier T, Jeanneret N, Dorador C, Molina V, Johnson DR,
838 Junier P. 2016. A Combination of Extreme Environmental Conditions Favor the

839 Prevalence of Endospore-Forming Firmicutes. Front. Microbiol. 7:1–11.
840 Fimlaid K a, Shen A. 2015. Diverse mechanisms regulate sporulation sigma factor
841 activity in the Firmicutes. Curr. Opin. Microbiol. 24:88–95.
842 Fimlaid KA, Bond JP, Schutz KC, Putnam EE, Leung JM, Lawley TD, Shen A. 2013.
843 Global Analysis of the Sporulation Pathway of Clostridium difficile. PLoS Genet.
844 9:e1003660.
845 Fimlaid KA, Jensen O, Donnelly ML, Siegrist MS, Shen A. 2015. Regulation of
846 Clostridium difficile Spore Formation by the SpoIIQ and SpoIIIA Proteins. PLoS
847 Genet. 11:1–35.
848 Flint JF, Drzymalski D, Montgomery WL, Southam G, Angert ER. 2005. Nocturnal
849 production of endospores in natural populations of epulopiscium-like
850 surgeonfish symbionts. J. Bacteriol. 187:7460–7470.
851 Fujita M, Losick R. 2005. Evidence that entry into sporulation in Bacillus subtilis is
852 governed by a gradual increase in the level and activity of the master regulator
853 Spo0A. Genes Dev. 19:2236–2244.
854 Fukushima T, Yamamoto H, Atrih A, Foster SJ, Sekiguchi J. 2002. A polysaccharide
855 deacetylase gene (pdaA) is required for germination and for production of
856 muramic delta-lactam residues in the spore cortex of Bacillus subtilis. J.
857 Bacteriol. 184:6007–6015.
858 Galperin MY. 2013. Genome Diversity of Spore-Forming Firmicutes. Microbiol.
859 Spectr. 1:1–15.
860 Galperin MY, Brover V, Tolstoy I, Yutin N. 2016. Phylogenomic analysis of the family
861 Peptostreptococcaceae (Clostridium cluster XI) and proposal for reclassification
862 of Clostridium litorale (Fendrich et al. 1991) and Eubacterium acidaminophilum
863 (Zindel et al. 1989) as Peptoclostridium litorale gen. nov. Int. J. Syst. Evol.
864 Microbiol. 66:5506–5513.
865 Galperin MY, Mekhedov SL, Puigbo P, Smirnov S, Wolf YI, Rigden DJ. 2012.
866 Genomic determinants of sporulation in Bacilli and Clostridia: towards the

867 minimal set of sporulation-specific genes. Environ. Microbiol. 14:2870–2890.
868 Ghosh J, Larsson P, Singh B, Pettersson BMF, Islam NM, Sarkar SN, Dasgupta S,
869 Kirsebom LA. 2009. Sporulation in mycobacteria. Proc. Natl. Acad. Sci.
870 106:10781–10786.
871 Guilhot E, Lagier JC, Raoult D, Khelaifia S. 2017. 'Prevotella ihumii' sp. nov. and
872 'Varibaculum timonense' sp. nov., two new bacterial species isolated from a
873 fresh human stool specimen. New Microbes New Infect. 18:3–5.
874 Haijema BJ, Hahn J, Haynes J, Dubnau D. 2001. A ComGA-dependent checkpoint
875 limits growth during the escape from competence. Mol. Microbiol. 40:52–64.
876 Hamilton-Brehm SD, Mosher JJ, Vishnivetskaya T, Podar M, Carroll S, Allman S,
877 Phelps TJ, Keller M, Elkins JG. 2010. Caldicellulosiruptor obsidiansis sp. nov.,
878 an anaerobic, extremely thermophilic, cellulolytic bacterium isolated from
879 obsidian pool, yellowstone National Park. Appl. Environ. Microbiol. 76:1014–
880 1020.
881 Han C, Gu W, Zhang X, Lapidus A, Nolan M, Copeland A, Lucas S, del Rio TG, Tice
882 H, Cheng JF, et al. 2010. Complete genome sequence of Thermaerobacter
883 marianensis type strain (7p75a T). Stand. Genomic Sci. 3:337–345.
884 He M, Sebaihia M, Lawley TD, Stabler RA, Dawson LF, Martin MJ, Holt KE, Seth-
885 Smith HMB, Quail MA, Rance R, et al. 2010. Evolutionary dynamics of
886 Clostridium difficile over short and long time scales. Proc. Natl. Acad. Sci.
887 107:7527–7532.
888 Henriques AO, Moran CP. 2007. Structure, assembly, and function of the spore
889 surface layers. Annu. Rev. Microbiol. 61:555–588.
890 Higgins D, Dworkin J. 2012. Recent progress in Bacillus subtilis sporulation. FEMS
891 Microbiol. Rev. 36:131–148.
892 Hong HA, Khaneja R, Tam NMK, Cazzato A, Tan S, Urdaci M, Brisson A, Gasbarrini
893 A, Barnes I, Cutting SM. 2009. Bacillus subtilis isolated from the human
894 gastrointestinal tract. Res. Microbiol. 160:134–143.

895 de Hoon MJL, Eichenberger P, Vitkup D. 2010. Hierarchical Evolution of the Bacterial
896 Sporulation Network. Curr. Biol. 20:R735–R745.
897 Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and
898 Visualization of Phylogenomic Data. Mol. Biol. Evol. 33:1635–1638.
899 Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T,
900 Mende DR, Sunagawa S, Kuhn M, et al. 2016. eggNOG 4.5: a hierarchical
901 orthology framework with improved functional annotations for eukaryotic,
902 prokaryotic and viral sequences. Nucleic Acids Res. 44:D286–D293.
903 Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield
904 CN, Hernsdorf AW, Amano Y, Ise K, et al. 2016. A new view of the tree of life.
905 Nat. Microbiol. 1:16048.
906 Illing N, Errington J. 1991. The spoIIIA operon of Bacillus subtilis defines a new
907 temporal class of mother-cell-specific sporulation genes under the control of the
908 sigma E form of RNA polymerase. Mol. Microbiol. 5:1927–1940.
909 Janoir C, Denève C, Bouttier S, Barbut F, Hoys S, Caleechum L, Chapetón-Montes
910 D, Pereira FC, Henriques AO, Collignon A, et al. 2013. Adaptive Strategies and
911 Pathogenesis of Clostridium difficile from In Vivo Transcriptomics. Infect.
912 Immun. 81:3757–3769.
913 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version
914 7: Improvements in performance and usability. Mol. Biol. Evol. 30:772–780.
915 Kim H, Hahn M, Grabowski P, McPherson DC, Otte MM, Wang R, Ferguson CC,
916 Eichenberger P, Driks A. 2006. The Bacillus subtilis spore coat protein
917 interaction network. Mol. Microbiol. 59:487–502.
918 Koonin E V., Makarova KS, Aravind L. 2001. Horizontal Gene Transfer in
919 Prokaryotes: Quantification and Classification. Annu. Rev. Microbiol. 55:709–
920 742.
921 Kroos L, Yu YT. 2000. Regulation of sigma factor activity during Bacillus subtilis
922 development. Curr. Opin. Microbiol. 3:553–560.

