Evolutionary Remodeling of the Sporulation Initiation Pathway

Philip Davidson
pdavidso@andrew.cmu.edu

August 4, 2017

Department of Biological Sciences
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Dannie Durand, Chair
Luisa Hiller
David Hackney
Michael T. Laub (Massachusetts Institute of Technology)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Keywords: comparative genomics, molecular evolution, protein domain, homology, sporulation initiation (Spo0) pathway

Abstract

Signal transduction pathways allow organisms to sense and respond appropriately to a complex bouquet of environmental cues. The molecular determinants of specificity are constrained by the demands of signaling fidelity, yet flexible enough to allow pathway remodeling to meet novel environmental challenges. A detailed picture of how these forces shape bacterial two-component signaling systems has emerged over the last decade. However, the tension between constraint and flexibility in more complex architectures has not been well-studied.

In this thesis, I combine comparative genomics and *in vitro* phosphotransfer experiments to investigate pathway remodeling using the Firmicutes sporulation initiation (Spo0) pathway as a model. The present-day Spo0 pathways in Bacilli and Clostridia share common ancestry, but possess different architectures. In Clostridia, a sensor kinase phosphorylates Spo0A, the master regulator of the sporulation, directly. In Bacilli, Spo0 is phosphorylated/activated indirectly via a four-protein phosphorelay. The presence in sister lineages of signaling pathways that activate the same response regulator and control analogous phenotypes, yet possess with different architectures, suggests a common ancestral pathway that evolved through interaction remodeling. The prevailing theory is that the ancestral pathway was a simpler, direct phosphorylation architecture; the more complex phosphorelay emerged within the Bacillar lineage.

In contrast to this prevailing view, my analysis of 84 representative genomes supports a novel hypothesis for the evolution of Spo0 architectures, wherein the two protein, direct phosphorylation architecture is a derived state, which arose from an ancestral Spo0 phosphorelay. The combination of my bioinformatic analysis and the first experimental characterization of a Clostridial phosphorelay provide evidence for the presence of functional phosphorelays in both classes Bacilli and Clostridia. Further, a cross-species complementation assay between phosphorelays from each class suggests that interaction specificity has been conserved since the divergence of this phylum, 2.7 BYA.

My results reveal a patchy phylogenetic distribution of both Spo0 pathway architectures, consistent with repeated remodeling events, in which a phosphorelay was replaced with a two protein, direct phosphorylation pathway. This remodeling likely occurred via acquisition of a sensor kinase with direct specificity for Spo0A.

Further, my analysis suggests that the unusual architectures of the Spo0 pathway and its striking tendency to gain and lose interactions may be due to the juxtaposition of three key properties: the maintenance of interaction specificity through molecular recognition; the ecological role of endosporulation; and the degeneracy of interaction space that permits the ongoing recruitment of kinases to recognize novel environmental signals.

Acknowledgments

Foremost, I would like to thank Dr. Dannie Durand for the support and mentorship she has given me as long as I've known her. From the first time that I stayed after her class to discuss the molecular genetics of the three bears that were suddenly stranded on different islands, I knew that Dannie would encourage me to dig deeper and approach scientific questions from every possible direction. These sentiments spurred me to join the Durand lab. Without Dannie's tutelage and feedback I would not have grown into the scientist and writer that I have become nor completed such a deep and interesting project.

My "co-"advisor, Dr. Luisa Hiller, was also incredibly supportive and especially helped me to understand the ins and outs of performing an experimental analysis. Further, Luisa's availability to troubleshoot, suggest next steps, and talk out just about any problem I had was essential in the completion of the experimental aspects of my research and to keep life in perspective (of course, all the coffee she gave me helped with this, too!).

I would like to thank the rest of my committee for their guidance and input, as well. Dr. David Hackney provided a keen look at the protein biochemistry involved in this project and always had insightful critiques during my committee meetings. Dr. Michael Laub's feedback and expertise on histidine-aspartate phosphotransfer interactions and their specificity helped me to define a feasible and compelling project.

In addition to my committee, this work would not have been possible without many other people. Members of the Durand and Hiller labs have provided with me a useful and actionable feedback that helped me to get through challenges in my computational and experimental work. Further, their intent listening and comprehensive critiques of my oral and poster presentations throughout the years have strengthened my communication skills. I would like to especially thank Han Lai, Maureen Stolzer, Collin McCormack, Deepa Sathaye, Rosie Alderson, Minli Xu, Patrick Ropp and Jayce Embry from the Durand lab and Anagha Kadam, Rory Eutsey, Evan Powell, Ben Janto, Rolando Cuevas, and Surya Dev Aggarwal from the Hiller lab (both at CMU and Allegheny Health Network). Each of these labmates have had immense and positive impacts on me throughout my academic career.

I could not have completed this PhD without the support of my friends and family. First, I'd like to thank my PhD Cohort who were in this with me every step of the way. I'd like to thank my friends for being the special people who were willing to listen. I could always count on Dan to open up new ways of looking at a problem or to encourage me to take a different approach. Karley was always willing to take a look at something I was writing, a presentation I was working on,

or to give me advice on whatever situation may have arisen. Nick has given me support going all the way back to high school and has been there for me whenever I need to talk something out or relieve some stress. Collin, Dan, Rory, Brendan, Teresa, and Anagha for providing consistent and wonderful company during lunch time nearly every day. Next, I'd like to thank my entire family, who do not get to see nearly enough, for the support that they provide and for coming to see my thesis defense. Finally, I'd like to thank my Fiancee, Chelsea Rose, for her unwavering support throughout my time obtaining my PhD. Chelsea is my biggest advocate and always encourages me to get enough sleep, figure out why I am stressing, and has done a heroic job as the cornerstone of my support network.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Signal transduction systems are essential for an organism to sense and respond to dynamic environments. In order to recognise new signals, these systems must also be able to adapt through the acquisition of new signal transduction pathways or the remodeling of existing pathways. Thus, adaptation to novel environmental challenges requires sufficient flexibility to allow the rapid remodeling of signal transduction systems. In bacteria, two-component signaling (TCS) systems, typically comprised of a histidine kinase (HK) and a cognate response regulator (RR), are a primary mechanism of environmental response (Fig 1.1A). Signal recognition by the N-terminal sensor region of the HK leads to the autophosphorylation of a conserved histidine residue in the so-called HisKA domain by the catalytic (HK_CA) domain. The signal is then transduced by phosphotransfer from the autophosphorylated HK to a conserved aspartate residue in the N-terminal receiver (REC) domain of the RR [Lewis et al., 1999]. Phosphorylation of the REC domain activates the C-terminal output domain of the RR, initiating a response to the recognized signal.

The modular encapsulation of sensing, transduction, and output functions into discrete regions within the two TCS proteins allows for specialized input and output modules that support recognition and response to a specific signal. Transduction between these domains is mediated by interaction at the interface of the HisKA and REC domains. This modularity allows for adaptation to novel environmental challenges through modification of the input module to recognize a new signal and modification of the output module to initiate a different set of output machinery [Galperin, 2005]. The variety of TCS proteins with distinct input and output modules and conserved interaction domains, 20-30 per organism on average, are a testament to this process [Szurmant and Hoch, 2010].

A set of non-contiguous, co-evolving residues at the interface of HK and RR proteins, six in

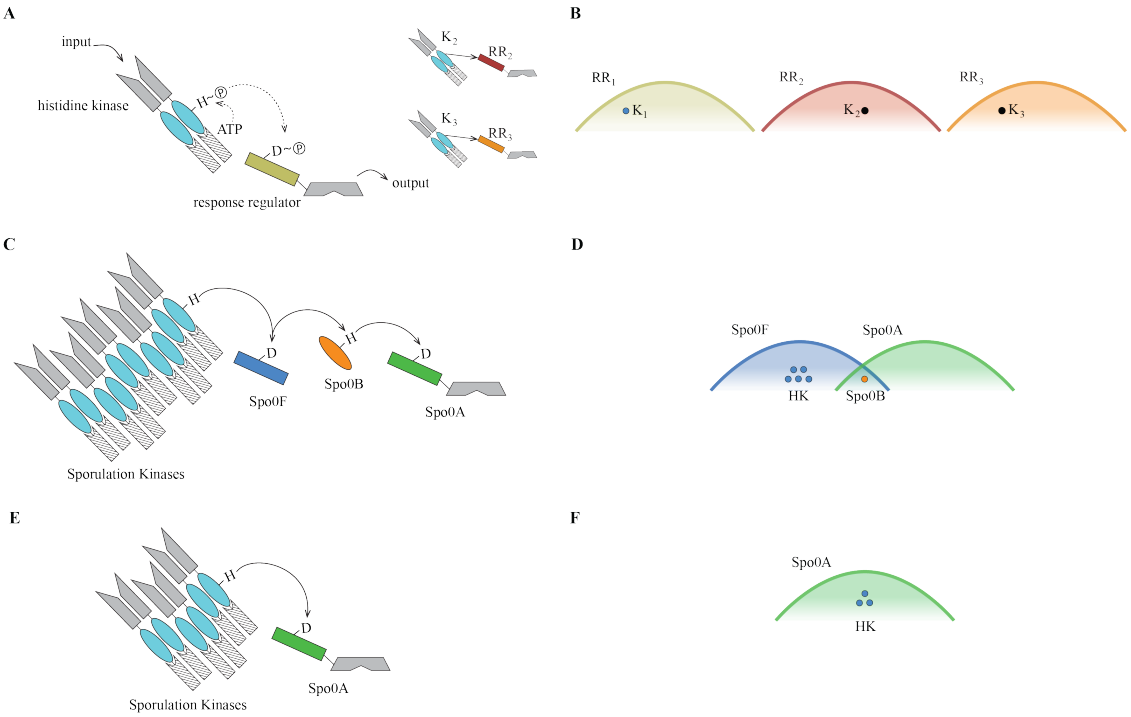


Figure 1.1: Histidine-aspartate phosphotransfer pathway architectures and qualitative representation of receiver interaction specificity. Protein interaction diagrams show simple and complex interaction architectures involving histidine-aspartate phosphotransfer domains, HisKA (teal oval), REC (blue and green rectangles), and Spo0B (orange oval). Grey domains are conserved aspects of these pathways that are not involved in interaction including sensor region of the histidine kinase (pointed rectangle), catalytic domain of the histidine kinase (rectangle with one curved side), and output domain of the response regulator (trapezoid with notch). Right side is a qualitative representation of interaction specificity where each colored semi-circle corresponds to the response regulator in the diagram in the left column. The horizontal axis represents the set of possible HK specificity signatures capable of phosphotransfer to a given RR (the RR specificity spectrum). The height of the spectrum indicates the strength of interaction. (A) Canonical TCS system architecture. Interaction between a histidine kinase and a response regulator (B) The interaction space of a set of three TCS system pathways showing the position of each HK within the specificity spectrum of its cognate RR. Although RRs with overlapping spectra could, in theory, be phosphorylated by the same HK, crosstalk would only occur if the genome encoded an HK with a specificity signature located in both spectra. Selection acts to separate these interactions to maintain signal fidelity and avoid crosstalk. Thus, each RR specificity spectrum must have a region that has no overlap with the spectrum of another response regulator, and its cognate histidine must occupy that non-overlapping region of the spectrum. (C) *B. subtilis* Spo0 phosphorelay. Multiple sporulation kinases (KinA-E) phosphorylate Spo0F. Subsequently, Spo0B transfers a phosphoryl group from Spo0F to Spo0A. (D) Phosphorelay interaction space. The phosphorelay interaction pattern requires that the spectra for the intermediate RR (Spo0F) and the terminal RR (Spo0A) overlap and that the phosphotransferase (Spo0B) be located in the overlap. Additionally, the specificity of the sporulation kinases must only be in the specificity spectrum of Spo0F or else crosstalk would occur to Spo0A. (E) *C. acetobutylicum* Spo0 direct phosphorylation architecture. Multiple sporulation kinases phosphorylate Spo0A. (F) Direct phosphorylation architecture specificity space. The multi-input direct phosphorylation specificity spectra only requires that all sporulation kinases be within the Spo0A specificity spectrum.

