**Contrasting evolutionary patterns of spore coat proteins in two *Bacillus* species groups are linked to a cellular structural difference.**

Hong Qina, \*, Adam Driksb

aDepartment of Biology, Spelman College, Atlanta, GA 30314

bDepartment of Microbiology and Immunology, Loyola University Medical Center, Maywood, IL. 60153

**Running Title:** Phylogenomics of Bacillus spore coat.

**Key words:** *Bacillus*, coat proteins, spore, phylogenetic profiles

\* Corresponding author:

Hong Qin

Department of Biology

Spelman College

350 Spelman Ln SW, Box 1183

Atlanta, GA 30314

Email: hqin@spelman.edu

Abstract

**Background:** When starved, bacteria of *Bacillus* genus produce a highly resistant cell type called a spore encased in a complex protective protein shell called the coat. The *B. subtilis* species group and the *B. cereus* species group are two well-studied groups in this genus. Spores of some species, including those of the *B.* *cereus*-group, contain an additional outer layer, the exosporium, that encircles the coat. The *B. subtilis* spore coat consists of an inner and an outer layer. Here, we compare coat protein evolution in the *Bacillaceae* and, in particular, between these two species groups, to understand the evolutionary impacts of structural differences in the spores.

**Results**: We curated a list of *B. subtilis* coat proteins and identified orthologs in a group of other *Bacillaceae* species. Phylogenic profiling showed that coat proteins can be grouped into conserved and labile ones. We found that coat protein composition is more conserved in the *B. subtilis* inner coat than the outer coat. By comparing the ratio of non-synonymous and synonymous substitution rates, we found that coat proteins have a higher ratio of nonsynonymous versus synonymous substitution rates than the rest of the genome only in the *B. subtilis*-group. In contrast, similar evolution patterns of coat proteins were not observed in the *B. cereus*-group.

**Conclusions:** We show that coat proteins are under contrasting selective pressures between *B. subtilis*-group species and *B. cereus*-group species. We speculate that this difference in selective pressure reflects the fact that in *B. cerues*-group species, the exosporium, and not the coat, is the outermost spore layer. We further speculate that differences in coat protein evolution between these groups of species are driven, at least in part, by the presence or absence of the exosporium in response to different ecological niches.

**1. Introduction**

The defining feature of bacteria of the family Bacillaceae (and the genus Bacillus in particular) is the ability to form a specialized alternate cell type, called the spore, which can withstand a wide range of environmental stresses, including toxic chemicals, heat, ultraviolet radiation and microbial predation [[1-4](#_ENREF_1" \o "Klobutcher, 2006 #159)]. The spore is essentially metabolically dormant and can remain in this state for extreme periods of time. Nonetheless, the spore can return to active growth once nutrient is available, in a process called germination [[5](#_ENREF_5" \o "Moir, 2006 #160)]. The ability of spores to remain dormant for long time periods and to resist extreme conditions has made this cell type a major model for studies of cellular defenses against stress.

The *Bacillaceae* thrive in essentially all environments, and have significant taxonomic and phylogenetic diversity, neither of which are fully characterized [[6](#_ENREF_6" \o "Fritze, 2004 #189)]. The vast majority of research on these organisms has focused on only two clades. The first of these is the “*B. cereus*-group”, which is comprised of the closely related species *Bacillus anthracis*, *Bacillus cereus*, *Bacillus thuringiensis*, *Bacillus mycoides*, *Bacillus pseudomycoides* and *Bacillus weihenstephanensis* [[7](#_ENREF_7" \o "Tourasse, 2006 #111)]. Of these, the best studied are *B. anthracis*, the causative agent of anthrax [[8](#_ENREF_8" \o "Mock, 2001 #176)], *B. cereus*, an important food-borne pathogen [[9](#_ENREF_9" \o "Stenfors Arnesen, 2008 #178)], and *B. thuringiensis*, which can produce an insect toxin and, therefore, be used for agricultural biocontrol [[10](#_ENREF_10" \o "Aronson, 2001 #173)]. The second clade is comprised of *Bacillus subtilis* and its close relatives, including *B. lichenniformis*, *B. pumilus*, *B. amyloliquefaciens*, *B. atrophaeus*, *B. mojavensis*, and *B. vallismortis*. Of these, only *B. subtilis* has received extensive study, making this species the primary model for Gram-positive bacteria and a major model for bacterial development [[11](#_ENREF_11" \o "Sonenshein, 2002 #190)]. Because the *B. cereus* -group and *B. subtilis* -group species comprise only a very small subset of the total diversity of the *Bacillaceae* [[12](#_ENREF_12" \o "Blackwood, 2004 #50)], the biology of the majority of these organisms remains poorly understood [[13](#_ENREF_13" \o "Driks, 2007 #6)].

A structure that is found in spores of all *Bacillaceae* (and, indeed, Clostridia as far as is known) is the coat, a protein shell that encapsulates and protects the spore [[14-17](#_ENREF_14" \o "Henriques, 2007 #55)]. In species where it is the outermost spore structure (see below), the coat has the important role of interacting directly with surfaces in the environment. For example, proteins on the coat surface play a critical role in the adhesive properties of the spore [[18](#_ENREF_18" \o "Chen, 2010 #236)]. It is likely that there are other roles for coat interactions with the environment but they remain undescribed [[15](#_ENREF_15" \o "Driks, 1999 #30), [18-22](#_ENREF_18" \o "Chen, 2010 #236)]. The coat has additional diverse functions, including roles in germination and resistance to environmental stresses, like small reactive molecules, degradative enzymes, microbial predation and UV radiation [[1](#_ENREF_1" \o "Klobutcher, 2006 #159), [15](#_ENREF_15" \o "Driks, 1999 #30), [19](#_ENREF_19" \o "Ragkousi, 2003 #235), [20](#_ENREF_20" \o "Riesenman, 2000 #234), [22](#_ENREF_22" \o "Behravan, 2000 #218), [23](#_ENREF_23" \o "Setlow, 2006 #228)]. It is plausible that any or all of these coat functions could differ among *Bacillaceae* species that inhabit various niches and the challenges faced by these spores may vary as well. These characteristics are among those making bacterial spores unique in nature and have motivated over 140 years of research [[11](#_ENREF_11" \o "Sonenshein, 2002 #190), [24](#_ENREF_24" \o "Cohn, 1876 #194), [25](#_ENREF_25" \o "Koch, 1876 #193)].

The coat varies significantly in structure among species [[15](#_ENREF_15" \o "Driks, 1999 #30), [26-28](#_ENREF_26" \o "Holt, 1969 #185)]. In *B. subtilis*, the coat has three major layers distinguishable by thin-section electron microscopy: a lightly staining inner coat and a darkly staining outer coat that encases a crust [[29](#_ENREF_29" \o "Warth, 1963 #179), [30](#_ENREF_30" \o "McKenney, 2010 #256)]. The crust is a recently identified structure that is distinct from the outer coat [[30](#_ENREF_30" \o "McKenney, 2010 #256)]. Other species, including those of the *B. cereus* -group, have a thinner coat [[27](#_ENREF_27" \o "Aronson, 1976 #195)]. The coat can also possess more complex features, such as the long filamentous structures in *Bacillus clausii* [[28](#_ENREF_28" \o "Traag, 2010 #237)]. *B. cereus* -group species, as well as other species including *B. megaterium*, *B. laterosporus* and *B. vedderi*, possess an additional structure that surrounds the coat, called the exosporium which also varies in structure among species [[14](#_ENREF_14" \o "Henriques, 2007 #55), [28](#_ENREF_28" \o "Traag, 2010 #237), [31](#_ENREF_31" \o "Hannay, 1957 #196), [32](#_ENREF_32" \o "Vary, 1994 #216)]. The exosporium is known to have roles in interacting with environmental surfaces and other cells [[18](#_ENREF_18" \o "Chen, 2010 #236), [33](#_ENREF_33" \o "Oliva, 2008 #214), [34](#_ENREF_34" \o "Bozue, 2007 #181)]. It may also have uncharacterized roles in protection of the spore. Importantly, the exosporium is not an impermeable barrier, as it allows passage of small molecules such as sugars and amino acids [[35](#_ENREF_35" \o "Ball, 2008 #197)].