923 Lai E-M, Phadke ND, Kachman MT, Giorno R, Vazquez S, Vazquez JA, Maddock
924 JR, Driks A. 2003. Proteomic Analysis of the Spore Coats of Bacillus subtilis
925 and Bacillus anthracis. J. Bacteriol. 185:1443–1454.
926 Lawley TD, Clare S, Walker AW, Goulding D, Stabler RA, Croucher N, Mastroeni P,
927 Scott P, Raisen C, Mottram L, et al. 2009. Antibiotic Treatment of Clostridium
928 difficile Carrier Mice Triggers a Supershedder State, Spore-Mediated
929 Transmission, and Severe Disease in Immunocompromised Hosts. Infect.
930 Immun. 77:3661–3669.
931 Lawson PA, Citron DM, Tyrrell KL, Finegold SM. 2016. Reclassification of
932 Clostridium difficile as Clostridioides difficile (Hall and O'Toole 1935) Prévot
933 1938. Anaerobe 40:95–99.
934 Lessa FC, Mu Y, Bamberg WM, Beldavs ZG, Dumyati GK, Dunn JR, Farley MM,
935 Holzbauer SM, Meek JI, Phipps EC, et al. 2015. Burden of Clostridium difficile
936 Infection in the United States. N. Engl. J. Med. 372:825–834.
937 Levdikov VM, Blagova E V., McFeat A, Fogg MJ, Wilson KS, Wilkinson AJ. 2012.
938 Structure of components of an intercellular channel complex in sporulating
939 Bacillus subtilis. Proc. Natl. Acad. Sci. 109:5441–5445.
940 Linke K, Rückerl I, Brugger K, Karpiskova R, Walland J, Muri-Klinger S, Tichy A,
941 Wagner M, Stessl B. 2014. Reservoirs of Listeria species in three environmental
942 ecosystems. Appl. Environ. Microbiol. 80:5583–5592.
943 Losick R, Stragier P. 1992. Crisscross regulation of cell-type-specific gene
944 expression during development in B. subtilis. Nature 355:601–604.
945 Mäder U, Schmeisky AG, Flórez LA, Stülke J. 2012. SubtiWiki--a comprehensive
946 community resource for the model organism Bacillus subtilis. Nucleic Acids Res.
947 40:D1278-87.
948 Maitra A, Dill KA. 2015. Bacterial growth laws reflect the evolutionary importance of
949 energy efficiency. Proc. Natl. Acad. Sci. 112:406–411.
950 Marin J, Battistuzzi FU, Brown AC, Hedges SB. 2017. The Timetree of Prokaryotes:

951 New Insights into Their Evolution and Speciation. Mol. Biol. Evol. 34:437–446.
952 Maughan H, Birky CW, Nicholson WL. 2009. Transcriptome Divergence and the Loss
953 of Plasticity in Bacillus subtilis after 6,000 Generations of Evolution under
954 Relaxed Selection for Sporulation. J. Bacteriol. 191:428–433.
955 Maughan H, Masel J, Birky CW, Nicholson WL. 2007. The roles of mutation
956 accumulation and selection in loss of sporulation in experimental populations of
957 Bacillus subtilis. Genetics 177:937–948.
958 Maughan H, Nicholson WL. 2011. Increased fitness and alteration of metabolic
959 pathways during bacillus subtilis evolution in the laboratory. Appl. Environ.
960 Microbiol. 77:4105–4118.
961 Meeske AJ, Rodrigues CDA, Brady J, Lim HC, Bernhardt TG, Rudner DZ. 2016.
962 High-Throughput Genetic Screens Identify a Large and Diverse Collection of
963 New Sporulation Genes in Bacillus subtilis. PLoS Biol. 14:1–33.
964 Meisner J, Wang X, Serrano M, Henriques AO, Moran CP. 2008. A channel
965 connecting the mother cell and forespore during bacterial endospore formation.
966 Proc. Natl. Acad. Sci. U. S. A. 105:15100–15105.
967 Mell JC, Redfield RJ. 2014. Natural competence and the evolution of DNA uptake
968 specificity. J. Bacteriol. 196:1471–1483.
969 Moeller R, Raguse M, Reitz G, Okayasu R, Li Z, Klein S, Setlow P, Nicholson WL.
970 2014. Resistance of Bacillus subtilis spore DNA to lethal ionizing radiation
971 damage relies primarily on spore core components and DNA repair, with minor
972 effects of oxygen radical detoxification. Appl. Environ. Microbiol. 80:104–109.
973 Moreno-Letelier A, Olmedo-Alvarez G, Eguiarte LE, Souza V. 2012. Divergence and
974 phylogeny of Firmicutes from the Cuatro Ciénegas Basin, Mexico: a window to
975 an ancient ocean. Astrobiology 12:674–684.
976 Morlot C, Rodrigues CDA. 2018. The New Kid on the Block: A Specialized Secretion
977 System during Bacterial Sporulation. Trends Microbiol. 26:663–676.
978 Mukherjee S, Barash D, Sengupta S. 2017. Comparative genomics and