the HisKA domain and seven in the REC domain, ensure specific interaction within each cognate pair [Capra et al., 2010; Casino et al., 2009; Procaccini et al., 2011; Skerker et al., 2008]. These specificity residues are partially degenerate: multiple sets of kinase specificity residues permit phosphotransfer to the same receiver (and vice versa [Podgornaia and Laub, 2015]), such that each receiver has a spectrum of kinase specificity with which it can interact (Fig 1.1B). To prevent deleterious crosstalk between non-cognate proteins [Capra et al., 2012b], selection acts to separate the spectra of two-component signaling pathways encoded in the same genome. Acquisition of novel pathways (e.g. through duplication or horizontal gene transfer) can cause conflicts in interaction space. The degeneracy of these interactions allows for repositioning in interaction space to eliminate crosstalk via mutational trajectories involving compensatory mutations in the cognate pair. However, in the absence of a perturbation, pathways likely inhabit the same region of interaction space [Capra et al., 2012b].

Histidine-aspartate phosphotransfer also admits more complex signal transduction architectures. Examples include multiple-input architectures, *e.g.* [Kaczmarczyk et al., 2014], multiple-output architectures, *e.g.* [Wuichet and Zhulin, 2010], and so-called phosphorelays comprising a sequence of phosphotransfer, interactions *e.g.* [Biondi et al., 2006; Burbulys et al., 1991; Ueki and Inouye, 2002]. For example, the sporulation initiation (Spo0) pathway is a multi-input phosphorelay characterized extensively in *Bacillus subtilis* [Burbulys et al., 1991; Jiang et al., 2000; Trach and Hoch, 1993] and also observed in closely related species [Bick et al., 2009; Brunsing et al., 2005; Park et al., 2012]. In this architecture, multiple sensor kinases phosphorylate Spo0F, a response regulator lacking an output domain; subsequently, that phosphoryl group is transferred via Spo0B, an intermediate histidine phosphotransferase, to Spo0A, a second response regulator that serves as the master regulator of sporulation (Fig 1.1C).

The maintenance of signal fidelity in these more complex pathways entails additional constraints on the genetic determinants of specificity, because a single protein must support multiple interactions. The interaction requirements of the Spo0 phosphorelay necessitate precise molecular recognition to allow both Spo0F and Spo0A to interact with Spo0B, but only Spo0F to accept a phosphoryl group from sporulation kinases (Fig 1.1D). The balance of flexibility and constraint that shapes molecular recognition in these complex architectures is not well understood.

To explore this issue, I present here an analysis of the evolution of the Spo0 pathway. The Spo0 pathway controls entrance into a developmental program that produces stress-resistant, dormant endospores. The ability to produce endospores is a common feature of the Firmicutes phylum, observed in numerous species throughout two anciently related classes, the Bacilli and Clostridia, suggesting that this survival mechanism is ancient [de Hoon et al., 2010; Galperin, 2013]. These

two classes are predicted to have diverged 2.7 billion years ago, coinciding with the atmospheric rise of oxygen during the great oxidation event [Battistuzzi et al., 2004]. The ancestral Firmicute was likely an obligate anaerobe, a trait that has been preserved in the present-day Class Clostridia, whereas the Bacilli are typically facultative aerobes. Many taxonomic families include both sporogenous and asporogenous species, suggesting that the ability to sporulate is frequently lost [Galperin et al., 2012] through adaptation to a stable niche where sporulation is unnecessary for survival [Maughan et al., 2009].

Strikingly, a comparison of the Spo0 pathways in the type species of the two Firmicutes classes, *B. subtilis* [Burbulys et al., 1991] and *Clostridium acetobutylicum* [Steiner et al., 2011], reveals that the sporulation initiation pathway is conserved in its output [Al-Hinai et al., 2015], but differs in terms of the input(s) and the signal transduction architecture. Spo0A, the terminal component of the pathway in both species, initiates spore development upon phosphorylation [Hoch, 1971; Wörner et al., 2006] and is encoded by all known sporulators [Galperin et al., 2012]. Spo0A is a canonical response regulator protein in its domain composition, including a REC domain [Lewis et al., 1999] and a highly conserved, DNA-binding output domain, Spo0A_C [Lewis et al., 2000]. Unlike Spo0A, which is likely orthologous in these distantly related species, the upstream signal transduction architectures are different. In contrast to the *B. subtilis* multi-input phosphorelay Spo0 architecture, *C. acetobutylicum* and other closely related species possess a multi-input architecture in which Spo0A is directly phosphorylated by multiple kinases [Mearls and Lynd, 2014; Steiner et al., 2011; Underwood et al., 2009] (Fig 1.1E).

Considering that these two different signal transduction architectures both orchestrate the initiation of sporulation through the phosphorylation of an orthologous regulator, they likely arose from a common ancestral pathway. How then, did different signaling architectures evolve in present day species? The prevailing view is that the ancestral Spo0 pathway had a two-component direct phosphorylation architecture and the more complex phosphorelay observed in B. subtillis is a derived state [Durre, 2011; Stephenson and Hoch, 2002; Talukdar et al., 2015]. This hypothesis was inspired by the apparent lack of Spo0F and Spo0B orthologs in the first Clostridium genome sequenced [Stephenson and Hoch, 2002]. The simplicity of the direct phosphorylation architecture and the similarly anaerobic lifestyles of the ancestral Firmicutes and present-day Clostridia, taken together, fueled predictions that the original Spo0 pathway also functioned through direct phosphorylation [Durre, 2011]. It was further proposed that the phosphorelay likely arose in the Bacillar lineage, possibly as the result of duplication of a cognate HK-RR pair [Stephenson and Lewis, 2005], and that the additional points of control associated with a phosphorelay may have contributed to adaptation to rising oxygen levels in early Bacilli [Durre, 2014].

The goal of this thesis is the characterization of the forces that govern the evolution and remodeling of bacterial histidine-aspartate phosphotransfer-based pathways using the Sporulation Initiation (Spo0) pathway as a case study. Regardless of the status of the ancestral pathway, some combination of gains and losses of interaction must have occurred to produce the distinct pathway architectures observed in present day species. I used a combination of computational and experimental techniques, based on a comparative genomics approach, leveraging the dramatically increased number of sequenced Firmicutes genomes available, to investigate these remodeling events. I provide evidence for the most comprehensive set of Spo0 pathway homologs, to date. Based on the distribution of these homologs, I generate a novel hypothesis for the evolution of the Spo0 pathway architectures. This evolutionary history implies a series of remodeling events that occured repeatedly in independent lineages, resulting in transition from an ancestral phosphorelay architecture to present-day direct phosphorylation architectures.

First, in Chapter 2, I review the histidine-aspartate phosphotransfer literature including the domains involved, common and complex architectures, specificity via molecular recognition and spatial tethering, and the evolutionary effects of signal transduction via molecular recognition in two-component signaling systems (Section 2.1). Next, I present a review of the knowledge regarding the Firmicutes phylum and endosporulation (Section 2.2). Finally, I review previous studies of the Spo0 pathway that have shaped our understanding of the two pathway architectures that are predicted to share a common ancestral pathway architecture(Section 2.3).

In Chapter 3, I investigate the present-day distribution of Spo0 architectures. The present-day Spo0 architectures was examined in a set of 84 Firmicutes genomes (Section 3.1). Although Spo0A was readily identifiable by a conserved output domain, Spo0F, Spo0B, and the candidate sporulation kinases were unable to be identified using typical homology identification approaches due to sequence similarity methods identifying too much. Instead, I defined specific criteria for each component, leveraging genome context and domain content, to identify a set of putative homologs throughout the genomes of the representative set. Based on similarity of specificity residues to experimentally verified Bacillar phosphorelays and a Clostridial phosphorelay experimentally verified in this work, the identified homologs are likely to function in Spo0 architectures. Specifically, I predict that if Spo0F and Spo0B are present, the Spo0 architecture of that genome has a phosphorelay architecture. In the absence of one or both of these intermediate phosphorelay proteins, the likely architecture is a direct phosphorylation of Spo0A.

Having presented evidence for the predicted architecture of Spo0 pathways in the representative set, I examine the phylogenetic distribution of these architectures (Section 3.4). This reveals a patchy distribution of both the phosphorelay and direct phosphorylation architectures. To assure

that this result is not due to phylogenetic error in the species tree that I constructed, I also present analyses of phylogenetic distribution in two other sets of Firmicutes reported in the literature, including one that is much larger (Section 3.4.1). Similar patchiness is observed in each of these alternate trees. This patchy distribution suggests that there have been multiple transitions from the ancestral architecture. The presence of Clostridial phosphorelays and the patchy distribution of Spo0 architectures in both classes calls for the reconsideration of the evolutionary history of the Spo0 pathway in the Firmicutes.

In Chapter 4, I compare specificity residues and interaction specificity of Bacillar Phosphorelays, Clostridial phosphorelays, and direct phosphorylation architectures to generate evidence regarding the evolutionary history of the Spo0 pathway. I provide evidence that specificity in phosphorelays is similarly encoded in both classes (Section 4.1) and that phosphorelays from Bacillar and Clostridial lineages are functionally interchangable (Section 4.2). These observervations suggest a single genesis of the Spo0 phosphorelay, from which all other Spo0 phosphorelays are descended. Based on the patchy distribution of phosphorelay architectures, I reason that this genesis must have occured prior to the divergence of Bacilli and Clostridia (in other words, in the Firmicutes ancestor), as all other scenarios containing a single generation of the Spo0 phosphorelay would require an unlikely series of horizontal gene transfer events.

If the ancestral Spo0 pathway was a phosphorelay, then the direct phosphorylation architectures, observed patchily throughout the phylum, must be the result of multiple, independent instances of remodeling, wherein an interaction was gained between a sensor kinase and Spo0A (Section 4.3.1. Comparison of heterologous interactions between these species provides insight into the possible mechanisms of remodeling (Section 4.3.2).