Understanding the forces that guide the evolution of the coat can provide unique insight into coat function and formation. For example, identifying highly conserved coat proteins may reveal those with important functions in coat assembly and function [[16](#_ENREF_16" \o "Driks, 2002 #17)]. This information, in turn, can help identify which coat proteins are more involved in adaptation. This is an especially interesting question given that the majority of the morphological variation among Bacillus spores is in the coat (as well as the exosporium) [[26-28](#_ENREF_26" \o "Holt, 1969 #185)]. Importantly, by measuring the degree of selection on a coat protein, it may be possible to show that coat proteins have evolutionarily important roles even when the corresponding coat protein gene mutants lack a detectable phenotype in the laboratory [[17](#_ENREF_17" \o "Driks, 2009 #212), [36](#_ENREF_36" \o "Driks, 2002 #217)].

In this work we first generated a curated list of *B. subtilis* spore coat proteins, allowing us to identify their orthologs based on phylogeny in a group of Bacillus species (10 fully-sequenced and 1 partially-sequenced). We then performed a detailed analysis of the molecular evolution of these proteins. Our results showed that evolutionary differences in spore coat proteins can reflect their locations in spore coat layers and differences in spore structure across species.

**2. Materials and Methods**

**2.1 Sequences**

Genomes analyzed in this study are summarized in Table 1. Most of the genomes are the species type-strains. We analyzed 5 *B. subtilis* -group genomes: *Bacillus subtilis* *subsp. subtilis str. 168*, *Bacillus mojavensis* RO-H-1, *Bacillus licheniformis* ATCC 14580, *Bacillus amyloliquefaciens* FZB42, and *Bacillus pumilus* SAFR-032. We analyzed 6 *B.* *cereus*-group genomes: *Bacillus anthracis str. Ames*, *Bacillus cereus* ATCC 10987, *Bacillus cereus* ATCC 14579, *Bacillus cereus* E33L, *Bacillus thuringiensis serovar konkukian*, and *Bacillus weihenstephanensis* KBAB4. We used genomes of *Bacillus clausii* KSM-K16 and *Bacillus halodurans* C-125 as outgroups. Genes of the draftgenome of *Bacillus mojavensis* RO-H-1 were predicted by GLIMMER [[37](#_ENREF_37" \o "Salzberg, 1998 #205)].

The rRNA sequences were aligned based on structural information using the Ribosomal Database Project II (Release 9.56, http://rdp.cme.msu.edu/). The annotation of the *B. subtilis* genome was based on SubtiList (http://genolist.pasteur.fr/SubtiList/) [[38](#_ENREF_38" \o "Moszer, 2002 #68)]. [[39](#_ENREF_39" \o "Kobayashi, 2003 #83)]. Spore coat genes in *B. subtlis* were annotated in the Driks group (see below). After excluding the spore coat genes and essential genes, the remaining genes are referred as non-coat and non-essential (nonCE) genes. The lists of *B. subtilis* spore coat, essential, and nonCE genes are provided in supplementary information.

**2.2 Orthologous identification**

We performed pairwise BLASTP search [[40](#_ENREF_40" \o "Altschul, 1997 #89)] for all genomes to find potential homologs. Because the BLAST E-value is influenced by particular database sizes [[41](#_ENREF_41" \o "Kerfeld, 2011 #291)], and because bacterial genomes usually contains about 5000 genes, we chose an E-value cutoff of 1/5000 that means approximately one false positive hit per bacterial genome. It is worth mentioning that E-value ranges would only affect a tiny fraction of the BLAST hits. For example, in *B. subtilis* 168 and *B. amyloliquefaciens* FZB42 comparison, the 90%, 95%, and 99.9% quantiles of BLAST E-values are 2.3x10-6, 3.0x10-5, and 2x10-4 respectively.

Potential orthologs were identified both by Markov clustering (MCL) [[42](#_ENREF_42" \o "Enright, 2002 #86)] and reciprocal best hits (RBH) [[43](#_ENREF_43" \o "Bork, 1998 #87), [44](#_ENREF_44" \o "Tatusov, 1997 #88)]. We iterated the Inflation parameter (I) of MCL from 1.1 to 8.0 to explore the granular effect on gene clusters. We started with 73 coat protein genes. We found that a value, I=3.1, is the smallest value that can give the largest number of orthologous groups in the coat protein genes, 70 clusters, (3 cluster contain duplicates even at very stringent I values). We distinguished orthologs from paralogs by comparing the bootstrapped neighbor-joining trees of the candidate orthologs to the species reference trees and its alternatives (see also Results section xxx). Topological comparison of trees was carried out using the program ‘treedist’ from the PHYLIP software package [[45](#_ENREF_45" \o "Felsenstein, 2005 #207)]. Examination of the multiple sequence alignments showed that many of the unresolved gene trees are due to repeat sequences, also known as low complexity regions (LCRs), in coat proteins. Because some coat proteins tend to contain a substantial number of repeats, filtering out these repeat regions during BLASTP searches would result in a reduction of the numbers of detectable hits [[46](#_ENREF_46" \o "Moreno-Hagelsieb, 2008 #91)]. To avoid this problem, we included repetitive sequence regions during BLASTP searches, used bootstrapping to mitigate the topological uncertainly due to alignment problems caused by repeats, compared the topology between gene trees and the species trees, and excluded the topological inconsistent hits as ‘false positives’. In addition, all of the phylogenies of coat protein orthologous groups were double-checked visually. This visual examination led to the identification of a split ORF in one coat protein gene (see supporting information).

For orthologous identification of non-coat protein genes, only automated analyses were used, but the repetitive sequences were filtered out during BLASTP searches to improve specificity.

We also attempted to use PSI-BLAST (with up to 5 iterations) [[40](#_ENREF_40" \o "Altschul, 1997 #89)] to search for remote orthologs of coat protein genes. This approach provided fewer than 1% more hits than BLAST that passed the phylogeny tests, and in many cases grouped together 1-2% of genes in the genome. Therefore, we did not find this approach to be useful for our purposes.

Among the 73 coat proteins in *B. subtilis*, six were closely related paralogs that cannot be separated into orthologous groups. Hence, we obtained 70 orthologous clusters (three clusters contain two orthologous groups). The pair BG13471 (CotU) and BG10492 (CotC) are so similar that their orthologs in *B. licheniformis* were arbitrarily chosen for further analysis.