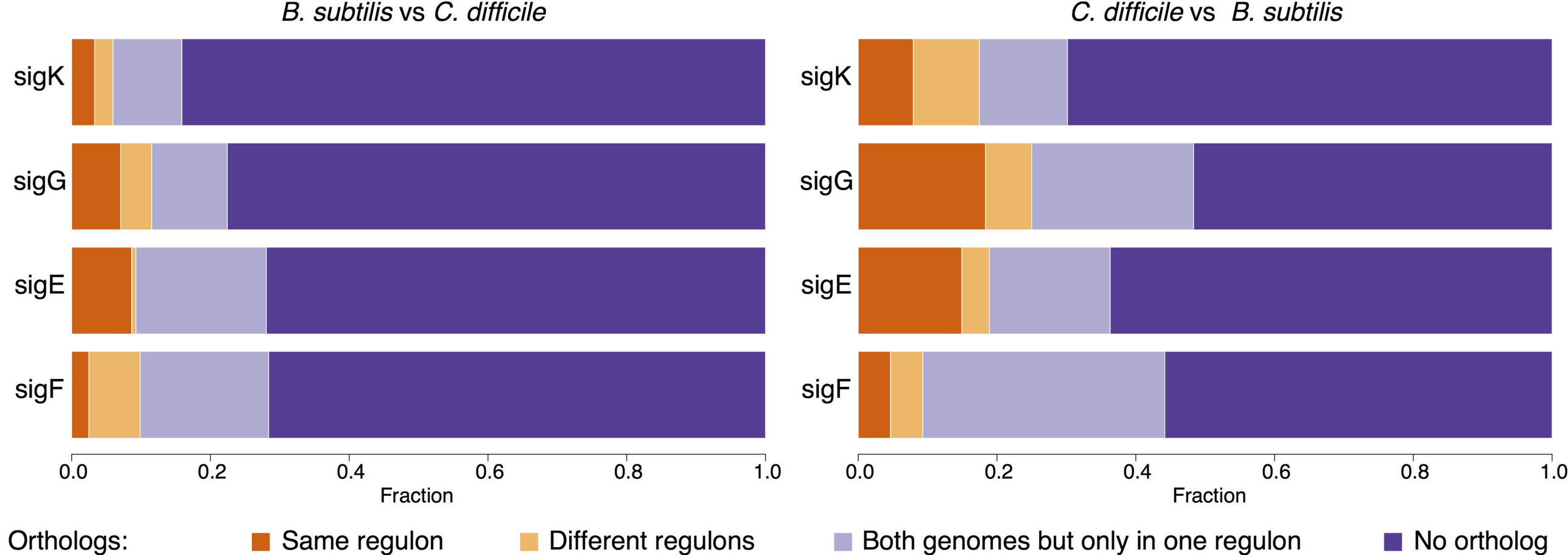
979 phylogenomic analyses of lysine riboswitch distributions in bacteria. PLoS One
980 12:e0184314.
981 Mukhopadhya I, Moraïs S, Laverde-Gomez J, Sheridan PO, Walker AW, Kelly W,
982 Klieve A V., Ouwerkerk D, Duncan SH, Louis P, et al. 2018. Sporulation
983 capability and amylosome conservation among diverse human colonic and
984 rumen isolates of the keystone starch-degrader Ruminococcus bromii. Environ.
985 Microbiol. 20:324–336.
986 Nicholson WL. 2012. Increased competitive fitness of Bacillus subtilis under
987 nonsporulating conditions via inactivation of pleiotropic regulators AlsR, SigD,
988 and SigW. Appl. Environ. Microbiol. 78:3500–3503.
989 Nicholson WL, Munakata N, Horneck G, Melosh HJ, Setlow P. 2000. Resistance of
990 Bacillus endospores to extreme terrestrial and extraterrestrial environments.
991 Microbiol. Mol. Biol. Rev. 64:548–572.
992 Notredame C, Higgins D, Heringa J. 2000. T-Coffee : A Novel Method for Fast and
993 Accurate Multiple Sequence Alignment. J. Mol. Biol. 302:205–217.
994 Onyenwoke RU, Brill JA, Farahi K, Wiegel J. 2004. Sporulation genes in members of
995 the low G+C Gram-type-positive phylogenetic branch (Firmicutes). Arch.
996 Microbiol. 182:182–192.
997 Orsburn BC, Melville SB, Popham DL. 2010. EtfA catalyses the formation of
998 dipicolinic acid in Clostridium perfringens. Mol. Microbiol. 75:178–186.
999 Paredes-Sabja D, Shen A, Sorg J a. 2014. Clostridium difficile spore biology:
1000 sporulation, germination, and spore structural proteins. Trends Microbiol.
1001 22:406–416.
1002 Paredes CJ, Alsaker K V, Papoutsakis ET. 2005. A comparative genomic view of
1003 clostridial sporulation and physiology. Nat. Rev. Microbiol. 3:969–978.
1004 Pereira FC, Saujet L, Tomé AR, Serrano M, Monot M, Couture-Tosi E, Martin-
1005 Verstraete I, Dupuy B, Henriques AO. 2013. The spore differentiation pathway
1006 in the enteric pathogen Clostridium difficile. PLoS Genet. 9:e1003782.

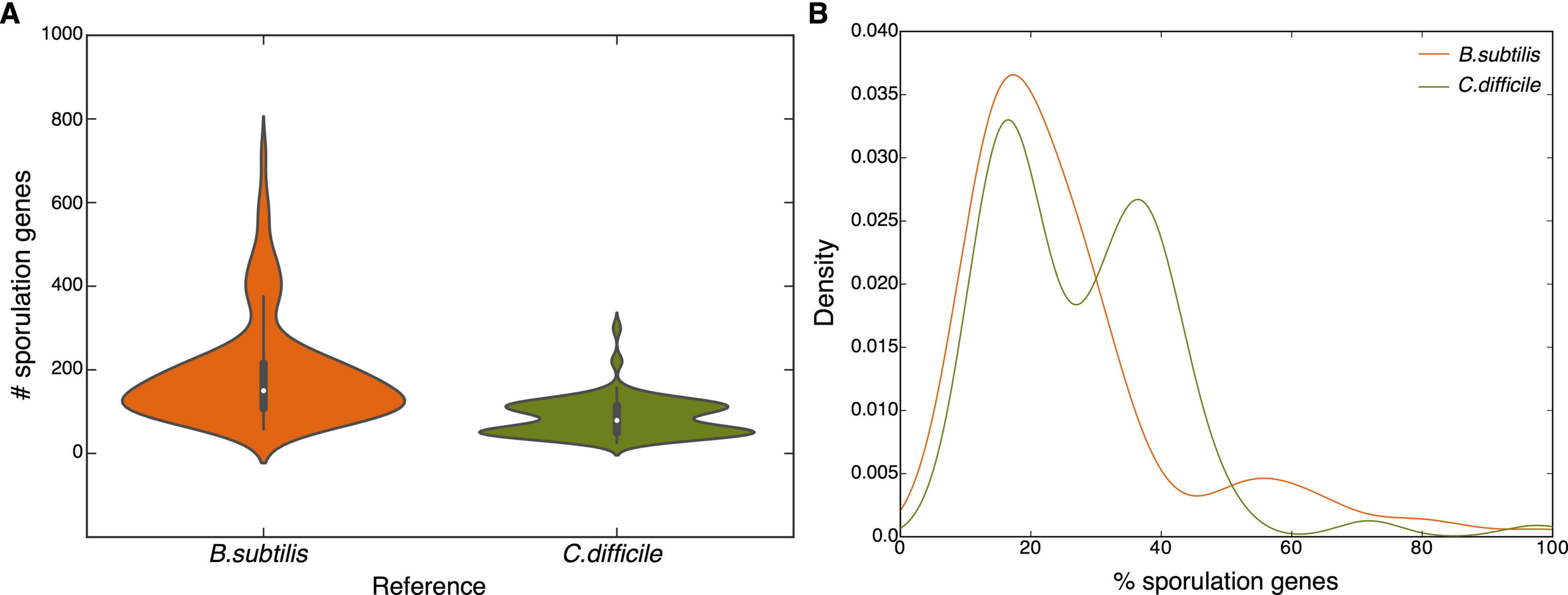
1007 Permpoonpattana P, Tolls EH, Nadem R, Tan S, Brisson A, Cutting SM. 2011.
1008 Surface Layers of Clostridium difficile Endospores. J. Bacteriol. 193:6461–6470.
1009 Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin
1010 TE. 2004. UCSF Chimera - A visualization system for exploratory research and
1011 analysis. J. Comput. Chem. 25:1605–1612.
1012 Pishdadian K, Fimlaid KA, Shen A. 2015. SpoIIID-mediated regulation of σ K function
1013 during C lostridium difficile sporulation. Mol. Microbiol. 95:189–208.
1014 Popham DL, Illades-Aguiar B, Setlow P. 1995. The Bacillus subtilis dacB gene,
1015 encoding penicillin-binding protein 5*, is part of a three-gene operon required for
1016 proper spore cortex synthesis and spore core dehydration. J. Bacteriol.
1017 177:4721–4729.
1018 Ramos-Silva P, Brito PH, Serrano M, Henriques AO, Pereira-Leal JB. 2015.
1019 Rethinking the Niche of Upper-Atmosphere Bacteria: Draft Genome Sequences
1020 of Bacillus aryabhattai C765 and Bacillus aerophilus C772, Isolated from Rice
1021 Fields. Genome Announc. 3:e00094-15.
1022 Redmond C. 2004. Identification of proteins in the exosporium of Bacillus anthracis.
1023 Microbiology 150:355–363.
1024 Rodrigues CDA, Marquis KA, Meisner J, Rudner DZ. 2013. Peptidoglycan hydrolysis
1025 is required for assembly and activity of the transenvelope secretion complex
1026 during sporulation in Bacillus subtilis. Mol. Microbiol. 89:1039–1052.
1027 Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B,
1028 Liu L, Suchard M a., Huelsenbeck JP. 2012. Mrbayes 3.2: Efficient bayesian
1029 phylogenetic inference and model choice across a large model space. Syst.
1030 Biol. 61:539–542.
1031 Rudner DZ. 2002. A sporulation membrane protein tethers the pro-sigma K
1032 processing enzyme to its inhibitor and dictates its subcellular localization. Genes
1033 Dev. 16:1007–1018.
1034 Rudner DZ, Losick R. 2001. Morphological coupling in development: lessons from