Finally, in Chapter 5, I discuss how flexibility and constraint have shaped the predicted evolutionary history of the Spo0 architectures. First, I present evidence that the repeated transitions were mediated by changes in kinase specificity while Spo0A specificity has remained relatively stable (Section 5.1.1). Considering the differences between interaction profiles that I observe, I discuss a series of possible mechanisms that result in the acquisition of a sensor kinase that specifically phosphorylate Spo0A (Section 5.1.2). This repeated remodeling suggests that phosphorelays which signal through molecular recognition are particularly susceptible to remodeling, compared to spatially tethered phosphorelays (Section 5.2). I hypothesize that this apparent susceptibility also permits the acquisition of new signal recognition capabilities, allowing sporulation in response to novel conditions. Experiments and future genomic studies motivated by this work are also discussed.

In summary, my results support a scenario in which the ancestral Spo0 pathway in the Firmi-

cutes ancestor was a phosphorelay, a contrasting hypothesis to the prevailing theory. The phylogenetic distribution of Spo0 architectures is patchy, consistent with several independent transitions from phosphorelay to direct phosphorylation architecture. My results further suggest that these transitions were mediated via changes in sensor kinases, while Spo0A specificity is conserved across the Firmicutes phylum. My findings provide a framework for reasoning about the forces that act to maintain the signaling fidelity in complex signal transduction pathways with multiple interactions.

Chapter 2

Background: Two-Component Systems and Spo0 Pathway History

2.1. Bacterial Signal Transduction Systems

In bacteria, the most common type of signal transduction system is the two-component system, consisting of a histidine kinase and a response regulator. Briefly, the histidine kinase reacts to a stimulus, autophosphorylates, and transfers that phosphoryl group to the response regulator. Upon receiving the phosphoryl group, the response regulator is activated to change cellular behavior, typically as a transcription factor. Examples of interacting pairs are typically encoded adjacently in so-called cognate pairs. Section 2.1.1 describes in detail the domain composition of each of these proteins and modularity in these pathways. This modularity further begets plasticity, which may be necessary for the evolution of these pathways (section 2.1.3). Typical bacterial cells encode 20-30 TCS systems, and therefore requires that interaction between cognate pairs is specific. The discovery of this specificity and its implications are discussed in section 2.1.2.

2.1.1. Two Component Signaling System Domain Content and Modularity

In bacteria, the most common type of signal transduction system is a two-component signaling (TCS) system, in which a Histidine Kinase (HK) autophosphorylates in response to a specific signal and then transfers the phosphoryl to a Response Regulator (RR). Each HK has a sensor region, which is N terminal and typically extracellular. The C terminal portion of an HK consists

of two domains: a HisKA domain and a catalytic domain (HK_CA). Upon detection of a signal by the N terminal domains, the HK_CA domain catalyzes the hydrolysis of an ATP to obtain the phosphate that will be transferred through the pathway. The HisKA domain contains the histidine that receives the phosphate during autophosphorylation. The RR is composed of two domains: the response receiver (REC) domain and an output domain. The receiver domain of the RR interacts specifically with the phosphorylated HisKA domain. Once the RR has been phosphorylated it will undergo a conformational change, allowing it to initiate a change in cellular processes in response to the signal detected by the HK.

A detailed explanation of the domains and interactions observed for these two proteins follows.

2.1.1.1. Histidine Kinase

The histidine kinase is composed of an N-terminal sensor region and C-terminal interaction and catalytic domains. The N-terminal sensor region is often separated from the C-terminal domains by a transmembrane region resulting in the sensor region being displayed on the outer membrane of a cell, allowing detection of the outside environment or cell deformation. However, cytoplasmic histidine kinases that react to changes in the environment within the cell have been observed .

Upon detection of a stimulus, a series of conformational changes occur that initiate signal transduction. Only limited studies have been performed to determine the mechanisms that allow conformational change in the N-terminal sensor region upon recognition of an environmental trigger. Based on these studies, sensor regions are specific for a signal due to the specific biochemistry of amino acids at the protein surface. The sensor regions undergo conformational changes under a specific condition that, in turn, cause conformational change in the C-terminal domains responsible for signal transduction. In some cases, the sensor may be formed by the coordination of histidine kinase dimers [Sánchez-Sutil et al., 2016]. For example, a specific, structural mechanism for signal recognition has been discovered for a histidine kinase responsible for the recognition of copper ions where a copper ion is coordinated in the sensor region which causes long range conformation changes [Sánchez-Sutil et al., 2016]. Although a specific signal recognition method may be challenging to identify, the stimuli associated with sensor domains can be studied using genetic experiments where a pathway is mutationally affected and growth or other responses are monitored in the condition of interest. Examples include the EnvZ-OmpR pathway that responds to oxygen [Shimada et al., 2015], the *E. coli* QseC pathway that senses epinephrine and norepinephrine [Clarke et al., 2006], and the *Myxococcus xanthus* CorS pathway that senses copper [Sánchez-Sutil et al., 2016]. Finally, the stimuli of a histidine kinase can be inferred by homology or domain conservation with the sensor region.

After the signal is detected, the catalytic domain (HK_CA) of the histidine kinase hydrolyzes a bound ATP and autophosphorylates a conserved histidine residue in the interaction domain, HisKA. Many histidine kinases have been found to be homo-dimeric. This dimerization is mediated by residues within the HisKA domain. This dimerization may be functionally relevant as many HKs have been found to "auto"-phosphorylate the HisKA domain in *trans* [Casino et al., 2009]. Histidine kinases have been classified into distinct sub-groups based on the phylogenetic signals of their C-terminal domains in diverse bacterial and fungal species [Catlett et al., 2003]. This subclassification is limited mainly to HKs involved in general TCS and those similar to chemotaxis kinase, CheA. Another subclassification focuses on the catalytic domain using hidden Markov models based on groups of similar kinases [Wuichet and Zhulin, 2010]. While these subclassifications of histidine kinases and their catalytic domains may be useful to distinguish evolutionarily distinct groups of kinases from each other, the basic function of the interaction and catalytic domains has been preserved.

2.1.1.2. Response Regulator

Once the kinase is phosphorylated, the response receiver transiently binds to the HisKA domain and the phosphate is transferred to a conserved aspartate residue in its own interaction domain, REC. Initial mutagenic and structural efforts examining the sensor histidine kinases and response receivers, individually, revealed that the interaction domains of each protein are structurally conserved and each has a specific histidine or aspartate residue where the signaling phosphoryl group is covalently bound [Mizuno, 1998]. The response receiver is typically composed of a REC domain and some output domain. Output domains are typically DNA-binding domains that allow the recruitment of other transcription machinery including the broad subtypes HTH_18 and Trans_reg_C. Other types of output domains that have enzymatic activity have been observed [Galperin, 2006].

Prior to phosphorylation, the REC domain sterically inhibits activity of the output domain. Once the receiver domain is phosphorylated, a conformational change occurs that allows the output domain to function. This rearrangement of the protein revolves around a conserved aromatic residue in the REC domain, commonly referred to as an "aromatic switch". (Note, some response regulators have two aromatic residues in this region of the response regulator.) Typically, in response regulators that have not been phosphorylated the aromatic switch residues will be pointing away from the core of the protein and the most energetically favorable conformations will include hydrophobic hiding of these residues by the C-terminal output domain. These conformations prevent the C-terminal output domain from performing its function. However, upon phosphorylation of the REC domain, conformational changes occur that activate the output domain [Lewis et al.,

2002]. Specifically, phosphorylation of the conserved aspartic acid causes the displacement of a conserved threonine or serine that moves towards the phosphoryl group. This movement has been termed the "threonine flip". Subsequently, the aromatic residue forming the aromatic switch, typically a phenylalanine or tyrosine residue, moves into the space created by this displacement (termed the "tyrosine tuck"). The "tucking" motion hides the exposed hydrophobicity of the aromatic residue, allowing conformations where the C-terminal output domain is not bound to the REC domain. This results in the activation of the output function. Some response regulators have also been demonstrated to dimerize at the site of the tucked tyrosine [Lewis et al., 2002].

2.1.2. Specificity in Two Component Systems

A bacterial organism can encode between up to two hundred [Alm et al., 2006] individual two component systems and typically encode between 20 and 30. Each of these pathways must remain specific for its intended target or cross-talk will develop and could reduce the efficacy of signal transduction or initiate an inappropriate response [Laub and Goulian, 2007a]. It is essential to understand how these proteins interact in order to understand how specificity is maintained. These interactions are likely transient as many sensor kinases are transmembrane while receivers that regulate genes are located on the bacterial chromosome, away from the surface.

Given the lack of variation in the shape of the domain, specificity must be achieved via the amino acids presented at the interfacial region of the interaction [Szurmant and Hoch, 2010]. The first structural analysis of these domains in complex revealed that not only do the amino acids surrounding the conserved phosphorylated residues interact, another set of residues away from the phosphorylation site also come in close contact [Zapf et al., 2000]. The non-phosphorylation site residue interactions are mainly found within the α1 of both RR and HK as well as α2 of HK which form the majority of the interfacial surface. Non-phosphorylation site residue interactions were observed in similar positions in more recent structures of HK and RR in complex [Casino et al., 2009; Yamada et al., 2009], despite different residues observed at those positions.

To elucidate how specificity is conferred, Skerker et al. [2005] sought sites that co-vary across pairs of cognate two component system proteins to find interprotein, co-evolving residues. This analysis revealed that a series of non-contiguous residues localized to the interaction domains of histidine kinases and response regulators (specifically, six in the HisKA domain and seven in the RR domains) have the highest interprotein covariance [Skerker et al., 2005]. Subsequent experiments showed that by swapping kinase interaction domains, the targeted response regulators were also swapped and the same effect was conferred by mutating as few as three of the six histidine kinase specificity residues [Skerker et al., 2008]. Such rational rewiring also had the same effect

for response regulators [Capra et al., 2010]. These experimental studies also show that it is difficult to predict the effects of individual specificity residues. Modifying only one specificity residue of a kinase, in some cases, resulted in gain of non-specific interactions to one or more partners, changed the strength of interaction observed, or did not change specificity at all.

To predict which histidine kinases would be specific for a certain response regulator, computational techniques were employed. One method used direct-coupling analysis [Procaccini et al., 2011] which determined 70 of the most coupled interdomain columns in an MSA of almost 9,000 concatenated cognate pairs. Cognate pairs were then scored by the set of paired residues of the columns discovered by DCA. Each set of paired residues were scored using a method based on the frequency of observing two residues paired and their biochemical properties. A favorable interaction would increase the score for that column. One such interaction might be the oppositely charged aspartate and arginine residues, which are observed to be frequently paired and would readily interact to "hide" their respective charge. Although this resulted in a fairly robust estimation of interaction, it was not extremely accurate for some of the examples they scored and the cut-off between interacting and not interacting is unclear.