**2.3 Implementation**

Statistical analyses and data visualization were largely performed in the R language and environment [[47](#_ENREF_47" \o "R Development Core Team, 2009 #206)]. Batch jobs were automated using PERL and shell scripts in LINUX/UNIX platforms. Phylogenies were mostly generated by neighbor-joining and evaluated by bootstraps in PHYLIP [[45](#_ENREF_45" \o "Felsenstein, 2005 #207)] and APE [[48](#_ENREF_48" \o "Paradis, 2004 #208)]. Visualization of phylogeny was done using MEGA [[49](#_ENREF_49" \o "Tamura, 2007 #93), [50](#_ENREF_50" \o "Kumar, 2008 #92)], Dendroscope [[51](#_ENREF_51" \o "Huson, 2007 #210)], and APE. Nested model tests on adaptive changes were performed using PAML [[52-54](#_ENREF_52" \o "Yang, 2006 #101)]. Initial clustering of sequence was done using MCL [[42](#_ENREF_42" \o "Enright, 2002 #86)] and PERL scripts. Protein statistics were calculated by pepstats from EMBOSS [[55](#_ENREF_55" \o "Rice, 2000 #211)]. Disordered regions in proteins were predicted using DisEMBL [[56](#_ENREF_56" \o "Linding, 2003 #243)]. LCR regions were calculated using xnu [[57](#_ENREF_57" \o "Claverie, 1993 #155)]. Handling of sequences and automation were done largely by PERL scripts in conjunction with Bioperl [[58](#_ENREF_58" \o "Stajich, 2002 #304)]. A small fraction of Python/BioPython codes were also used, especially for the topological analysis [[59](#_ENREF_59" \o "Talevich, 2012 #305), [60](#_ENREF_60" \o "Cock, 2009 #306)].

**2.4 Public project repository**

In addition to the supplementary information, we created a GitHub repository, https://github.com/hongqin/BacillusSporeCoat. This GitHub repository contains the list of annotated spore coat genes, their sequences in FASTA formats, gene trees, running results, and key scripts for data analysis.

# 3. Results

**3.1 Inference of the species reference tree**

A defined species reference tree is important in phylogenetic analysis [[52](#_ENREF_52), [53](#_ENREF_53)]. However, species trees of bacteria are difficult to construct [[61](#_ENREF_61)]. The *B. cereus* sensu lato group is known to be very closely related. Sequence variations suggest that the *B. cereus* sensu lato group is a group of asexual clonal lineages [[62](#_ENREF_62)]. The *B. cereus* is also known to be an intermingled cluster of genetically diverse strains [[63](#_ENREF_63)]. To facilitate appropriate molecular evolution analysis, our aim here is to infer a species reference tree only for the strains whose complete genomes are available to our study.

We tried both the 16s rRNA approach and the multi-locus approach [[61](#_ENREF_61)]. The 16s rRNA is often used for species identification in the *Bacillus* genus [[12](#_ENREF_12), [64-66](#_ENREF_64)]. We curated 148 sequences of the 16s ribosomal RNA sequences for *Bacillaceae* and their related species. *Alicyclobacillus acidocaldariu* and *Geobillus kaustophilus* were used outgroups. Structure-based alignment of these rRNAs was obtained from the Ribosomal Database Project (http://rdp.cme.msu.edu/) [[67](#_ENREF_67)]. Trees were generated using neighbor-joining, maximal parsimony and Bayesian approaches [[49](#_ENREF_49), [50](#_ENREF_50), [68-70](#_ENREF_68)]. Neighbor-joining tree was evaluated by bootstrap [[71](#_ENREF_71)]. Positioning of *B. cereus ATC 14579* is inconsistent among neighbor-joining and parsimony, and Bayesian trees. Hence, we found that 16sRNAs could not resolve the branch patterns within the *B. cereus*-group, even though this partially resolved species tree is generally in agreement with previous results using the 16s rRNAs [[12](#_ENREF_12), [65](#_ENREF_65), [66](#_ENREF_66)].

For the multi-locus approach, we curated a list of 34 essential genes in *B. subtilis* that had unequivocally single-orthologs in all of the genomes under study and could yield resolved phylogenetic trees based on bootstrap resample of their protein sequences. We concatenated coding sequences of these 34 genes and obtained 11 super-genes with about 36.6 Kb in length. The neighbor-joining tree of these concatenated sequences is 100% supported by bootstrap resampling (Figure 1) and is used as the resolved species reference tree. This concatenation based approach is the state-of-art method for multi-locus tree inference [[72](#_ENREF_72)]. In this resolved tree, the ATCC 14579 type strain of *B. cereus* is positioned next to *B. weihenstephanesis* KBAB4, and *B. anthracis* and *B.* *thuringiensis* konkukian are next to each other, which is similar to a neighbour-joining tree based on concatenated sequences of 7 house-keeping genes [[73](#_ENREF_73)]. This species tree is further supported by our clustering results, based on the coat protein phylogenetic profile (Figure 2). *B.* *thuringiensis* konkukian is also reported to be close to *B. anthracis* [[74](#_ENREF_74)].

We are aware that many genes in microbial genomes can have different gene trees, which can pose a problem if we use only one reference gene tree for ortholog identification. Based on the neighbor-joining trees of individual spore coat genes, we observed that alternative branching patterns frequently occur within the *B. subtilis* and *B. cereus* clades, but not between these two clades. Hence, we tested whether these alternative tree topologies are statistically acceptable using CONSEL [[75](#_ENREF_75)] in the 34 essential genes (Table 2). A total of 10 topologies were tested using the AU-test provided by CONSEL. Overall, alternative branching patterns within the two major clades are often accepted, but those occur between the two major clades (such as the 10th tree topology) are consistently rejected at p-value = 0.05 level. The species reference tree in Figure 1 (the 1st tree in Table 2) is ranked the highest 20 out of 33 times, and is only rejected 1 out 34 times at p-value = 0.05 level. Therefore, for orthologous identification, we accepted trees with alternative branching patterns within the two major clades.

**3.2 Phylogenetic profiling of spore coat proteins**

Previous work shows there are at least forty-five coat proteins in *B. subtilis* [[76](#_ENREF_76" \o "Kim, 2006 #9)]. Using sequence similarity criteria, and data from microarrays studies identifying genes of unknown function that are expressed late in sporulation [[77](#_ENREF_77" \o "Eichenberger, 2004 #198), [78](#_ENREF_78" \o "Eichenberger, 2003 #215)], we compiled an expanded list of 73 genes (see supplementary data), that includes genes we regard as strong candidates for coat protein genes [[76](#_ENREF_76" \o "Kim, 2006 #9), [77](#_ENREF_77" \o "Eichenberger, 2004 #198)]. Over 80% (60 out of 73) of these genes were annotated as spore coat protein genes independently by another group [[14](#_ENREF_14" \o "Henriques, 2007 #55)].

Analysis of the distributions of protein orthologs among species, i.e. the phylogenetic profile, can give important insights into evolution of these proteins and help identify those proteins with essential functional roles. Previous profile analysis of spore coat genes were based on [[14](#_ENREF_14" \o "Henriques, 2007 #55)]Because orthologs are genes in different species that are derived from a single ancestral gene [[79](#_ENREF_79" \o "Fitch, 2000 #99)], an orthologous relationship is by definition determined by phylogeny, using molecular evolutionary measures of gene distances [[71](#_ENREF_71" \o "Li, 1997 #97)]. Consequently, we used a phylogeny-based approach to identify orthologous distributions of coat proteins (the set of coat protein orthologs among species with sequenced genomes) in 11 *Bacillus* species.

The resulting coat protein phylogenetic profiles suggest that coat proteins can be partitioned into those that are evolutionarily conserved or labile (Figure 2). The orthologous distribution for each coat protein orthologous group (named after the *B. subtilis* (Bsu) gene IDs) was generated by assigning 1 to each species with detectable orthologous hits and assigning 0 otherwise. The dissimilarities in the coat protein orthologous distributions are significant enough that they can be used to cluster the species in a manner in agreement with the species reference tree (Figure 1).