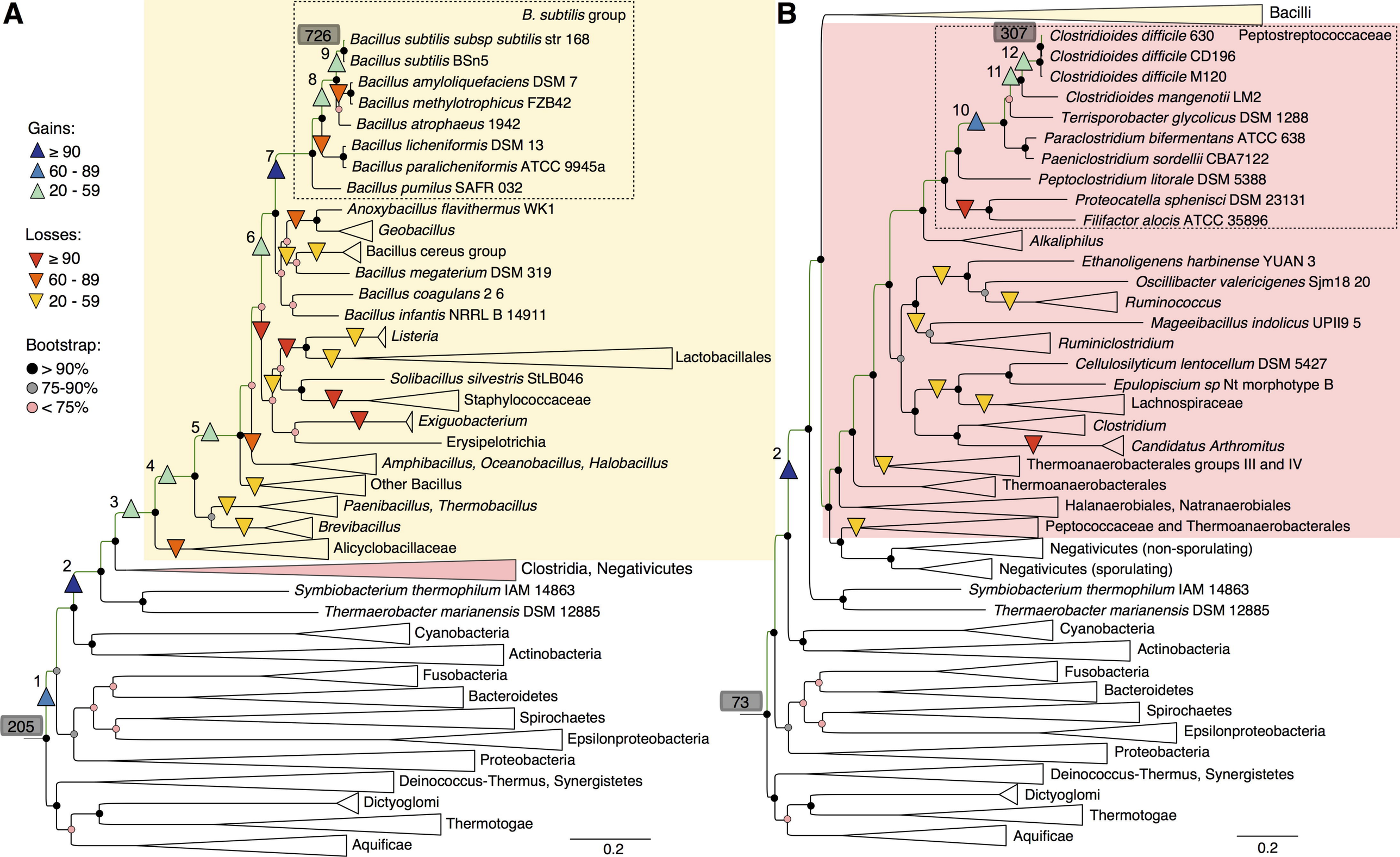
1035 prokaryotes. Dev. Cell 1:733–742.
1036 Sauders BD, Overdevest J, Fortes E, Windham K, Schukken Y, Lembo A, Wiedmann
1037 M. 2012. Diversity of Listeria species in urban and natural environments. Appl.
1038 Environ. Microbiol. 78:4420–4433.
1039 Saujet L, Pereira FC, Serrano M, Soutourina O, Monot M, Shelyakin P V, Gelfand
1040 MS, Dupuy B, Henriques AO, Martin-Verstraete I. 2013. Genome-wide analysis
1041 of cell type-specific gene transcription during spore formation in Clostridium
1042 difficile. PLoS Genet. 9:e1003756.
1043 Schloss PD, Iverson KD, Petrosino JF, Schloss SJ. 2014. The dynamics of a family's
1044 gut microbiota reveal variations on a theme. Microbiome 2:1–13.
1045 Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson
1046 NR, Roberts AP, Cerdeño-Tárraga AM, Wang H, et al. 2006. The multidrug-
1047 resistant human pathogen Clostridium difficile has a highly mobile, mosaic
1048 genome. Nat. Genet. 38:779–786.
1049 Serrano M, Crawshaw AD, Dembek M, Monteiro JM, Pereira FC, Pinho MG,
1050 Fairweather NF, Salgado PS, Henriques AO. 2016. The SpoIIQ-SpoIIIAH
1051 complex of Clostridium difficile controls forespore engulfment and late stages of
1052 gene expression and spore morphogenesis. Mol. Microbiol. 100:204–228.
1053 Setlow P. 2014. Spore Resistance Properties. Microbiol. Spectr. 2.
1054 Slieman TA, Nicholson WL. 2001. Role of Dipicolinic Acid in Survival of Bacillus
1055 subtilis Spores Exposed to Artificial and Solar UV Radiation. Appl. Environ.
1056 Microbiol. 67:1274–1279.
1057 Smits WK, Eschevins CC, Susanna KA, Bron S, Kuipers OP, Hamoen LW. 2005.
1058 Stripping Bacillus: ComK auto-stimulation is responsible for the bistable
1059 response in competence development. Mol. Microbiol. 56:604–614.
1060 Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-
1061 analysis of large phylogenies. Bioinformatics 30:1312–1313.
1062 Steil L, Serrano M, Henriques AO, Völker U. 2005. Genome-wide analysis of