2.1.2.1. Interaction Space

In two-component systems, the genetic determinants of interaction are partially degenerate; multiple sets of kinase specificity residue signatures permit phosphotransfer to the same receiver (and vice versa [Podgornaia and Laub, 2015]). The specificity spectrum of each RR is therefore defined by the specificity residues of the set of histidine kinases that are able to phosphorylate it.

To prevent deleterious crosstalk between non-cognate proteins [Capra et al., 2012b], selection acts to separate the spectra of two-component signaling pathways encoded in the same genome. For the set of TCS proteins encoded in a single genome, these constraints require that the space of all possible interactions is partitioned, i.e. each HK must interact with only one RR. This partition requires that each specificity spectrum must have a region with no overlap and its cognate HK

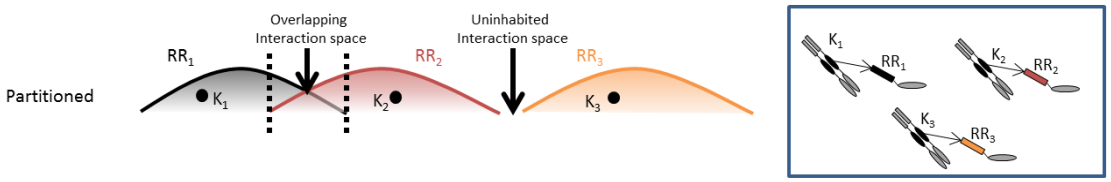


Figure 2.1: Example specificity spectra. The interaction space of the three partitioned RRs with their cognate histidine kinases is depicted. The set of histidine kinases that do not have phosphotransfer to any re in a given cell are depicted as blank space between spectra (arrow).

must occupy that part of the spectrum. The degeneracy of specificity allows for some plasticity; mutations are tolerated if interaction is maintained without the generation of crosstalk. Acquisition of novel pathways (e.g. through duplication or horizontal gene transfer) can cause conflicts in interaction space that require co-evolutionary shifts of specificity spectra. However, in the absence of a perturbation, pathways likely inhabit the same region of interaction space [Capra et al., 2012b].

2.1.3. Expansion of TCS through remodeling

The diverse repertoire of TCS proteins that respectively sense and respond to diverse stimuli are linked by the conserved signal transduction mechanism of phosphotransfer. The variety of these so-called cognate pairs are a testament to the rapid expansion of TCS pathways [Szurmant and Hoch, 2010]. Several processes, including domain replacement (or shuffling) and duplication/divergence events, contribute to the expansion of TCS systems [Alm et al., 2006].

Duplication and divergence allows the generation of new TCS pathways from existing ones [Alm et al., 2006]. TCS proteins are typically encoded adjacently, leading to an increased likelihood of being duplicated together. Following duplication, the resulting two identical pathways must accrue mutations such that fitness is provided in the cell or else they may be lost. The pairing of histidine kinase and response regulator results in additional constraints. First, the two identical pathways must accrue sufficient mutations within their interaction domains to insulate from the other pathway (and also must avoid other existing pathways, see 2.1.2) such that signaling fidelity is maintained. Next, the sensor and output domains may also need to accrue mutations such that a novel signal and novel response are generated. In addition to mutations within the output domain and in order to provide a new response, this evolutionary process likely plays out in mutations throughout the genome in the form of mutations in non-coding DNA to create transcription factor binding sites (in the common case that the output domain is a transcription factor) and the proteins that will be regulated by this new output domain.

The modularity afforded by encapsulation of function into discrete domains allows specialization to sense and respond to new signals through domain shuffling and replacement. As the N-terminal sensing region of the HK and the output domain of the RR have been demonstrated not to affect the ability of the interaction domains to perform phosphotransfer [Procaccini et al., 2011; Skerker et al., 2005], they could be replaced by their counterparts from different TCS proteins. Such replacements are possible through recombination and other domain shuffling events [Long and Thornton, 2001; Long et al., 2003]. The result of replacing the sensing or output modules with another similar domain could pair a novel response to a stimulus. A similar result could be obtained by replacing one of the interaction domains in either the HK or RR, resulting in a new

pairing of sensing and output modules. The new pairing could result in a beneficial response that could provide a fitness advantage, leading to selection and proliferation.

2.1.3.1. Specificity following Duplication of the Ntr pathway

Study of the duplication/divergence of the genetic determinants of interaction in the NtrBY and NtrCX pathways across the alpha-, beta-, and gammaproteobacteria lineages revealed that, in the alphaproteobacteria lineage, the specificity of the PhoRB pathway was impacted [Capra et al., 2012b]. This hypothesis is supported by lineage-specific differences in the specificity residues of both NtrB and PhoR that translate into differences in specificity. When heterologously combined, the gammaproteobacterial PhoR was able to target the alphaprotebacterial NtrX while the autologous PhoR was unable to interact with it. Given that NtrX is not encoded by gammaproteobacteria, this suggests that the duplication and subsequent divergence of NtrX in the alphaproteobacteria required changes in specificity residues for PhoR.

This study demonstrates that the constraints of a partitioned specificity space both restrict and drive changes in signal transduction over the course of evolution. Perhaps on a larger or more immediate scale, these constraints additionally affect the specificity of duplicated pathways. In particular, the divergence within interaction space is the result of selection to insulate the two new pathways from each other to maintain signal fidelity. These changes in specificity will manifest as differences between the specificity residues of the paralogous proteins [Capra et al., 2012b]. These changes that prevent such interactions can occur in either the HisKA or REC domains. Changes to additional unrelated pathways may occur if crosstalk develops to an unrelated pathway during the mutational trajectory to insulate the duplicated pair from each other. This study highlights the role that flexibility of interaction and the constraints of the partitioned interaction space play in the evolution of TCS proteins.

2.1.4. Complex Histidine-Aspartate Phosphotransfer Architectures

For the canonical two-component signaling system, the architecture is simply direct phosphorylation by phosphotransfer between the histidine kinase and response regulator interaction domains.

However, more complicated signal transduction architectures have also been observed. These include multiple-input architectures, where more than one kinase is specific for the same RR, (e.g. the general stress response of Alphaproteobacter [Kaczmarczyk et al., 2014]). The level of signal integration among the signals transduced by the kinases in these systems is unclear. A single HK can also be specific for multiple RRs in a multiple-output architecture (e.g. the signal transduction

proteins that control chemotaxis in many organisms [Wuichet and Zhulin, 2010] or the response to adrenergic hormones in *E. coli* [Parker et al., 2017]).

Multiple phosphotransfer interactions are observed in phosphorelay architectures. The interaction domains involved in a phosphorelay can be encoded within the same protein (e.g. hybrid HK observed in stalk biogenesis of Caulobacter [Biondi et al., 2006] or hybrid RR observed in flagellar gene expression in Geobacter [Ueki and Inouye, 2002]) or distributed across multiple proteins (e.g. RedCDEF in *Myxococcus xanthus* [Jagadeesan et al., 2009]).

The major difference between these two types of phosphorelay architectures is the way signal fidelity is managed. Interaction in distributed pathways must still rely on molecular recognition, restricting certain mutational trajectories where interaction is lost or cross-talk is gained (as discussed in 2.1.2.1). The maintenance of signal fidelity in these more complex pathways entails additional constraints on the genetic determinants of specificity because a single protein must support multiple interactions. The interaction requirements of the distributed-domain phosphorelay, for example, necessitate precise molecular recognition to allow both the intermediate and terminal response regulator to interact with histidine phosphotransferase. However, additional constraints are made to the specificity of the initiating histidine kinase, which, to maintain signal fidelity, must only interact with the intermediate response regulator and not the terminal response regulator.

However, for the co-located phosphorelay interactions, specificity is controlled by spatial tethering, relaxing constraints on specificity residues [Capra et al., 2012a]. The specificity residues of co-located interaction domains lack a strong mutual information signal suggesting that they are under weaker selection than domains where specificity is maintained by molecular recognition. Further, when the REC domain is removed from a hybrid histidine kinase, it displays phosphotransfer to several autologous response regulators, *in vitro* (and sometimes does not even interact with its co-located domain!). This suggests that the co-location of the REC domain insulates the HisKA domain from interaction with other RRs as, based on the constraints placed on TCS by the partitioning of specificity space, these promiscuous and strong interactions would likely be deleterious. This insulation is likely incurred by the effective concentration of the interaction REC domain (or, in other words, the co-located REC domain is phosphorylated before interaction with a different REC domain, due to its proximity).

Considering the effect of spatial tethering in the broader context of the other TCS proteins in a given genome, the evolution and specificity of the co-located interaction is not governed by selection which acts to create a partitioned TCS system to maintain signal fidelity and avoid crosstalk. Therefore, the HisKA specificity may be specific for one or more of the response regulators encoded by that genome and may not be specific for its co-located REC domain. Loss of the REC

domain from a hybrid histidine kinase may result in a remodeling event whereby the newly minted HK is wired up to a separate response regulator that it is able to phosphorylate.

2.2. Firmicutes Phylum and Endosporulation

The Firmicutes (which loosely translates to s̈trong skinïn Latin) phylum is a taxonomic delineation of mainly gram positive bacteria with low %G+C [Vos et al., 2009]. Most members of this phylum have either cocci or rod-shaped cells. One of the key characteristic phenotypes of this phylum is the production of endospores, a cellular morphotype that is dormant and protected from the environment. Endosporulation is observed widely and scattered throughout the phylum. The major divisions in this phylum, the classes Bacilli and Clostridia, are contrasted by their preference for anoxic versus oxygenated environments. The Clostridia are obligate anaerobes, while the Bacilli are facultative aerobes. Due to this major ecological difference as well as phylogenetic inference [Battistuzzi et al., 2004], the Bacilli are thought to have diverged from the Clostridia around the time of the Great Oxygenation Event, 2.3 to 2.7 billion years ago. The Bacilli are likely descendants of a population of ancestral Firmicutes exposed to Earth's atmosphere that evolved the ability to survive in increasingly oxygenated conditions. A discussion of the phylogenetic organization within this phylum follows (Section 2.2.3).

2.2.1. Ecology

The Firmicutes are ecologically diverse. Firmicutes species have been found in a large variety of environments including aquatic, mammalian, soil, and food [Mandic-Mulec et al., 2015; Vos et al., 2009]. Members of both classes also include extremophilic species including from those tolerant of extreme heat (thermophilic) [Juturu and Wu, 2017; Kozianowski et al., 1997], cold (psychrophilic) [Larkin and Stokes, 1967], and salinity (halophilic) [Mavromatis et al., 2009]. Many species inhabit specific niches and require certain substrates for their continued survival and growth. For example, *Desulfotomaculum acetoxidans* was demonstrated to perform redox reactions necessary to survival using acetate or butyrate as a substrate (and as a by-product produces hydro-sulfuric acid!) [Widdel and Pfennig, 1977].

Several species within both classes are also medically relevant as many of the earliest discovered and studied species within the Firmicutes are pathogenic. Further, transmission is commonly linked to the production of antibiotic and disinfectant resistant endospores. Notably, this phylum includes the leading cause of hospital-associated diarrhoea, *Clostridiodes difficile*, and many other species that are associated with human or livestock disease [Ontiveros Corpus et al., 2008].