**3.3 The *B. subtilis* inner coat is more conserved than outer coat.**

We speculated that proteins comprising the outermost structures of the spore would be more evolutionarily labile, since these proteins would be most likely to make direct contact with the environment. If so, this lability might be reflected in the coat protein gene phylogenetic profiles. Specifically, we expected to find that coat proteins closer to the spore surface would be more labile than coat proteins at more interior locations. To test this hypothesis, we first analyzed the phylogenetic profiles of the coat proteins in *B. subtilis*, because it is already known that that many if not most of the outermost proteins in *B. subtilis* are among the already identified outer coat proteins (or outer coat protein candidates) [[14](#_ENREF_14" \o "Henriques, 2007 #55), [17](#_ENREF_17" \o "Driks, 2009 #212), [76](#_ENREF_76" \o "Kim, 2006 #9), [80](#_ENREF_80" \o "Driks, 2002 #20)]. We note that proteins designated in the literature or in genome annotations as members of the outer coat could also be present in the crust, the recently identified and still poorly characterized coat layer surrounding the outer coat [[30](#_ENREF_30" \o "McKenney, 2010 #256)]. Although, in the present study, we chose to avoid confusion with the existing literature by retaining the designation “outer coat proteins” to refer to any coat proteins in layer(s) surrounding the inner coat, we emphasize that future studies are likely to assign at least some of them to the crust, in addition to or instead of the outer coat.

We first tested whether the coat protein phylogenetic profiles are associated with their known (or likely) sub-locations within inner or outer coat layers by constructing a two-by-two table and then analyzing the statistical associations (Table 3). In Table 3, the conserved coat proteins in the *B. cereus*-group are those with orthologous hits in all four species, and the labile coat proteins in the *B. cereus*-group are those missing at least one orthologous hit in the *B. cereus*-group. Consistent with our hypothesis, the conserved coat proteins are enriched in the inner coat, and the labile coat proteins are enriched in the outer coat (Fisher-exact test, p= 0.025).

We are aware that the test in Table 3 can be influenced by the partitioning of coat proteins into conserved and labile categories. To avoid this caveat, we chose to compare the orthologous hits directly (Figure 3). Based on the pooled orthologous hits in the four *B. cereus* genomes, orthologous hits are significantly over-represented among the inner coat proteins.

Based on the above two analyses, we concluded that protein compositions are more conserved in the inner coat layer than the outer coat layer between *B. subtilis*-group and *B. cereus*-group species. We speculate that in allthe species analyzed above, the relatively greater lability of the outer layer protein composition is due to an important role for this layer in adaptation to specific niches. It is possible that the adaptive features of the spore outer layers are, at least in part, a consequence of many coat protein working together, for example, by contributing a particular chemical property to the spore surface [[18](#_ENREF_18" \o "Chen, 2010 #236)]. This could lead to complex adaptive evolutionary events, including positive selection, relaxed negative selection, and loss of negative selection at gene levels. The loss of negative selection on some genes is supported by their lack of orthologous hits in some species.

**3.4 Relatively higher dN/dS ratio of spore coat genes in the *B. subtilis*-group.**

If the diversity in *Bacillaceae* spore coat morphology reflects adaptation of these species to a range of environments, then we may be able to detect signatures of selection from the perspective of molecular evolution. We chose to address this by estimating the ratio of non-synonymous (dN) to synonymous (dS) substitution rates, ω, a proxy for selective pressure [[81](#_ENREF_81" \o "Hurst, 2002 #303)]. An increase in ω can suggest a relatively faster non-synonymous substitution rate due to either relaxed negative selection or positive selection in divergent species [[81](#_ENREF_81" \o "Hurst, 2002 #303)].

First, we compared ω between coat proteins and non-coat non-essential (nonCE) proteins among the 10 species with complete genomes (Figure 4A). We calculated p-values using the one-sided Wilcoxon non-parametric test (alternative hypothesis: coat ω > nonCE ω ). Although simple pairwise comparisons usually can not narrow down evolutionary events to specific branches, the matrix approach used here can detect differences between clades. In Figure 4A, the patterns in the *B. subtilis*-group and the *B. cereus*-group are clearly opposite. In the *B. subtilis*-group, the p-values are mostly less than 0.05 and the alternative hypothesis is accepted. Hence, coat proteins show higher ω than do nonCE proteins. In the *B. cereus*-group, the p-values are mostly greater than 0.95, which means the opposite alternative hypothesis, coat ω < nonCE ω, should be accepted. Hence, the patterns of coat protein gene evolution differ between the *B. subtilis*- and *B. cereus*-groups. These contrasting ω patterns held when additional *B. cereus* genomes were included in the analysis (Figure S5A??). As expected, the contrasting evolutionary patterns of coat protein genes are not pronounced in pairwise tests of dN measures, and are absent in pairwise tests of dS measures (Figure S5B?? and S5C??). For comparison, the ω of essential genes are significantly lower than those of nonCE genes (with an exception in the *B. weihenstephanensis* lineage) (Figures 4B and S5D??), further validating this pairwise matrix approach. These results show that negative selection pressure on coat proteins is significantly stronger in the *B. cereus*-group than in the *B. subtilis*-group.

Second, we calculated the likelihood of different evolutionary patterns, designed in nested branch models in *codeml*, and applied likelihood ratio tests (LRTs) [[53](#_ENREF_53" \o "Bielawski, 2005 #82)]. We are aware that the LRT method detects changes only within each gene family (not between two different groups of genes), and are therefore more conservative than the pairwise analysis. Meaningful LRTs should be done using the same model, i.e, the same tree topology, which constrained us to focus LRTs on conserved genes. We selected 1174 conserved gene families whose neighbor-joining gene trees agree with the species reference tee, and also contain an orthologous hit in the outgroup *B. halodurans*. These conserved gene families include 19 coat and 182 essential gene families. We then calculated their likelihood for four nested branch models: H0, H1c, H1s, and H2 (Figure 5A) using *codeml* [[53](#_ENREF_53" \o "Bielawski, 2005 #82)]. The results at a false-discovery rate of 0.05 are summarized in Figure 5B. LRT show that xx genes (including 5 spore coat protein genes) show significantly different ω in the *B. cereus*-group (model H1c), and xx genes (including 8 spore coat protein genes) show significantly different ω in the *B. subtilis*-group (model H1s). We did not observe any spore coat gene family with ω greater than 1 in any branches. Hence, this result is more consistent with the argument of relaxation of negative selection.

In summary, by comparing the levels of selective pressure on coat protein genes versus nonCE genes among various *Bacillus* species and along phylogeny, we can argue there are contrasting patterns of coat protein gene evolution between the *B. subtilis-* and *B. cereus*-groups. Because *B. subtilis*-group spores lack the exosporium, the significantly relaxed negative selection pressure in coat genes can be attributed to the removal of a structural constraint, and is probably driven by a change of ecological niche, ultimately. Overall, we postulate that differential selection pressure on coat protein genes reflect differences between the ecological life styles of *B. subtilis*- and *B. cereus*-group species.

**4. Discussion**

We demonstrated a strong association between the structural diversity of the spore coat and the evolutionary patterns of its protein components between the *B. subtilis*-group and *B. cereus*-group (Figure 6), by two lines of evidences: 1) In *B. subtilis*, the inner coat is more conserved than the outer coat based on phylogenetic profiles (Table 3 and Figure 3); 2) Coat protein genes have significantly higher ratio of nonsynonymous versus synonymous substitution rates than nonCE genes in *B. subtilis*-group but not in the *B. cereus*-group (Figure 4), which can be attributed to relaxed negative selection as suggested by nested model analysis (Figure 5). Because species in the *B. subtilis*-group lack the exosporium, the coat might have more freedom to evolve. This is an appealing possibility given its likely importance in the interaction with environment species without exosporia (Figure 6). Even in exosporium-bearing species, the coat still makes significant (albeit indirect) contact with the environment, since the exosporium permits diffusion of small molecules. Nonetheless, in the absence of the exosporium, the coat surface likely has direct roles in adhesion to surfaces in the environment. As already discussed, *B. subtilis* possesses a recently discovered outermost coat layer called the crust, which is composed, at least in part, of proteins presently designated as outer coat proteins [[30](#_ENREF_30" \o "McKenney, 2010 #256)]. The current ambiguity in assignment of coat proteins to the crust or outer coat layer does not affect the conclusions of our work. However, as the composition of the crust becomes clarified in future studies, we may learn that its evolutionary history has features that distinguish it from the true outer coat.