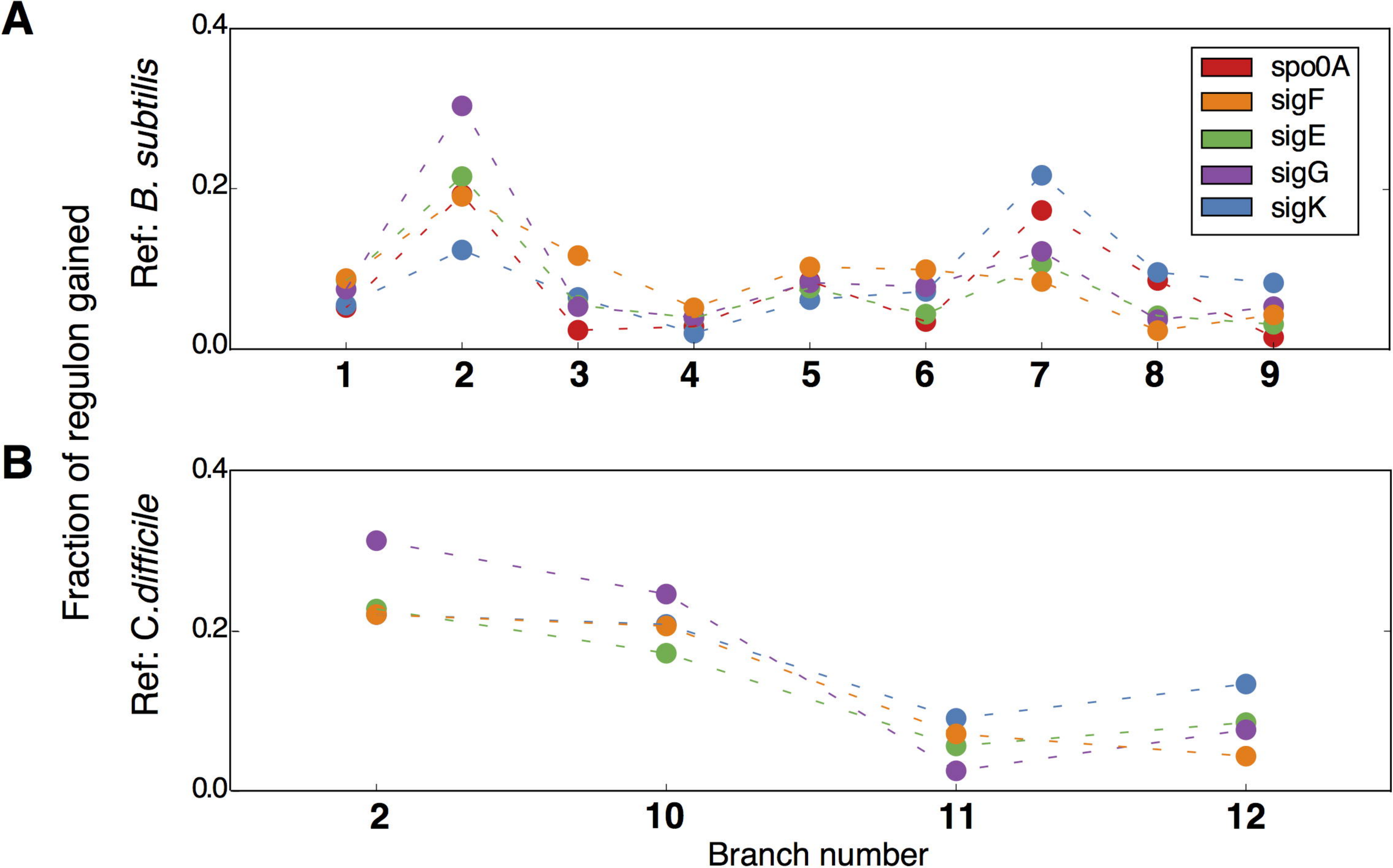
1063 temporally regulated and compartment-specific gene expression in sporulating
1064 cells of Bacillus subtilis. Microbiology 151:399–420.
1065 Stragier P, Losick R. 1996. Molecular genetics of sporulation in Bacillus subtilis.
1066 Annu. Rev. Genet. 30:297–41.
1067 Thompson BM, Hoelscher BC, Driks A, Stewart GC. 2012. Assembly of the BclB
1068 glycoprotein into the exosporium and evidence for its role in the formation of the
1069 exosporium “cap” structure in Bacillus anthracis. Mol. Microbiol. 86:1073–1084.
1070 Traag BA, Driks A, Stragier P, Bitter W, Broussard G, Hatfull G, Chu F, Adams KN,
1071 Ramakrishnan L, Losick R. 2010. Do mycobacteria produce endospores? Proc.
1072 Natl. Acad. Sci. 107:878–881.
1073 Ueda K, Yamashita A, Ishikawa J, Shamada M, Watsuji TO, Morimura K, Ikeda H,
1074 Hattori M, Beppu T. 2004. Genome sequence of Symbiobacterium
1075 thermophilum, an uncultivable bacterium that depends on microbial
1076 commensalism. Nucleic Acids Res. 32:4937–4944.
1077 Urios L. 2004. Tepidibacter formicigenes sp. nov., a novel spore-forming bacterium
1078 isolated from a Mid-Atlantic Ridge hydrothermal vent. Int. J. Syst. Evol.
1079 Microbiol. 54:439–443.
1080 Vishnivetskaya TA, Kathariou S. 2005. Putative transposases conserved in
1081 Exiguobacterium isolates from ancient Siberian permafrost and from
1082 contemporary surface habitats. Appl. Environ. Microbiol. 71:6954–6962.
1083 Vishnivetskaya TA, Kathariou S, Tiedje JM. 2009. The Exiguobacterium genus:
1084 Biodiversity and biogeography. Extremophiles 13:541–555.
1085 Vos M. 2009. Why do bacteria engage in homologous recombination? Trends
1086 Microbiol. 17:226–232.
1087 Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria
1088 and archaea. ISME J. 3:199–208.
1089 Wang ST, Setlow B, Conlon EM, Lyon JL, Imamura D, Sato T, Setlow P, Losick R,
1090 Eichenberger P. 2006. The Forespore Line of Gene Expression in Bacillus