Many infections by these bacteria are typified by their production of toxins and neurotoxins such as those that result in paralysis (*Clostridium botulinum* [Maignel-Ludop et al., 2017]), muscle spasms (*C. tetanus* [Ahnert-Hilger and Bigalke, 1995]), or cause cell death (*Bacillus anthracis* group [Mourez et al., 2002]). In general, bacterial pathogenesis is a substantial medical issue; the medical challenges when faced with Firmicutes pathogens are enhanced by toxinogenesis and sporulation. Sporulation in particular can allow the pathogen to persist despite the arsenal of medical treatments available.

The Firmicutes phylum is taxonomically subdivided into two major classes, the Bacilli and the Clostridia. Although these classes are each diverse, they are typically characterized by aerotolerance; most Clostridia are obligate anaerobes and most Bacilli are facultative aerobes. This distinction, along with phylogenetic evidence [Battistuzzi et al., 2004], have led to the hypothesis that the Bacilli and Clostridia diverged around the time of the Great Oxidation Event (GOE), when the atmospheric oxygen levels began to rise due to the advent of photosynthetic cyanobacteria in the oceans.

2.2.2. Production and lifecycle of Endospores

Endosporulation is a survival response to adverse environmental conditions observed broadly scattered throughout the Firmicutes phylum [Galperin et al., 2012]. The specific conditions wherein sporulation is activated vary between species and may be lineage specific [Auchtung and Grossman, 2008; Baril et al., 2012]. Heat and starvation are conditions typically associated with the production of spores. As such, many of the assays that are used to study the production of spores use media with reduced nutrients or subject the organisms to extreme heat [Burns and Minton, 2011]. However, other lineage-specific sporulation signals or conditions have been discovered such as quorum sensing in *Bacillus subtilis* [Auchtung and Grossman, 2008] and the presence of acetate in *Desulfotomaculum acetoxidans* [Widdel and Pfennig, 1977].

Although this survival response ensures the continuation of the lineage of an organism, it is a metabolically expensive [de Hoon et al., 2010] and intricate process involving up to 150 different genes [Galperin et al., 2012; Molle et al., 2003]. If sporulation is interrupted by irregular expression of a sporulation gene or not enough energy is available to complete the process, the resulting spore may not be protected from the environment and could result in loss of both the mother and daughter cell. To assure that spores are only produced in the correct conditions, tight control of the initiation of sporulation is advantageous [Higgins and Dworkin, 2012].

The production of endospores is split into five "stages" and the genes controlling it are labeled by the stage in which they are involved. Initiation of sporulation is also called sporulation stage 0

and all involved proteins carry the prefix Spo0.

The initiation of sporulation is controlled by a histidine kinase-based signal transduction system in all species in which it has been studied [Burbulys et al., 1991; Steiner et al., 2011]. The result of activation of this pathway is the phosphorylation of Spo0A, a response regulator that acts as the master regulator of sporulation, controlling many of the genes in the initial stages of spore production [Asayama et al., 1995].

The general scheme of the sporulation process begins with standard mitosis, but diverges at the conclusion of anaphase II. At this point, rather than splitting into two cells, the daughter cell is engulfed by the mother cell, where it is coated with protecting proteins and lipids [Tocheva et al., 2011]. In many species, endosporulation uses up the remaining energy stores of the mother cell, which typically undergoes apoptosis, releasing the spore into the environment. However, there are some species which are known to produce multiple spores before apoptosing, such as *B. polyspora* [Flint et al., 2005]. Once the spore is created it will remain dormant until it is reactivated in response to environmental triggers [Paredes-Sabja et al., 2011].

The result of the endosporulation process is a dormant cell surrounded by a double membrane, protectively coated with a mixture of proteins and lipids [Tocheva et al., 2011]. Typical spores are protected from the gamut of environmental stressors including heat, radiation, antibiotics, and toxins, by their thick, decorated membranes [Setlow, 2014]. Through these resistances, spores are able to persist until conditions improve, at which point they will germinate, beginning the bacterial lifecycle anew. Germination conditions are likely to be niche-specific [Paredes-Sabja et al., 2011]; anaerobic spore-formers that colonize the human gut may germinate in response to tachoaic acid [Francis et al., 2013] .

2.2.2.1. Relationship between Negativicutes and Sporulation

Unlike most other Firmicutes, the so-called Negativicutes have a double cell-membrane and are Gram-stain negative. Recent work in *Acetonema longum* by [Tocheva et al., 2011, 2013, 2016] has provided evidence that the diderm characteristic may be related to endosporulation. Fully mature spores of the Negativicutes, *Bacillus*, and *Clostridium*, all share the same morphology: one cell membrane derived from the forespore and one derived by engulfment by the mother cell. When spores germinate, those of Bacilli and Clostridia lose their outer membrane, while in *A. longum* the outer membrane persists and is further elaborated.

2.2.2.2. Challenges in identifying sporogenous species

The spore formation phenotype is observed scattered throughout the Firmicutes phylum, implying that sporulation has been independently lost in many lineages. It has been suggested that adaptation to a stable environment will lead to the eventual loss of the ability to sporulate [Maughan et al., 2009]. Specific adaptations, where possible, to the dynamics of a specific environment may be more cost-effective than the expensive and intricate sporulation process. Consistent with this hypothesis, there are two large clades within the Firmicutes phylum, the Lactobacillaceae and the Veillonellaceae, that have adapted to specific, stable environments and spore-formation is not observed. Both of these families are commonly associated with commensal or pathogenic inhabitation of animal niches [Ludwig et al., 2015].

For the purposes of this thesis, sporulation status of a species has implications on (what can be learned) from the Spo0 components. If a species is asporogenous, its Spo0 pathway may not be under selective pressure and could be in the process of being lost or lost altogether. The formation of spores can be measured in a variety of ways and, for sporogenous species, is typically reported as part of the initial description of a species or alongside the characterization of a species during whole genome sequencing. However, asporogeneity is challenging to predict. Several stress conditions have been used to test for the presence of spores including growth in reduced-nutrient media (starvation), heat-shock, or exposure to oxygen [Burns and Minton, 2011]. After induction of sporulation, spore-formation has been measured by observation of spore-forms using microscopes and viability following stress condition. The latter method is based on the implicit assumption that only spores would survive the stress condition. Measurement of spore-formation (or in some cases, the prevalence of spore-formation) is also used in the study of the genetic basis of sporulation (e.g. characterization of mutants and knockouts of sporulation kinases).

While determining that a species is able to form spores may be trivial using standard methods, assessment of inability to form spores is challenging. Typically, if no spores or viability following stress conditions is observed, the species will be reported to be asporogenous. However, as spore-production has been observed in a variety of specific conditions and may not be initiated by the conditions tested, it is possible that some reports of asporogenous species are false-negatives. Further, the methods used to detect spore-formation could vary based implementation. Differences in spore detection and counting technique, for example, has lead to vast variability in reporting the frequency of *Clostridiodes difficile* spores [Burns and Minton, 2011].

Another strategy to determine sporogeneity leverages sporulation genotypes. Specifically, a number of studies have sought the minimal set of sporulation genes that genes are necessary and sufficient for the production of endospores [Abecasis et al., 2013; Browne et al., 2016; Galperin

et al., 2012]. For example, Galperin et al. [2012] examines the presence of four essential genes to provide evidence for sporogeneity. If the genome is found not encode Spo0A, evidence suggests that it will not to form spores. Consistent with this hypothesis, in studies where Spo0A is knocked out or mutated, an asporogenous phenotype is observed [Obana et al., 2017; Trach and Hoch, 1993]. However, if Spo0A is encoded but there is no evidence of spore-formation, the presence and absence of downstream sporulation genes may not be sufficient to predict the (lack of) ability to form spores. This may be further complicated by variations in the proteins involved in the downstream mechanisms of spore-production, given the observed differences between evolutionarily distant species [Al-Hinai et al., 2015; Dalla Vecchia et al., 2014; de Hoon et al., 2010].

Finally, a recent study on the prevalence of spore-formers in the human anaerobic microbiome suggests that the number of sporogenous species within the Ruminoccaceae, Lachnospiraceae, and Erysipelotrichaceae has been underestimated [Browne et al., 2016]. This study grew microbes from human faecal samples to determine which species compose the intestinal microbiome and, specifically, what the prevalence of spore formation in this microbiome is. To assess sporulation capability, growth was compared with and without exposure to ethanol and oxygen. Single-founder colonies were generated from combined samples after exposure to the combined stress condition. The viability of these colonies is likely the result spore-formation. To determine the identity of these species that were able to survive the stress conditions, whole genome sequencing was performed on the single-founder colonies. They found that several of the species they isolated had greater than 95% similarity to species that had not been reported to form spores, but clearly survived the stress condition.

There are several possibilities for these conflicting results. First, the conditions tested in this study differ from those tested in prior studies. Next, some of the species identified in this study may be able to withstand the ethanol stress condition via other mechanisms [Bravo-Ferrada et al., 2015]. Finally, there could be strain differences between a species identified as a non-spore-former in the lab versus that retrieved from the "wild" such that the lab strain is a *bonafide* non-spore-former while the environmental strain is able to form spores. At 95% sequence similarity, for example, the isolate may even be a distinct species from the sequenced species to which it shares similarity. While these results are intriguing, we look forward to more detailed characterization of the strains identified including microscopy, spore-formation in isolation, and further phylogenetic classification before species definitions are altered.

2.2.3. Phylogenetic Context and Taxonomy

The Firmicutes phylum is thought to have diverged from the high-GC actino- and cyano-bacteria approximately 3 billion years ago [Battistuzzi et al., 2004]. Within the Firmicutes phylum, the currently accepted hypothesis is that the classes Clostridia and Bacilli diverged approximately 2.3-2.7 billion years ago, coinciding with the atmospheric rise of oxygen [Battistuzzi et al., 2004].

A definitive species tree for the Firmicutes phylum has been elusive due to the nature of the time-scale, in addition to the likely occurrence of horizontal gene transfer that may cloud phylogenetic signal. As more Firmicutes genomes have become available, efforts have been made to generate Firmicutes species trees that are more accurate and comprehensive [Antunes et al., 2016; Ludwig et al., 2015; Yutin and Galperin, 2013].

Early phylogenetic efforts included trees built to determine the relationships within the two major classes of the Firmicutes. These notably include a phylogeny by Collins et al. [1994] based on 16S rRNA that grouped the species of Class Clostridia into 24 groups, commonly referred to as "Cluster #" Clostridia, representing groups of bacteria that are genetically similar and inhabit similar niches. In this tree, the *Clostridium sensu stricto* were established as Cluster 1 Clostridia.

A set of comprehensive trees were published in Bergey's Manual of Systematic Bacteriology [Ludwig et al., 2015] and based on 16S rRNA. This tree suggests a topology (summarized in Fig 2.2A) where many genomes of "early-branching" Clostridia diverged prior to the separation of the Bacilli class. These early-branching genomes include the families Thermoanaerobacter and several families of syntrobiotic or symbiobiotic Firmicutes. Relationships within Class Clostridia are largely unresolved in this tree.