In the interpretations just described, we have assumed that  is an accurate reflection of the strength of selection. However, other interpretations are possible. The genomes of the *B. cereus*-group are relatively closely related, whereas genomes in *B. subtilis*– group are more divergent. In closely related bacteria, increased  are often observed, which can be attributed to changes in effective population size, relaxation of negative selection, differences in divergent time, or limitations of parametric evolution models [[82](#_ENREF_82" \o "Rocha, 2006 #152)]. For closely related genomes of asexual organisms, negative selection will not have enough time to “purify” the deleterious mutations and thereby lead to relatively high . This is similar to the mistreatment of standing polymorphism as fixed change in diploid sexual organisms. This problem is at least partially due to a bias in current genome sequencing efforts towards those genomes with perceived medical relevance. Moreover, it is important to emphasize that species identification remains a commonly encountered and significant challenge in bacterial genome analysis. Species misidentification can lead to mistreating polymorphism as divergence which, in turn, leads to false-positive signatures of selection. We have sought to mitigate this problem by focusing on the genomes of well-established species-type strains. We are aware that genomes of many more Bacillus strains have been sequenced recently. However, most of these are assigned to the species that have been studied here, and nucleotide changes in many of these genomes should be treated as polymorphisms.

An important outcome from our studies is the finding that most if not all coat proteins are under significant selective pressure. Consequently, even though only a small fraction of coat protein gene mutations have phenotypes that are readily detectable in the laboratory [[15](#_ENREF_15" \o "Driks, 1999 #30), [28](#_ENREF_28" \o "Traag, 2010 #237)], it is reasonable to assert that most or all coat proteins have functions in the spore and are not, therefore, redundant. We were unable to find a correlation between the known phenotype of each coat protein gene mutation and its degree of conservation. However, this is not surprising, as coat protein gene mutants are rarely if ever analyzed using ecologically realistic assays [[17](#_ENREF_17" \o "Driks, 2009 #212)].

The prevalence of disordered regions in coat proteins is notable and raises the possibility that these regions provide a function needed by most or all coat proteins. We do not know whether this is so, but we note that in some cases (such as the muscle protein titin [[83](#_ENREF_83" \o "Lee, 2010 #251)]) disordered regions can act effectively as springs that unfold when stretched by an external force. If this were true for the disordered regions within coat proteins, then such regions could at least partially explain the ability of the coat to fold and unfold like the pleats of an accordion, since this dynamic action likely involves stretching and contract the coat at specific points within each fold (Sahin and Driks, unpublished observations) [[84-89](#_ENREF_84" \o "Chada, 2003 #14)]. Interestingly, intrinsically disordered regions of proteins can evolve more rapidly than other regions, because disordered regions have fewer constraints than ordered regions [[90](#_ENREF_90" \o "Brown, 2010 #241)]. However, because the alignment between sequences with intrinsic disorder may be poorer than other regions, evolutionary rates can appear faster than they actually are. Further analysis will be needed to determine whether these regions within coat proteins are, indeed, evolving faster than other regions.

Our work raises several intriguing questions for future studies. First, what are the broader biological and functional implications of the different evolutionary patterns of spore coat proteins among different *Bacillaceae* clades? Second, does exosporium protein evolution follow a trend that is similar to the outer coat in *B. subtilis*, as we would predict? In future studies, we will apply the approach described here to those proteins, to determine not only whether they evolve more rapidly than coat proteins, but also whether different rates of evolution can be detected within the exosporium sublayers.

We searched for adaptive changes in coat proteins in the two species groups using the site-model implemented in PAML [[52](#_ENREF_52" \o "Yang, 2006 #101), [53](#_ENREF_53" \o "Bielawski, 2005 #82)]. The current number species in each group is too small to offer a reasonable statistical power. We hope to revisit this question when genomes of many more Bacillus species become available.

One of the most interesting consequences of this work is the likely role for the outer coat and crust proteins in variation among spores of the Bacillaceae. The phylogenomic approach employed in this study is likely to be very useful to further investigations into the divergent ecological histories and patterns of adaptation among spore-forming bacteria. We hope that this work prompts deeper investigations into poorly studied species with intriguing lifestyles and poorly studied ecological niches [[13](#_ENREF_13" \o "Driks, 2007 #6)].

**5 Authors' contributions.**

HQ and AD designed the study, HQ performed the study, and HQ and AD wrote the manuscript.

**6 Acknowledgements.**

We thank Michele Mock and Marie Moya-Nilges for comments, Richard Schultz for enlightening discussions, Emmanuel Paradis for helps on the APE package in R, and Hongwe Wu on usage of MCL. HQ was partially supported by a NCMHD grant (5P20MD000215-05) given to the Spelman Center for Health Disparities Research and Education and a seed grant from the Spelman ASPIRE program.

# 7. References

**8 Links.**

Sequences, alignment for the annotated spore coat proteins can be downloaded at the following site:

GitHub xxxxx

**Table 1**. Key genomes analyzed in this study.

|  |  |  |
| --- | --- | --- |
| Genome | Abbreviation | Source |
| *Bacillus amyloliquefaciens* FZB42 | Bam | Genbank NC\_009725 |
| *Bacillus anthracis* Ames | Ban | JCVI CMR |
| *Bacillus cereus* ATCC 14579 | Bce | JCVI CMR |
| *Bacillus cereus* ATCC 10987 | Bce87 | JCVI CMR |
| *Bacillus cereus* ATCC E33L | Bce3L | JCVI CMR |
| *Bacillus clausii* KSM-K16 | Bcl | JCVI CMR |
| *Bacillus halodurans* C-125 | Bha | JCVI CMR |
| *Bacillus licheniformis* ATCC14580 | Bli | JCVI CMR |
| *Bacillus pumilus* SAFR-032 | Bpu | Genbank NC\_009848 |
| *Bacillus subtilis* 168 | Bsu | SubtiList |
| *Bacillus thuringiensis* konkukian | Bth | JCVI CMR |
| *Bacillus weihenstephanensis* KBAB4 | Bwe | Genbank NC\_010184 |
| *Bacillus mojavensis* RO-H-1 | Bmo | JCVI CMR |

SubtiList: http://genolist.pasteur.fr/SubtiList/ [[38](#_ENREF_38)]; JCVI CMR: http://cmr.jcvi.org/. Our analysis of the partial genome of *Bacillus mojavensis* RO-H-1 is limited to coat protein orthologs. Except Bce87 and Bce3L, all other genomes are type strains of each species.

**Table 2**. Topologies tested by CONSEL in 34 essential genes. Alternative branching patterns with the *B subtilis*-clade or *B cereus*-clade were often accepted. Alternative branching patterns violating the two major clades, such as in the 10th tree, was consistently rejected.