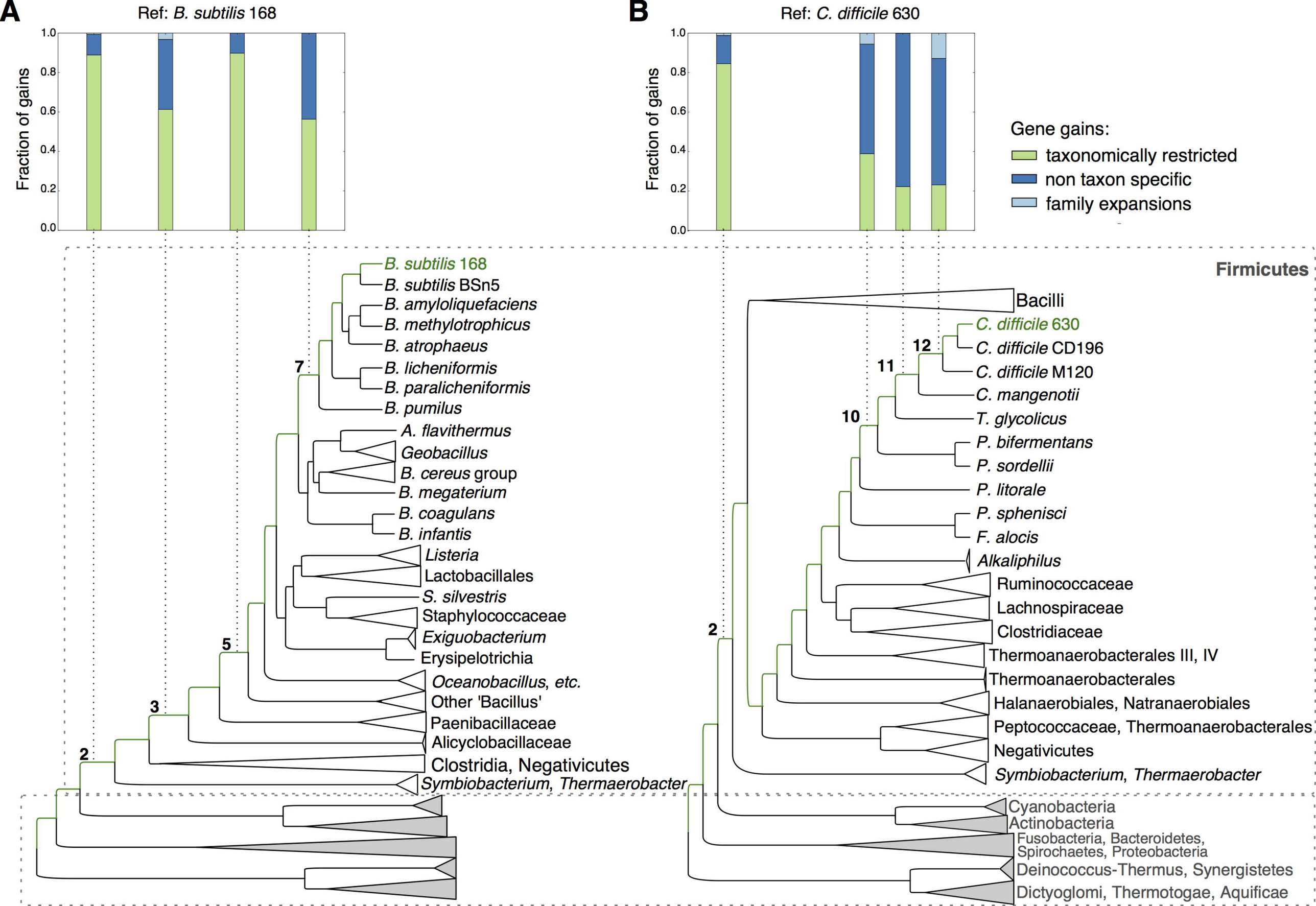
1091 subtilis. J. Mol. Biol. 358:16–37.
1092 Waterhouse AM, Procter JB, Martin DM a, Clamp M, Barton GJ. 2009. Jalview
1093 Version 2--a multiple sequence alignment editor and analysis workbench.
1094 Bioinformatics 25:1189–1191.
1095 Van De Werken HJG, Verhaart MRA, VanFossen AL, Willquist K, Lewis DL, Nichols
1096 JD, Goorissen HP, Mongodin EF, Nelson KE, Van Niel EWJ, et al. 2008.
1097 Hydrogenomics of the extremely thermophilic bacterium Caldicellulosiruptor
1098 saccharolyticus. Appl. Environ. Microbiol. 74:6720–6729.
1099 Yano JM, Yu K, Donaldson GP, Shastri GG, Ann P, Ma L, Nagler CR, Ismagilov RF,
1100 Mazmanian SK, Hsiao EY. 2015. Indigenous Bacteria from the Gut Microbiota
1101 Regulate Host Serotonin Biosynthesis. Cell 161:264–276.
1102 Zeytuni N, Flanagan KA, Worrall LJ, Massoni SC, Camp AH, Strynadka NCJ. 2018.
1103 Structural and biochemical characterization of SpoIIIAF, a component of a
1104 sporulation-essential channel in Bacillus subtilis. J. Struct. Biol. 204:1–8.
1105 Zeytuni N, Hong C, Flanagan KA, Worrall LJ, Theiltges KA, Vuckovic M, Huang RK,
1106 Massoni SC, Camp AH, Yu Z, et al. 2017. Near-atomic resolution cryoelectron
1107 microscopy structure of the 30-fold homooligomeric SpoIIIAG channel essential
1108 to spore formation in Bacillus subtilis. Proc. Natl. Acad. Sci. 114:E7073–E7081.
1109
1110