Although 16S rRNA has commonly been used to determine species phylogenies, protein-based phylogenies have recently come in to favor. Amino acid sequences have more phylogenetic information, more accurate models, larger available data sets, and more efficient algorithms than rRNA. Further, the use of amino acid data, which may also require increased memory and calculation speed, is facilitated by increases in computing power. For species trees, more recent approaches have used ribosomal protein data to build species-level phylogenies [Antunes et al., 2016; Kunisawa, 2015; Yutin and Galperin, 2013]. Ribosomal proteins may provide the benefits of amino acid data and are thought to be rarely transferred or duplicated. Here I focus on two of the most recent and comprehensive Firmicutes species tree studies based on this type of data.

The Yutin and Galperin [2013] tree was built on 70 Firmicutes with the goal of placing the Veillonellaceae in the context of the phylum and to re-classify several species of mis-placed *Clostridium* [Yutin and Galperin, 2013]. This tree was created using a concatenated sequence alignment of 50 ribosomal proteins as data. To root this tree, two outgroup species, *Leptotrichia buccalis* and

Fusobacterium nucleatum, were included.

Antunes et al. [2016] was constructed using a larger set of taxa (205 genomes) using multiple tree-building methods and a ribosomal protein dataset similar to that used in the Yutin and Galperin [2013] tree. These trees were built to confidently place the Selenomonadales/Veillonellaceae group and the Halanaerobiales/Natranaerobiales group for the purposes of determining the ancestral state and evolutionary history of monoderm and diderm cellular envelopes. To root these trees, thirteen outgroup species were also included in the analysis. This study presents species trees built using both maximum likelihood and a Bayesian approach. Trees built using either method are topologically similar and well-supported.

Comparison of the Yutin and Antunes trees reveals similar topology at the highest ranks of taxonomy (Figure 2.2). These trees differ in their placement of the anaerobic halophiles (Halanaerobiales and Natranaerobiales) — in the Antunes tree, this clade is predicted to have diverged prior to the divergence of the two Firmicutes classes, Clostridia and Bacilli; in the Yutin tree, it is predicted that this clade diverged from Class Clostridia after the divergence of the Bacilli, but the relationship to other clades within Class Clostridia is not resolved in this tree. The resolution of families within Class Clostridia in the Antunes tree compared to the Yutin tree also leads to other differences, such as the prediction that families Lachnospiraceae and Ruminoccocaceae are sister taxa and the placement of the order Thermoanaerobacterales is basal to those two families and the Clostridiaceae. This ordering is consistent with the unresolved branches of the Yutin tree.

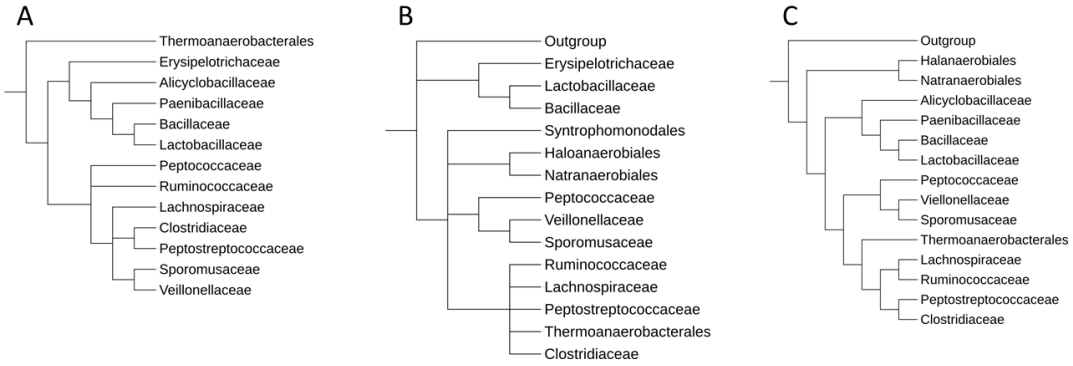


Figure 2.2: Comparison of three Firmicutes phylogenies. Phylogenetic summaries show family and order levels of taxonomy based on trees presented in the main text of (A) Bergey's Manual of Systematic Bacteriology (16S rRNA). (B) Yutin and Galperin [2013] tree (50 ribosomal proteins). (C) Antunes et al. [2016] (46 ribosomal proteins).

2.2.3.1. Recent reclassifications within the Firmicutes phylum

In general, taxonomic nomenclature does not always agree with phylogeny. Systematic information is only one of several criteria used in the classification of bacteria, which, prior to the advent of systematics mainly classified and named bacterial species based on phenotypic characteristics, including cellular morphology, niche, and unique metabolic tendencies.

The taxonomic nomenclature and classification within the Firmicutes phylum is currently in flux [Galperin et al., 2016; Lawson et al., 2016]. This is partially due to the historical classification of spore-formers based on aerotolerance. Prior to the 1990s: nearly all anaerobic spore-formers were assigned to the *Clostridium* genus while aerobic spore-formers were assigned to the *Bacillus* genus [Parte, 2014]. With the advent of whole genome sequencing and more sophisticated phylogenetic reconstruction methods, a better understanding of phylogenetic relationships within the Firmicutes is emerging [Kunisawa, 2015]. As such, efforts are being made to reclassify species that were formerly assigned by these general criteria into a more descriptive taxonomic organization. New classifications aim to group species by genetic similarity, metabolic capability, and environmental niche at all levels of taxonomy. Typically, upon reclassification, new names are generated that aim to provide more characteristic information. In cases where it is believed that re-classification is necessary but a new taxonomic organization has not been determined or agreed upon, species are re-named using the convention "*New-genus-name [historical-genus-name] species*".

For example, a case that has been particularly contentious is *Clostridioides difficile*, formerly known as *Clostridium difficile*. In recent phylogenetic reconstructions using both 16S rRNA and ribosomal proteins [Galperin et al., 2016; Lawson et al., 2016], this medically relevant species has been placed with other members of the family Peptostreptococcaceae. Yutin and Galperin [2013] suggested that this species be named "*Peptoclostridium difficile*." However databases and later journal articles rejected this name as this medically-relevant species is colloquially referred to as "C. diff" within medical communities and the general zeitgeist. The "validly described" reclassification of this species is *Clostridiodes difficile*, which preserves the initial "C" while differentiating the species from the *Clostridium sensu stricto* [Lawson et al., 2016].

For a more in-depth look at the current issues in *Clostridium* reclassification, please refer to [Galperin et al., 2016; Lawson et al., 2016; Vos et al., 2009; Yutin and Galperin, 2013]; the most recent major reclassifications within the Bacilli are found in Vos et al. [2009]. Further complicating these reclassification efforts, different databases that contain taxonomic information, such as LPSN [Parte, 2014] and the NCBI taxonomy database [Sayers et al., 2009], have different criteria that must be met for a suggested change in taxonomic assignment or name to be reflected in their

databases.

2.3. Sporulation Initiation (Spo0) Pathway

The sporulation initiation pathway controls the entrance into the cellular differentiation program that results in the production of endospores. This pathway has different signal transduction architectures in two evolutionarily distant genera, the *Clostridium* and *Bacillus*. Although the architectures differ, both pathways result in the phosphorylation of the response regulator, Spo0A, which has been called the "master regulator of sporulation" [Galperin et al., 2012] (Section 2.3.1).

In *B. subtilis*, the sporulation initiation (Spo0) pathway is a multi-input phosphorelay (see Section 2.1.4) characterized extensively [Burbulys et al., 1991; Jiang et al., 2000; Trach and Hoch, 1993] and also observed in closely related species [Bick et al., 2009; Brunsing et al., 2005; Park et al., 2012]. In this architecture, multiple sensor kinases phosphorylate Spo0F, a response regulator lacking an output domain; subsequently, that phosphoryl group is transferred via Spo0B, an intermediate histidine phosphotransferase, to Spo0A, a second response regulator that serves as the master regulator of sporulation (Fig 2.3). Spo0F and Spo0B were not identified in early genomes of *Clostridium* [Stephenson and Hoch, 2002] and one of these genomes was later found to encode a multi-input direct phosphorylation architecture [Steiner et al., 2011].

2.3.1. Regulation of sporulation via the Spo0 pathway

Spo0A, the terminal component of the pathway in both species, initiates spore development upon phosphorylation in both classes [Hoch, 1971; Wörner et al., 2006] and is encoded by all known sporulators [Galperin et al., 2012]. Spo0A is a canonical response regulator protein in its domain composition, including a REC domain [Lewis et al., 1999] and a highly conserved, DNA-binding output domain, Spo0A_C [Lewis et al., 2000]. As is typical in response regulator proteins, phosphorylation of Spo0A activates the output functionality. Specifically, the phosphorylated form of Spo0A "unhinges" (as described in 2.1.1.2) allowing the Spo0A_C output domain to act as a DNA-binding transcription factor. This conformational change also exposes residues that coordinate dimerization [Lewis et al., 2000], which has been implicated in the regulatory role of Spo0A.

The Spo0A_C output domain binds to so-called 0A-boxes that are found upstream of the genes controlled by Spo0A [Burbulys et al., 1991]. For successful sporulation, proteins involved in the machinery that produces spores must be precisely expressed. For many of these proteins, phosphorylated Spo0A plays a role in transcriptional regulation. Phosphorylated Spo0A proteins bind a specific upstream site, consisting of seven basepairs. It has been suggested that the configuration

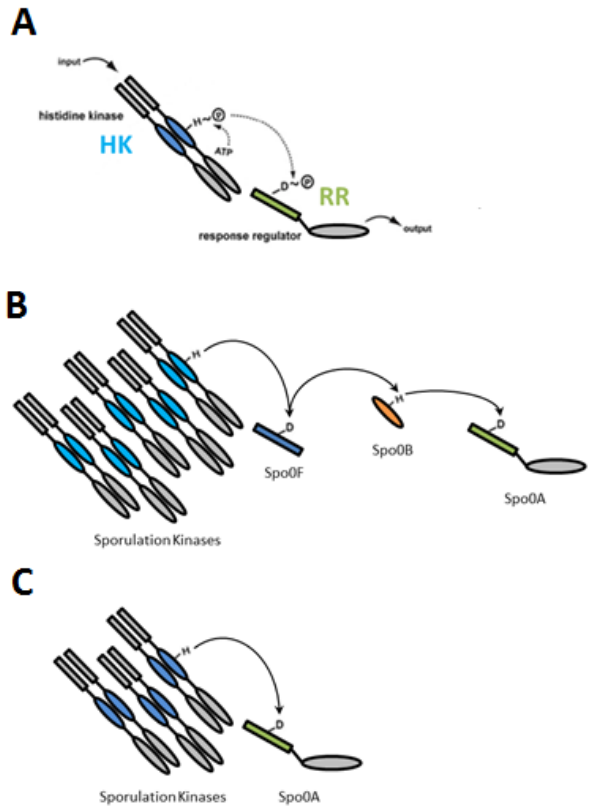


Figure 2.3: Simple and complex interaction architectures involving histidine-aspartate phosphotransfer domains, HisKA (teal oval), REC (blue and green rectangles), and Spo0B (orange oval). A) Canonical TCS system architecture. Interaction between a histidine kinase and response regulator B) *B. subtilis* Spo0 phosphorelay. Multiple sporulation kinases (KinA-E) phosphorylate Spo0F. Subsequently, Spo0B transfers phosphate from Spo0F to Spo0A. C) *C. acetobutylicum* Spo0 direct phosphorylation architecture. Multiple sporulation kinases phosphorylate Spo0A.