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Trees in Newick format** | **AU test summary** | |
| 1 | ((Bha,Bcl), (Bpu,(Bli,(Bam,(Bmo,Bsu)))), (Bwe,(Bce,(Ban,Bth)))); | Ranked highest in 20 genes. | Accepted in 33 genes |
| 2 | ((Bha,Bcl), (Bpu,(Bli,(Bam,(Bmo,Bsu)))), ((Bwe,Bce),(Ban,Bth))); | Ranked highest in 7 genes. | Accepted in 33 genes |
| 3 | ((Bha,Bcl), (Bpu,(Bli,(Bam,(Bmo,Bsu)))), (Bwe,(Bth,(Bce,Ban)))); |  | Accepted in 5 genes |
| 4 | ((Bha,Bcl), ((Bpu,Bli),(Bam,(Bmo,Bsu))), (Bwe,(Bth,(Bce,Ban)))); |  | Accepted in 7 genes |
| 5 | ((Bha,Bcl), ((Bpu,Bli),(Bam,(Bmo,Bsu))), ((Bwe,Bth),(Bce,Ban))); |  | Accepted in 4 genes |
| 6 | ((Bha,Bcl), (Bli,(Bpu,(Bam,(Bmo,Bsu)))), (Bwe,(Bce,(Ban,Bth)))); | Ranked highest in 6 genes. | Accepted in 26 genes |
| 7 | ((Bha,Bcl), (Bli,(Bpu,(Bam,(Bmo,Bsu)))), ((Bwe,Bce),(Ban,Bth))); | Ranked highest in 1 gene. | Accepted in 22 genes |
| 8 | ((Bha,Bcl), (Bam,(Bli,(Bpu,(Bmo,Bsu)))), (Bwe,(Bce,(Ban,Bth)))); |  | Accepted in 5 genes |
| 9 | ((Bha,Bcl), (Bam,(Bpu,(Bli,(Bmo,Bsu)))), (Bwe,(Bce,(Ban,Bth)))); |  | Accepted in 4 genes |
| 10 | ((Bha,Bcl), (Bwe,(Bli,(Bam,(Bmo,Bsu)))), (Bpu,(Bce,(Ban,Bth)))); | Ranked lowest in 33 genes. | Accepted in 0 genes |

**Table 3** Inner coat genes are more conserved than outer coat genes, as shown by the spore coat location of the evolutionary conserved and labile coat genes.

|  |  |  |
| --- | --- | --- |
| Coat Proteins | Inner Coat | Outer Coat |
| Conserved in the *B. cereus*-group | 17 | 8 |
| Labile in the *B. cereus*-group | 6 | 12 |
| Subtotal | 23 | 20 |
| One sided Fisher’s exact test, p-value=0.025 | | |

The “conserved” coat genes are those with orthologous in all of the four species in the *B*. *cereus*-group. The “labile” coat genes are those with at least one missing orthologous hit in the *B. cereus*-group.



Figure 1, species reference tree

**Figure S1 The** species/reference tree of the *Bacillus* genomes under study based on 34 concatenated essential genes. The Newick format of the species tree is (((Bpu,(Bli,(Bam,(Bmo,Bsu)))),(Bwe,(Bce,Ban,Bth))),(Bha,Bcl)). The evolutionary history was inferred using the neighbor-joining method [[91](#_ENREF_91), [92](#_ENREF_92)]. The optimal tree with the sum of branch length = 1.58 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the JTT matrix-based method and are in the units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated from the dataset (Complete deletion option). There were a total of 11302 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 [[49](#_ENREF_49)].



Figure 2. Orthologous profiling of coat proteins. Blue indicates ‘1’ (the presence of orthologous hits) and red indicates ‘0’ (the absence of detectable orthologous hits). Hierarchical clustering using average linkage and hamming distances is applied both by row, which yields groups of coat proteins, and by column, which show relatedness of species. Each row is for one spore coat protein. (add conserved and labile grouping into the figure).

Figure 3 Comparison of orthologous hits per gene between inner and outer coat layers. For each *B. subtilis* coat protein, t Each bin represent the number of B. subtilis coat protein with the indicated number of otherohlogus hits. (XXXXXXwork here)

Figure 4. The dN/dS ratio, ω, of coat proteins is higher than average in the *B. subtilis*-group but not in the *B. cereus*-group. (**A**) The p-values of pairwise Wilcoxon non-parametric tests of coat proteins between genomes (Alternative hypothesis: coat protein gene ω > nonCE gene ω). Coat proteins have higher ω than non-coat non-essential proteins in most genomes but lower ω than non-coat non-essential proteins in *B. cereus*-group genomes. Each cell represents the p-value of a one-sided Wilcoxon non-parametric test between two genomes. Most cells within the *B. subtilis*-group are in red with p-values less than 0.05, indicating the coat protein gene ω > the nonCE gene ω. Most cells within the *B. cereus*-group are green with p-values greater than 0.95, indicating the opposite is significant: coat protein gene ω < nonCE gene ω. The diagonal cells would suggest self-comparison and, hence, are excluded. (**B**) The p-values of pairwise Wilconxon non-parametric tests of essential genes between genomes (Alternative hypothesis: essential gene ω > nonCE gene ω). Essential genes generally have smaller ω and evolve slower than nonCE genes. An exception occurs in the *B. weihenstephanensis* branch, where essential genes have higher ω than nonCE genes.

Figure 5, Differential selective pressures between the *B. subtilis*-group and the *B. cereus*-group using nested models and likelihood ratio tests. (A) Specification of ω along branches. (B) The number of genes that accept the alternative models (H1c, H1s, H2) are listed in their circles, respectively. XXXXadd more here on LRT results. Explain dashed line, circles, etc. XXXX

Figure 6. Structural differences in outer-spore layers are plausible causes for the observed evolutionary differences in coat protein genes between the *B. subtilis*-group and *B. cereus*-group.

1. Klobutcher LA, Ragkousi K, Setlow P: **The *Bacillus subtilis* spore coat provides "eat resistance" during phagocytic predation by the protozoan *Tetrahymena thermophila***. *Proc Natl Acad Sci U S A* 2006, **103**(1):165-170.

2. Nicholson WL, Munakata N, Horneck G, Melosh HJ, Setlow P: **Resistance of *Bacillus* endospores to extreme terrestrial and extraterrestrial environments**. *Microbiol Mol Biol Rev* 2000, **64**(3):548-572.

3. Losick R, Youngman P, Piggot PJ: **Genetics of endospore formation in *Bacillus subtilis***. *Annu Rev Genet* 1986, **20**:625-669.

4. Claus D, Berkeley RCW: **Genus *Bacillus* Cohn 1872**. In: *Bergey's Manual of Systematic Bacteriology.* Edited by P.H.A. S, Mair NS, Sharpe ME, G. HJ, vol. 2. Baltimore: Williams & Wilkins; 1986: 1105-1139.

5. Moir A: **How do spores germinate?** *J Appl Microbiol* 2006, **101**(3):526-530.

6. Fritze D: **Taxonomy of the Genus *Bacillus* and Related Genera: The Aerobic Endospore-Forming Bacteria**. *Phytopathology* 2004, **94**:1245-1248.

7. Tourasse NJ, Helgason E, Okstad OA, Hegna IK, Kolsto AB: **The *Bacillus cereus* group: novel aspects of population structure and genome dynamics**. *J Appl Microbiol* 2006, **101**(3):579-593.

8. Mock M, Fouet A: **Anthrax**. *Annu Rev Microbiol* 2001, **55**:647-671.

9. Stenfors Arnesen LP, Fagerlund A, Granum PE: **From soil to gut: *Bacillus cereus* and its food poisoning toxins**. *FEMS Microbiol Rev* 2008, **32**(4):579-606.

10. Aronson AI, Shai Y: **Why *Bacillus thuringiensis* insecticidal toxins are so effective: unique features of their mode of action**. *FEMS Microbiol Lett* 2001, **195**(1):1-8.

11. Sonenshein AL, Hoch JA, Losick R: ***Bacillus subtilis* and its closest relatives**. Washington: American Society for Microbiology; 2002.