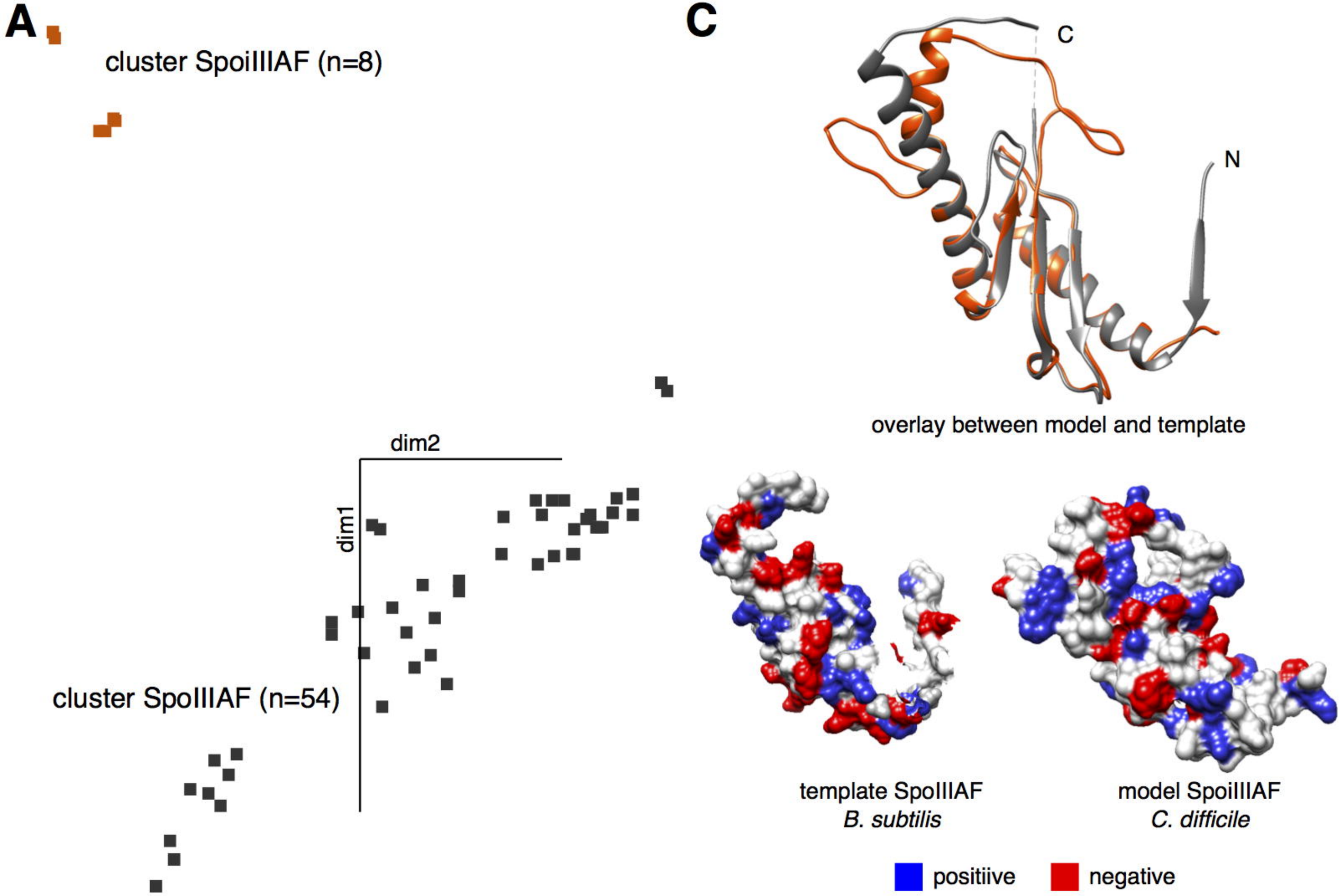

B. subtilis vs C. difficile
C. difficile vs B. subtilis
sigK
sigG
sigE
sigF
0.0
0.2
0.4
0.6
0.8
1.0
Fraction
Orthologs:
Same regulon
Different regulons
Both genomes but only in one regulon
No ortholog