of these so-called 0A-boxes can lead to different expression profiles that may be required for the regulation of Spo0 machinery. The nucleotide identity of the variable basepairs of this motif can vary the affinity of Spo0A binding. For example, in the instance of *B. subtilis* Spo0F, a rise-and-fall expression pattern is achieved through multiple upstream 0A-boxes in a specific pattern [Asayama et al., 1995]. The binding of two high affinity 0A-boxes, by dimerized Spo0A is hypothesized to cause a conformation change in the surrounding DNA, which may promote the recruitment of a sigma factor, (σH), that, in turn, recruits polymerase machinery [reviewed in de Hoon et al., 2010]. As the concentration of phosphorylated Spo0A increases, an 0A-box with weaker affinity that is further upstream from the two high-affinity 0A boxes is bound and inhibits the binding of the sigma factor, reducing the expression of Spo0F. This weaker 0A-box is hypothesized to be bound

by non-dimerized Spo0A that may accrue as a result of high concentrations of phosphorylated Spo0A [Asayama et al., 1995]. Evidence for this transcriptional control through dimerized, phosphorylated Spo0A is corroborated by the reduction of DNA-binding [Muchová et al., 2004] and sporulation observed when the dimerization residues of Spo0A are mutated to abolish dimerization between either the N- or C-terminal domains [Seredick et al., 2009].

Regulation of sporulation initiation by phosphorylated Spo0A is modulated by proteins that interact with the Spo0 pathway proteins to remove phosphoryl groups from the system or to inhibit interactions. These include aspartyl-phosphatases, such as *B. subtilis* RapA, B, and E, which target Spo0F, and YisI, YnzD, and Spo0E, which target Spo0A [de Hoon et al., 2010, reviewed in]. In some cases, these phosphatases are additionally inhibited by small peptides: RapA, for example, is inhibited by the small peptide PhrA [Diaz et al., 2012]. There are also proteins that inhibit phosphorylation by direct interaction with Spo0 proteins. *B. subtilis* Sda, for example, binds to the HisKA domain of sporulation kinases to prevent interaction with Spo0F [Rowland et al., 2004]. Intriguingly, Sda is upregulated in response to DNA damage, suggesting that this protein acts as a checkpoint before sporulation can proceed. Finally, proteins such as SinR [Mandic-Mulec et al., 2015] and AbrB [Zuber and Losick, 1987] act as transcriptional repressors in the Spo0A regulon genes. As with many concepts related to sporulation, regulation of the Spo0 pathway was primarily characterized in *B. subtilis*. Although comparative genomic studies have revealed sets of Spo0 regulators in other species [Galperin et al., 2012], the complete set of regulators and their concerted mechanisms in species other than *B. subtilis* is poorly understood.

Regulation of sporulation following the production of phosphorylated Spo0A is mainly through recruitment of specific sporulation sigma factors (mediated by phosphorylated Spo0A). These sigma factors are temporally and locally expressed in either the forespore or the mother cell of the sporulating bacteria [de Hoon et al., 2010]. For example, in the earliest stages of sporulation, Spo0A recruits sigH in the mother cell and sigF in the forespore.

The study of sporulation kinases in the *Clostridium* genus has revealed another apparent method of sporulation regulation: histidine phosphatases. These proteins are similar to histidine kinases in their domain content, though rather than phosphorylating a response regulator they remove the phosphoryl group from response regulators [Casino et al., 2010]. Although it has not been explicitly studied, based on the domain content similarities to histidine kinases, it is intriguing to consider that the specificity of this de-phosphorylation interaction may be governed by the same specificity residues as histidine kinases. Further, these phosphatases may also require stimuli to initiate their phosphatase activity. No specific histidine phosphatases have been reported in Spo0 phosphorelays, although many kinases are suspected to act as both kinases and phosphatases [Casino et al.,

2010].

2.3.1.1. Sporulation and the cell cycle

Sporulation initiation must be tightly linked to the end of the cell cycle, as it is a derivative process from mitosis. Further, the appropriate expression of sporulation proteins in both the mother cell and the forespore is essential to the sporulation differentiation process. Consistent with this logic, inability to fully transfer DNA to the prespore has been observed to result in stage III sporulation arrest with a partially formed forespore [Wu et al., 1995]. These results suggest that if the chromosomal copy is not available, sporulation will not proceed. In turn, this suggests that cellular regulation of the Spo0 pathway must result in syncing the availability of the regulating, phosphorylated Spo0A with the end of the cell cycle when the chromosomal copy is available, or else sporulation will not proceed appropriately. In *B. subtilis* and other organisms that control sporulation through phosphorelays, this timing control may be partially performed by inhibitive small peptides that are independently linked to the cell cycle, such as NprX and Sda that block Spo0F from interacting with sporulation kinases [Perchat et al., 2016]. It is also intriguing to consider that the kinetics of the phosphorelay may result in syncing with the cell cycle. It has been suggested that the availability of Spo0F may temporally inhibit sporulation kinase activity in a concentration dependent manner [Narula et al., 2015] — as such, overexpression of Spo0F leads to decreased sporulation [Grimshaw et al., 1998].

The coordination of the cell cycle and sporulation in species where Spo0A is directly phosphorylated is not well-studied. There is evidence to support the hypothesis of regulation via a feedback loop wherein sporulation kinases are controlled by phosphorylated Spo0A [Steiner et al., 2011].

2.3.2. Experimental Characterization of the Spo0 Pathway

In this section, I review the experimental evidence of Spo0 components, their interaction, and the resulting pathway architectures. The sporulation kinases, and other Spo0 component homologs identified in these studies are catalogued in Table 1 along with their specificity residue signature and PAS domain content.

The well-characterized *B. subtilis* Spo0 pathway has a multi-input phosphorelay architecture. Evidence from experiments in other species in Class Bacilli suggest that their Spo0 pathways are similar in architecture to the *B. subtilis* Spo0 pathway [Bick et al., 2009; Brunsing et al., 2005; Park et al., 2012].

In contrast to these multi-input phosphorelay Spo0 architectures, *C. acetobutylicum* and other

closely related species possess a multi-input architecture in which Spo0A is directly phosphorylated by multiple kinases [Steiner11, Underwood09, Mearls14].

2.3.2.1. Experiments Demonstrating Phosphorelay Architecture of Spo0 pathways

2.3.2.1.1. *B. subtilis* The Spo0 pathway was initially discovered in *Bacillus subtilis*, a model organism for the gram-positive eubacteria (later known as Firmicutes). Determination of the mechanisms responsible for the production of endospores was a primary focus for early geneticists working in this system given its association with infection in both humans and livestock. As such, analysis of mutations that caused reduction or elimination of sporulation was common. These mutations were categorized by the stage in which sporulation was arrested. Any mutation that was demonstrated to occur prior to formation of the prespore and excretion of antibiotics and proteases was considered a mutant in Stage 0, or sporulation initiation. Mutants that appeared in different regions of the chromosome were given a letter to denote that region.

By 1976, nine Spo0 genes had been identified among several hundred mutations that were mapped [Hoch, 1976]. As understanding of the Spo0 pathway became better understood, the number of Spo0 pathway proteins would decrease to five. Two mutations were found to be within previously characterized Spo0 components: Spo0D and Spo0C were found to be mutations within *spo0B* and *spo0A* respectively. Other genes initially given the Spo0 prefix were better characterized and found to not participate in the histidine-aspartate phosphotransfer pathway, including *spo0H*, which was found to be a sigma factor and renamed σ^H [Dubnau et al., 1988]; *spo0K*, an oligopeptide permease that likely plays a role in small-peptide import that may include a sporulation stimulus [Rudner et al., 1991]; and finally, *spo0J*, a ParB family protein involved in chromosomal segregation during the initiation of sporulation [Ireton et al., 1994].

A combination of mutants was used to elucidate a possible interaction pattern at a time when little was known about the structure of these proteins. Specifically, a particular mutant of *B. subtilis* Spo0A, sof-1, was demonstrated to recover sporulation in the presence of mutations in either Spo0F, Spo0B, or Spo0E [Hoch et al., 1985]. The rescue of these other mutations implies that the other Spo0 proteins "interact in a sequential or concerted fashion to effect the activity of the Spo0A gene product" [Hoch et al., 1985]. This sporulation defect suppressor mutation was mapped to the 12th codon of the Spo0A locus, altering an asparagine to a lysine (N12K).

We now know that the sof-1 Spo0A mutant altered the first specificity residue from N to K, likely introducing direct phosphorylation of Spo0A by either a sporulation kinase or some other kinase [LeDeaux et al., 1995]. Alternatively, this amino acid change may have solely or additionally changed the dynamics of dephosphorylation by Spo0E, allowing the usually low rate of

Spo0A phosphorylation by KinA and other sporulation kinases to be sufficient to initiate sporulation. Although not included in the study published by LeDeaux et al. [1995], the associated thesis demonstrated that KinC, later found to be a sporulation kinase [Jiang et al., 2000], directly phosphorylates the sof-1 version of Spo0A, suggesting that direct phosphorylation is indeed occurring.

Each of the products of genes in which a Spo0 mutation was identified were further characterized by sequence analysis in subsequent studies. While the primary sequence of *spo0B* did not reveal a function for the gene product [Ferrari et al., 1985a], the primary sequences of both Spo0F and Spo0A were identified to be homologous to OmpR, a regulatory protein [Ferrari et al., 1985b; Trach et al., 1985]. However, Spo0F notably lacks the C-terminal region typically associated with this protein family [Trach et al., 1985] (see Section 2.1.1.2). This homology suggested that there should be a corresponding "transmitter kinase" that interacts with these proteins to activate them.

A potential transmitter kinase was identified as the *spo0IIJ* gene [Antoniewski et al., 1990]. This gene was categorized as a sporulation stage II gene, as mutations to this gene caused sporulation arrest prior to the transcription of several other SpoII genes. However, such arrest at stage II was intermittent, even with the same strains harboring mutations within this gene, and stage 0 blockage was also inconsistently observed [Antoniewski et al., 1990, unpublished A. Ryter, cited in]. Sequence analysis revealed that the product of the *spo0IIJ* gene shares homology with the transmitter kinase family, especially in the C-terminal transmitter region. In contrast to typical transmitter kinases, which are encoded adjacent to a cognate response regulator, this protein is an orphan kinase, lacking such an adjacent interaction partner. This led to speculation that *spoIIJ* encodes a kinase that may be specific for Spo0F or Spo0A, given the occasional effect of Spo0IIJ mutations to block spore production at stage 0. The Spo0IIJ gene product, renamed KinA, was later demonstrated to phosphorylate Spo0F, and, to a lesser extent, Spo0A, confirming a role in in sporulation initiation [Perego et al., 1989]. Note, however, that these interactions were observed after an incubation period of 1 hour at 37C with autophosphorylated KinA and may not represent physiological conditions.