12. Blackwood KS, Turenne CY, Harmsen D, Kabani AM: **Reassessment of sequence-based targets for identification of *Bacillus* species**. *J Clin Microbiol* 2004, **42**(4):1626-1630.

13. Driks A: **Surface appendages of bacterial spores**. *Mol Microbiol* 2007, **63**(3):623-625.

14. Henriques AO, Moran CP, Jr.: **Structure, assembly, and function of the spore surface layers**. *Annu Rev Microbiol* 2007, **61**:555-588.

15. Driks A: ***Bacillus subtilis* spore coat**. *Microbiol Mol Biol Rev* 1999, **63**(1):1-20.

16. Driks A: **Maximum shields: the assembly and function of the bacterial spore coat**. *Trends Microbiol* 2002, **10**(6):251-254.

17. Driks A, Mallozzi M: **Outer structures of the *Bacillus anthracis* spore.** In: *Bacillus anthracis and Anthrax.* Edited by Bergman N. New Jersey: John Wiley & Sons; 2009.

18. Chen G, Driks A, Tawfig K, Mallozzi M, Patil S: ***Bacillus anthracis* and *Bacillus subtilis* Spore surface properties and transport.** *Colloids and Surfaces B: Biointerfaces* 2010, **76**(2):512-518.

19. Ragkousi K, Eichenberger P, van Ooij C, Setlow P: **Identification of a new gene essential for germination of *Bacillus subtilis* spores with Ca2+-dipicolinate**. *J Bacteriol* 2003, **185**(7):2315-2329.

20. Riesenman PJ, Nicholson WL: **Role of the spore coat layers in *Bacillus subtilis* spore resistance to hydrogen peroxide, artificial UV-C, UV-B, and solar UV radiation**. *Appl Environ Microbiol* 2000, **66**(2):620-626.

21. Setlow B, Atluri S, Kitchel R, Koziol-Dube K, Setlow P: **Role of dipicolinic acid in resistance and stability of spores of *Bacillus subtilis* with or without DNA-protective alpha/beta-type small acid-soluble proteins**. *J Bacteriol* 2006, **188**(11):3740-3747.

22. Behravan J, Chirakkal H, Masson A, Moir A: **Mutations in the gerP locus of *Bacillus subtilis* and *Bacillus cereus* affect access of germinants to their targets in spores**. *J Bacteriol* 2000, **182**(7):1987-1994.

23. Setlow P: **Spores of *Bacillus subtilis*: their resistance to and killing by radiation, heat and chemicals**. *J Appl Microbiol* 2006, **101**(3):514-525.

24. Cohn F: **Studies on the biology of the *Bacilli***. *Beiträge zur Biologie der Pflanzen* 1876, **2**:249-276.

25. Koch R: **The etiology of anthrax, based on the life history of *Bacillus anthracis***. *Beiträge zur Biologie der Pflanzen* 1876, **2**:277-310.

26. Holt SC, Leadbetter ER: **Comparative ultrastructure of selected aerobic spore-forming bacteria: a freeze-etching study**. *Bacteriol Rev* 1969, **33**(2):346-378.

27. Aronson AI, Fitz-James P: **Structure and morphogenesis of the bacterial spore coat**. *Bacteriol Rev* 1976, **40**:360-402.

28. Traag BA, Driks A, Stragier P, Bitter W, Broussard G, Hatfull G, Chu F, Adams KN, Ramakrishnan L, Losick R: **Do mycobacteria produce endospores?** *Proc Natl Acad Sci U S A* 2010, **107**(2):878-881.

29. Warth AD, Ohye DF, Murrell WG: **Location and composition of spore mucopeptide in *Bacillus* species**. *J Cell Biol* 1963, **16**:593-609.

30. McKenney PT, Driks A, Eskandarian HA, Grabowski P, Guberman J, Wang KH, Gitai Z, Eichenberger P: **A distance-weighted interaction map reveals a previously uncharacterized layer of the *Bacillus subtilis* spore coat**. *Curr Biol* 2010, **20**(10):934-938.

31. Hannay CL: **The parasporal body of *Bacillus laterosporus* Laubach**. *J Biophys Biochem Cytol* 1957, **3**:1001-1010.

32. Vary PS: **Prime time for *Bacillus megaterium***. *Microbiology* 1994, **140 ( Pt 5)**:1001-1013.

33. Oliva CR, Swiecki MK, Griguer CE, Lisanby MW, Bullard DC, Turnbough CL, Jr., Kearney JF: **The integrin Mac-1 (CR3) mediates internalization and directs *Bacillus anthracis* spores into professional phagocytes**. *Proc Natl Acad Sci U S A* 2008, **105**:1261-1266.

34. Bozue J, Moody KL, Cote CK, Stiles BG, Friedlander AM, Welkos SL, Hale ML: ***Bacillus anthracis* spores of the bclA mutant exhibit increased adherence to epithelial cells, fibroblasts, and endothelial cells but not to macrophages**. *Infect Immun* 2007, **75**(9):4498-4505.

35. Ball DA, Taylor R, Todd SJ, Redmond C, Couture-Tosi E, Sylvestre P, Moir A, Bullough PA: **Structure of the exosporium and sublayers of spores of the *Bacillus cereus* family revealed by electron crystallography**. *Mol Microbiol* 2008, **68**:947-958.

36. Driks A: **Proteins of the spore core and coat**. In: *Bacillus subtilis and its closest relatives.* Edited by Sonenshein AL, Hoch JA, Losick R. Washington, D.C: American Society for Microbiology; 2002: 527-536.

37. Salzberg SL, Delcher AL, Kasif S, White O: **Microbial gene identification using interpolated Markov models**. *Nucleic Acids Res* 1998, **26**(2):544-548.

38. Moszer I, Jones LM, Moreira S, Fabry C, Danchin A: **SubtiList: the reference database for the *Bacillus subtilis* genome**. *Nucleic Acids Res* 2002, **30**(1):62-65.

39. Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P *et al*: **Essential *Bacillus subtilis* genes**. *Proc Natl Acad Sci U S A* 2003, **100**(8):4678-4683.

40. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

41. Kerfeld CA, Scott KM: **Using BLAST to teach "E-value-tionary" concepts**. *PLoS biology* 2011, **9**(2):e1001014.

42. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families**. *Nucleic Acids Res* 2002, **30**(7):1575-1584.

43. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y: **Predicting function: from genes to genomes and back**. *J Mol Biol* 1998, **283**(4):707-725.

44. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families**. *Science* 1997, **278**(5338):631-637.

45. Felsenstein J: **PHYLIP (Phylogeny inference package) version 3.6**. *Distributed by the author Department of Genome Sciences, University of Washington, Seattle* 2005.

46. Moreno-Hagelsieb G, Latimer K: **Choosing BLAST options for better detection of orthologs as reciprocal best hits**. *Bioinformatics* 2008, **24**(3):319-324.

47. R Development Core Team: **R: A language and environment for statistical computing**. [*http://wwwR-projectorg*](http://wwwR-projectorg)2009.

48. Paradis E, Claude J, Strimmer K: **APE: analyses of phylogenetics and evolution in R**. *Bioinformatics* 2004, **20**:289-290.

49. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0**. *Mol Biol Evol* 2007, **24**(8):1596-1599.

50. Kumar S, Nei M, Dudley J, Tamura K: **MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences**. *Brief Bioinform* 2008, **9**(4):299-306.

51. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R: **Dendroscope: An interactive viewer for large phylogenetic trees**. *BMC Bioinformatics* 2007, **8**:460.

52. Yang Z: **Computational Molecular Evolution**. Oxford: Oxford University Press; 2006.

53. Bielawski JP, Yang Z: **Maximum Likelihood Methods for Detecting Adaptive Protein Evolution**. In: *Statistical Methods in Molecular Evolution.* Edited by Nielsen R: Springer; 2005: 103-124.

54. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood**. *CABIOS* 1997, **13**:555-556.

55. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet* 2000, **16**(6):276-277.

56. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB: **Protein disorder prediction: implications for structural proteomics**. *Structure* 2003, **11**(11):1453-1459.

57. Claverie JM, States D: **Information enhancement methods for large scale sequence analysis**. *Computers and Chemistry* 1993, **17**:191-201.

58. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H *et al*: **The Bioperl toolkit: Perl modules for the life sciences**. *Genome research* 2002, **12**(10):1611-1618.

59. Talevich E, Invergo BM, Cock PJ, Chapman BA: **Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython**. *BMC Bioinformatics* 2012, **13**:209.

60. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B *et al*: **Biopython: freely available Python tools for computational molecular biology and bioinformatics**. *Bioinformatics* 2009, **25**(11):1422-1423.

61. Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL *et al*: **Opinion: Re-evaluating prokaryotic species**. *Nat Rev Microbiol* 2005, **3**(9):733-739.

62. Priest FG, Barker M, Baillie LW, Holmes EC, Maiden MC: **Population structure and evolution of the *Bacillus cereus* group**. *J Bacteriol* 2004, **186**(23):7959-7970.

63. Vilas-Boas GT, Peruca AP, Arantes OM: **Biology and taxonomy of *Bacillus cereus*, *Bacillus anthracis*, and *Bacillus thuringiensis***. *Can J Microbiol* 2007, **53**(6):673-687.

64. Priest FG, Kaji DA, Rosato YB, Canhos VP: **Characterization of *Bacillus thuringiensis* and related bacteria by ribosomal RNA gene restriction fragment length polymorphisms**. *Microbiology* 1994, **140 ( Pt 5)**:1015-1022.

65. Goto K, Omura T, Hara Y, Sadaie Y: **Application of the partial 16S rDNA sequence as an index for rapid identification of species in the genus *Bacillus***. *J Gen Appl Microbiol* 2000, **46**(1):1-8.

66. Xu D, Cote JC: **Phylogenetic relationships between *Bacillus* species and related genera inferred from comparison of 3' end 16S rDNA and 5' end 16S-23S ITS nucleotide sequences**. *Int J Syst Evol Microbiol* 2003, **53**(Pt 3):695-704.

67. Cole JR, Chai B, Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, McGarrell DM, Bandela AM, Cardenas E, Garrity GM, Tiedje JM: **The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data**. *Nucleic Acids Res* 2007, **35**(Database issue):D169-172.

68. Swofford DL: **PAUP\*. Phylogennetic Analysis Using Parsimony(\* and Other Methods). Version 4.** In*.* Sunderland, Massachusetts: Sinauer Assocates; 2002.

69. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models**. *Bioinformatics* 2003, **19**(12):1572-1574.

70. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees**. *Bioinformatics* 2001, **17**(8):754-755.

71. Li W-H: **Molecular Evolution**. Sunderland, Massachusetts: Sinauer Associates; 1997.

72. Rannala B, Yang Z: **Phylogenetic inference using whole genomes**. *Annu Rev Genomics Hum Genet* 2008, **9**:217-231.

73. Rasko DA, Ravel J, Okstad OA, Helgason E, Cer RZ, Jiang L, Shores KA, Fouts DE, Tourasse NJ, Angiuoli SV *et al*: **The genome sequence of *Bacillus cereus* ATCC 10987 reveals metabolic adaptations and a large plasmid related to *Bacillus anthracis* pXO1**. *Nucleic Acids Res* 2004, **32**(3):977-988.

74. Han CS, Xie G, Challacombe JF, Altherr MR, Bhotika SS, Brown N, Bruce D, Campbell CS, Campbell ML, Chen J *et al*: **Pathogenomic sequence analysis of *Bacillus cereus* and *Bacillus thuringiensis* isolates closely related to *Bacillus anthracis***. *J Bacteriol* 2006, **188**(9):3382-3390.

75. Shimodaira H, Hasegawa M: **CONSEL: for assessing the confidence of phylogenetic tree selection**. *Bioinformatics* 2001, **17**(12):1246-1247.

76. Kim H, Hahn M, Grabowski P, McPherson DC, Otte MM, Wang R, Ferguson CC, Eichenberger P, Driks A: **The *Bacillus subtilis* spore coat protein interaction network**. *Mol Microbiol* 2006, **59**(2):487-502.

77. Eichenberger P, Fujita M, Jensen ST, Conlon EM, Rudner DZ, Wang ST, Ferguson C, Haga K, Sato T, Liu JS *et al*: **The program of gene transcription for a single differentiating cell type during sporulation in *Bacillus subtilis***. *PLoS Biol* 2004, **2**:e328.

78. Eichenberger P, Jensen ST, Conlon EM, van Ooij C, Silvaggi J, Gonzalez-Pastor JE, Fujita M, Ben-Yehuda S, Stragier P, Liu JS *et al*: **The sigmaE regulon and the identification of additional sporulation genes in *Bacillus subtilis***. *J Mol Biol* 2003, **327**(5):945-972.

79. Fitch WM: **Homology a personal view on some of the problems**. *Trends Genet* 2000, **16**(5):227-231.

80. Driks A: **Overview: Development in bacteria: spore formation in *Bacillus subtilis***. *Cell Mol Life Sci* 2002, **59**(3):389-391.

81. Hurst LD: **The Ka/Ks ratio: diagnosing the form of sequence evolution**. *Trends Genet* 2002, **18**(9):486.

82. Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, Feil EJ: **Comparisons of dN/dS are time dependent for closely related bacterial genomes**. *J Theor Biol* 2006, **239**(2):226-235.

83. Lee EH, Hsin J, von Castelmur E, Mayans O, Schulten K: **Tertiary and secondary structure elasticity of a six-Ig titin chain**. *Biophys J* 2010, **98**(6):1085-1095.

84. Chada VG, Sanstad EA, Wang R, Driks A: **Morphogenesis of bacillus spore surfaces**. *J Bacteriol* 2003, **185**(21):6255-6261.

85. Driks A: **The dynamic spore**. *Proc Natl Acad Sci U S A* 2003, **100**(6):3007-3009.

86. Plomp M, Leighton TJ, Wheeler KE, Pitesky ME, Malkin AJ: ***Bacillus atrophaeus* outer spore coat assembly and ultrastructure**. *Langmuir* 2005, **21**(23):10710-10716.

87. Plomp M, Leighton TJ, Wheeler KE, Malkin AJ: **The high-resolution architecture and structural dynamics of *Bacillus* spores**. *Biophys J* 2005, **88**(1):603-608.

88. Plomp M, Leighton TJ, Wheeler KE, Malkin AJ: **Architecture and high-resolution structure of *Bacillus thuringiensis* and *Bacillus cereus* spore coat surfaces**. *Langmuir* 2005, **21**(17):7892-7898.

89. Westphal AJ, Price PB, Leighton TJ, Wheeler KE: **Kinetics of size changes of individual *Bacillus thuringiensis* spores in response to changes in relative humidity**. *Proc Natl Acad Sci U S A* 2003, **100**(6):3461-3466.

90. Brown CJ, Johnson AK, Daughdrill GW: **Comparing models of evolution for ordered and disordered proteins**. *Mol Biol Evol* 2010, **27**(3):609-621.

91. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees**. *Mol Biol Evol* 1987, **4**(4):406-425.

92. Studier JA, Keppler KJ: **A note on the neighbor-joining algorithm of Saitou and Nei**. *Mol Biol Evol* 1988, **5**(6):729-731.