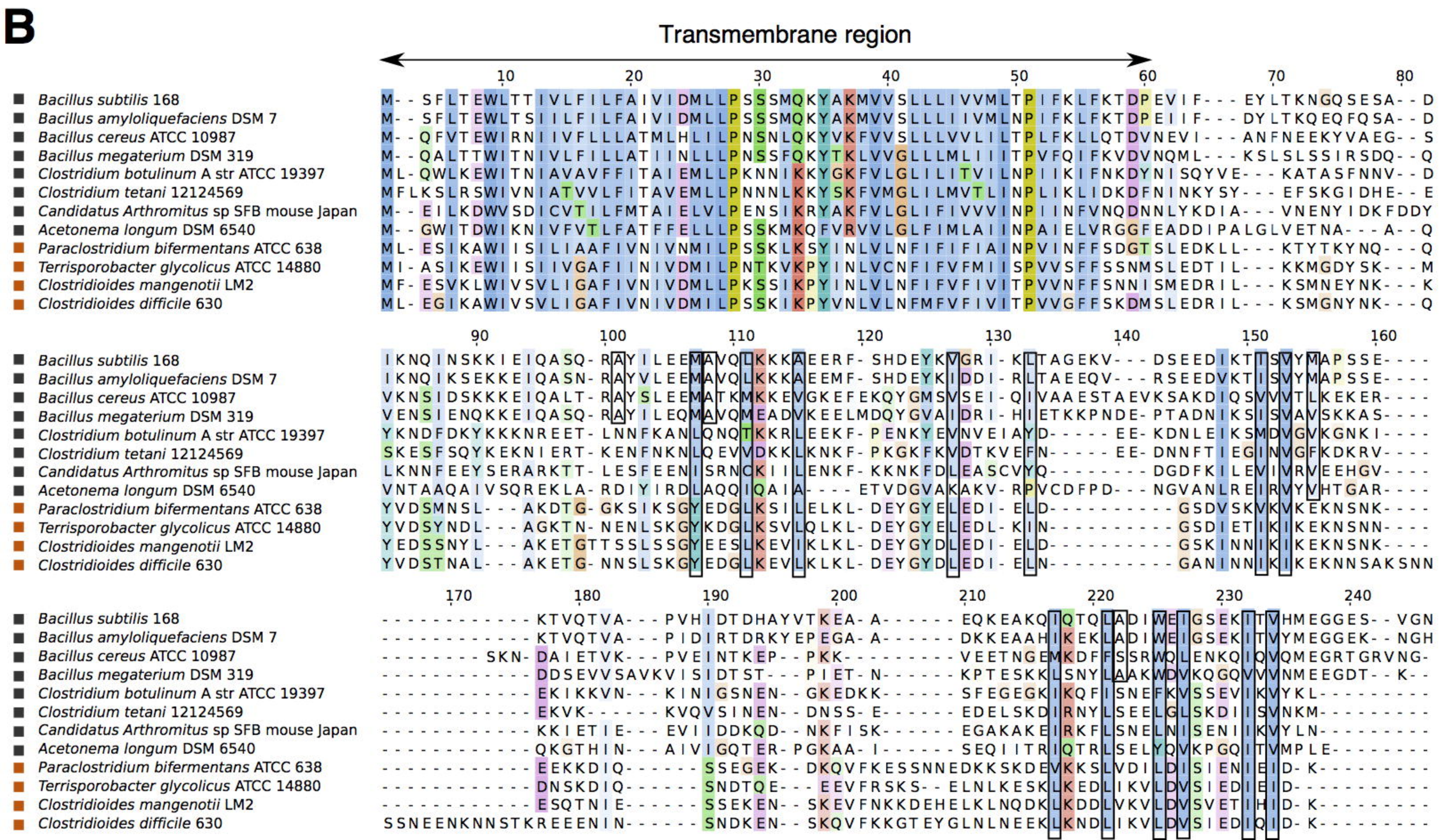

A
1000
800
600
400
200
0
sporulation genes
B.subtilis
C.difficile
Reference
B
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.0
Density
B.subtilis
C.difficile
0
20
40
60
80
100
% sporulation genes


A
Gains:
≥ 90
60 - 89
20 - 59
Losses:
≥ 90
60 - 89
20 - 59
Bootstrap:
> 90%
75-90%
< 75%
B. subtilis group
726
Bacillus subtilis subsp subtilis str 168
Bacillus subtilis BSn5
Bacillus amyloliquefaciens DSM 7
Bacillus methylotrophicus FZB42
Bacillus atrophaeus 1942
Bacillus licheniformis DSM 13
Bacillus paralicheniformis ATCC 9945a
Bacillus pumilus SAFR 032
Anoxybacillus flavithermus WK1
Geobacillus
Bacillus cereus group
Bacillus megaterium DSM 319
Bacillus coagulans 2 6
Bacillus infantis NRRL B 14911
Listeria
Lactobacillales
Solibacillus silvestris StLB046
Staphylococcaceae
Exiguobacterium
Erysipelotrichia
Amphibacillus, Oceanobacillus, Halobacillus
Other Bacillus
Paenibacillus, Thermobacillus
Brevibacillus
Alicyclobacillaceae
Clostridia, Negativicutes
Symbiobacterium thermophilum IAM 14863
Thermaerobacter marianensis DSM 12885
Cyanobacteria
Actinobacteria
Fusobacteria
Bacteroidetes
Spirochaetes
Epsilonproteobacteria
Proteobacteria
Deinococcus-Thermus, Synergistetes
Dictyoglomi
Thermotogae
Aquificae
205
0.2
B
Bacilli
307
Clostridioides difficile 630
Clostridioides difficile CD196
Clostridioides difficile M120
Clostridioides mangenotii LM2
Terrisporobacter glycolicus DSM 1288
Paraclostridium bifermentans ATCC 638
Paeniclostridium sordellii CBA7122
Peptoclostridium litorale DSM 5388
Proteocatella sphenisci DSM 23131
Filifactor alocis ATCC 35896
Peptostreptococcaceae
Alkaliphilus
Ethanoligenens harbinense YUAN 3
Oscillibacter valericigenes Sjm18 20
Ruminococcus
Mageeibacillus indolicus UPII9 5
Ruminiclostridium
Cellulosilyticum lentocellum DSM 5427
Epulopiscium sp Nt morphotype B
Lachnospiraceae
Clostridium
Candidatus Arthromitus
Thermoanaerobacterales groups III and IV
Thermoanaerobacterales
Halanaerobiales, Natranaerobiales
Peptococcaceae and Thermoanaerobacterales
Negativicutes (non-sporulating)
Negativicutes (sporulating)
Symbiobacterium thermophilum IAM 14863
Thermaerobacter marianensis DSM 12885
Cyanobacteria
Actinobacteria
Fusobacteria
Bacteroidetes
Spirochaetes
Epsilonproteobacteria
Proteobacteria
Deinococcus-Thermus, Synergistetes
Dictyoglomi
Thermotogae
Aquificae
73
0.2


A
B
Fraction of regulon gained
Ref: B. subtilis
Ref: C.difficile
0.4
0.2
0.0
spo0A
sigF
sigE
sigG
sigK
1
2
3
4
5
6
7
8
9
2
10
11
12
Branch number


A
Ref: B. subtilis 168
Fraction of gains
1.0
0.8
0.6
0.4
0.2
0.0
B
Ref: C. difficile 630
Gene gains:
taxonomically restricted
non taxon specific
family expansions
Firmicutes
B. subtilis 168
B. subtilis BSn5
B. amyloliquefaciens
B. methylotrophicus
B. atrophaeus
B. licheniformis
B. paralicheniformis
B. pumilus
A. flavithermus
Geobacillus
B. cereus group
B. megaterium
B. coagulans
B. infantis
Listeria
Lactobacillales
S. silvestris
Staphylococcaceae
Exiguobacterium
Erysipelotrichia
Oceanobacillus, etc.
Other 'Bacillus'
Paenibacillaceae
Alicyclobacillaceae
Clostridia, Negativicutes
Symbiobacterium, Thermaerobacter
7
5
3
2
Bacilli
C. difficile 630
C. difficile CD196
C. difficile M120
C. mangenotii
T. glycolicus
P. bifermentans
P. sordellii
P. litorale
P. sphenisci
F. alocis
Alkaliphilus
Ruminococcaceae
Lachnospiraceae
Clostridiaceae
Thermoanaerobacterales III, IV
Thermoanaerobacterales
Halanaerobiales, Natranaerobiales
Peptococcaceae, Thermoanaerobacterales
Negativicutes
Symbiobacterium, Thermaerobacter
12
11
10
2
Cyanobacteria
Actinobacteria
Fusobacteria, Bacteroidetes,
Spirochaetes, Proteobacteria
Deinococcus-Thermus, Synergistetes
Dictyoglomi, Thermotogae, Aquificae


A
cluster SpoIIIAF (n=8)
cluster SpoIIIAF (n=54)
dim1
dim2
C
C
N
overlay between model and template
template SpoIIIAF
B. subtilis
model SpoIIIAF
C. difficile
positiive
negative


B
Transmembrane region
Bacillus subtilis 168
Bacillus amyloliquefaciens DSM 7
Bacillus cereus ATCC 10987
Bacillus megaterium DSM 319
Clostridium botulinum A str ATCC 19397
Clostridium tetani 12124569
Candidatus Arthromitus sp SFB mouse Japan
Acetonema longum DSM 6540
Paraclostridium bifermentans ATCC 638
Terrisporobacter glycolicus ATCC 14880
Clostridioides mangenotii LM2
Clostridioides difficile 630