Finally, a phosphotransfer assay was performed using KinA, Spo0F, Spo0B, and Spo0A purified from *B. subtilis* that demonstrated that these proteins interact sequentially [Burbulys et al., 1991]. This series of interactions was termed a "multi-component phosphorelay". This study additionally confirmed that the primary target of KinA is Spo0F and that Spo0B transfers a phosphoryl group from Spo0F to Spo0A, likely through a mechanism similar to that of the transmitter kinases.

Four additional orphan transmitter kinases were also found to be specific for Spo0F [Jiang et al., 1999, 2000; Trach and Hoch, 1993]. Each of these kinases were identified by homology to KinA and orphan status; there is experimental evidence for interaction with Spo0F for each

of these kinases by phosphotransfer assay, though KinB [Trach and Hoch, 1993] was initially identified by a sporulation efficiency assay in a double deletion mutant. Of these five sporulation kinases, the largest reduction of sporulation efficiency is observed for KinA, followed by KinB, and the remaining kinases contribute substantially less. It may be interesting to study the sporulation kinases that effect sporulation less in different conditions, as these may be sensitive to different stimuli not commonly tested in sporulation efficiency assays.

2.3.2.1.2. *Bacillus anthracis* The sporulation kinases of *Bacillus anthracis* were identified by sequence similarity to the interaction domain of the *B. subtilis* sporulation kinases. These nine kinases were identified using a BLASTP search with *B. subtilis* KinA as a query. Next, these kinases were experimentally characterized by their ability to drive sporulation initiation in B. subtilis when under their own promoter or the promoter of BA4223, which was found to be highly active in the backgrounds tested. Although orphan kinases were not the target of this BLAST search, all nine kinases identified were also orphans. Further inspection reveals that this is the full complement of orphan kinases in *B. anthracis*. Of those tested, only BA4223 and BA2291 restored sporulation to significant levels in a Δ KinA and Δ KinB background. BA1356 also restored sporulation in the double kinase knockout but to a much lesser degree and increased sporulation much more in a single knockout (δ KinA) mutant. Expression of BA5029 under the promoter of BA4223 resulted in sporulation rescue under the single knockout (δ KinA). Expression of BA1351 in the wildtype background eradicated sporulation ability, suggesting it is a phosphatase and possibly acting as a Spo0E homolog. Overexpression of BA2291 also reduced sporulation percentage nearly to zero, suggesting it has phosphatase activity at high concentrations, though this could be an artifact of overexpression. I suggest an alternative hypothesis: BA2291 has some affinity for *B. subtilis* Spo0A and its overexpression resulted in inappropriate crosstalk to Spo0A such that sporulation was not able to proceed as normal. These mixed results suggest that these four kinases and one phosphatase are involved in sporulation in *B. anthracis*.

The remaining kinases, BA2636, BA3702, BA1478, BA5029, and BA2644 did not show any rescue of sporulation deficiency. Other experiments were done to examine the activity of the promoters of these proteins and sporulation ability after single knockout in *B. anthracis*. These experiments do not directly characterize these proteins as sporulation kinases, though the authors suggest that the knockout experiments show that there is no kinase that regulates sporulation initiation more than any other (unlike in *B. subtilis* where knockouts of Spo0A have greatly reduced sporulation).

Although the downstream components of the phosphorelay were not directly tested by the Brunsing et al. [2005] study, it is mentioned that they are extremely similar in sequence to *B.*

subtilis including "identity throughout the interaction surface."

2.3.2.1.3. *Paenibacillus polymyxa*

Spo0F and Spo0A of *P. polymyxa* were identified to be highly conserved with those of *B. subtilis*. Although *P. polymyxa* Spo0B is conserved with *B. subtilis* throughout its region of interaction, it is varied throughout the rest of the protein, sharing only 14% identity. Additional evidence that this protein is a Spo0B homolog was the presence of an upstream ObgE protein.

P. polymyxa Spo0A was shown to partially recover sporulation in the Spo0A knockout strain of *B. subtilis*. It is unclear if this recovery is partial due to the inability of the Spo0A_C domain to interact with the "0A box" promoter regions of *B. subtilis* or the inability of the *P. polymyxa* Spo0A interaction domain to receive phosphate from *B. subtilis* Spo0B. Later results suggest the latter.

Sporulation kinases in *P. polymyxa* were identified by sequence comparison of orphan kinases to *B. subtilis* KinA and KinB. Five orphan kinases were tested by expression in a KinA and KinB deletion strain of *B. subtilis*. Only two kinases, PP_1038 and PP_1377 were able to restore sporulation in *B. subtilis*. PP_1038 was additionally able to restore sporulation in a Spo0B deletion strain of *B. subtilis* suggesting that it is able to directly phosphorylate *B. subtilis* Spo0A. Surprisingly, sporulation was restored to nearly wild-type levels in a Spo0B knockout strain of *B. subtilis* expressing *P. polymyxa* PP_1038 and *P. polymyxa* Spo0A. These results suggest that, at least in the *B. subtilis* background, PP_1038 is able to directly phosphorylate its own Spo0A.

2.3.2.1.4. *Geobacillus stearothermophilus*

One orphan kinase (WP_013143992.1) of *G. stearothermophilus* was shown to interact with its own Spo0F by *in vitro* phosphotransfer [Bick et al., 2009]. Spo0B and Spo0A were not reported in this study, though the presence of Spo0F and its interaction with an orphan kinase suggest that this organism encodes a phosphorelay.

2.3.2.2. Direct Phosphorylation Architecture

Spo0F and Spo0B were found to be absent from the early genomes of Class Clostridia [Stephenson and Hoch, 2002]. As the absence of these proteins precludes the phosphorelay architecture, this initially this led to myriad theories, perhaps best captured by Paredes et al. [2005], including "an unknown phosphorelay, spontaneous acetyl/butyl phosphotransfer, or direct phosphotransfer." However, with the recent development of tools for genetic manipulation of Clostridial species [Kuehne et al., 2011], experimental characterization of the Spo0 pathway in *Clostridium* species has become possible [Steiner et al., 2011].

2.3.2.2.1. *C. botulinum* Research into the Spo0 pathway of the Clostridia began with characterization of the *Clostridium botulinum* Spo0A in a *B. subtilis* background [Wörner et al., 2006]. Unfortunately, the negative results of rescue and transcriptional activation experiments suggest that the *C. botulinum* Spo0A (NEFISDD) is not phosphorylated in the *B. subtilis* background. The addition of a *C. botulinum* orphan kinase and the *C. botulinum* Spo0A to a Δ Spo0A *B. subtilis* background, proved lethal. The orphan kinase on its own in this background did not result in lethality, suggesting that phosphorylated *C. botulinum* Spo0A caused the lethality. However, a chimeric Spo0A with the REC domain of *B. subtilis* Spo0A and the *C. botulinum* Spo0A_C domain was able to sporulate. These mixed results suggest that the Spo0A_C domain of *C. botulinum* can initiate transcription at "0A" boxes, but the effect of phosphorylation of *C. botulinum* Spo0A REC domain remains unclear.

2.3.2.2.2. *C. difficile* Two sporulation kinases in *Clostridioides difficile* were experimentally verified using two different methods: CD2492 by reduction of sporulation in a single deletion mutant and CD1579 by *in vitro* phosphotransfer. The authors of this study note a third *C. difficile* sporulation kinase is suggested by orphan status and sequence homology in the HisKA domain, though it was not tested.

2.3.2.2.3. *C. acetobutylicum* Sporulation efficiency and *in vitro* phosphotransfer experiments were performed in *Clostridia acetobutylicum*, demonstrating that three orphan kinases are individually sufficient for the phosphorylation of Spo0A [Steiner et al., 2011].

Single deletion mutants of each of the five orphan kinases in C. acetobutylicum, were screened in a sporulation assay. Three of those five kinase knockouts showed a reduction in sporulation efficiency, suggesting that these three kinases are sporulation kinases. Another of the five kinases showed an increase in sporulation yield, suggesting that it is a phosphatase, preventing activation of the sporulation pathway. In a set of double knockouts, two of the kinases showed even greater sporulation yield, suggesting that these kinases work in conjunction.

Based on these *in vivo* results, two of the three kinases were tested for *in vitro* phosphotransfer to *C. acetobutylicum* Spo0A. They additionally tested those kinases for phosphotransfer with B. subtilis Spo0F and Spo0A. Results were mixed for the two *C. acetobutylicum* kinases tested: both kinases were able to phosphorylate *B. subtilis* Spo0A and Ca_C3319 was additionally able to phosphorylate *B. subtilis* Spo0F (though relatively weakly).

2.3.2.2.4. *R. thermocellum* Finally, sporulation efficiency in single deletion mutants of three orphan kinases and Spo0A was measured in *Ruminoclostridium thermocellum* [Mearls and

Lynd, 2014]. In each knock out, sporulation was eradicated, suggesting that these proteins work in concert to initiate sporulation. As *in vitro* phosphotransfer experiments were not performed, the underlying architecture of this Spo0 pathway is unclear. Further, one of the kinases tested, clo1313_1942, is a hybrid RR, consisting of an N-terminal REC domain and C-terminal HisKA and HK_CA domains. It is intriguing to consider that this hybrid kinase may require phosphorylation of the REC domain for kinase activity to proceed — if this is the case, the result would be an alternative phosphorelay architecture compared to that of the Spo0 pathway in *B. subtilis*.

2.3.2.2.5. Enigmatic impact of Sporulation kinases in direct phosphorylation architecture The dynamics of sporulation kinases in Clostridia species remains enigmatic. Two of the three *C. acetobutylicum* sporulation kinases appear to work together, while the third is independent. In *C. difficile* knockout of CD2492 results in sporulation rates dropping from 14% to 4%, implying other kinases additionally phosphorylate Spo0A, such as CD1579. Finally, *R. thermocellum* sporulation kinases are each necessary, as knockout of any of the three results in reduction of sporulation equal to knockout of Spo0A.

2.3.2.2.6. Bioinformatic identification of Spo0B and Spo0F in Class Clostridia Spo0B homologs have been computationally predicted using a combination of BLAST and genomic region conservation [Mattoo et al., 2008] in species ranging from Class Bacilli to several examples in Class Clostridia. None of these except *B. subtilis* have been experimentally investigated for interaction with Spo0F and Spo0A. In another study, putative phosphorelay proteins have been identified in the genus *Desulfotomaculum* [Dalla Vecchia et al., 2014], using sequence similarity methods to identify potential sporulation kinases and Spo0F as well as genomic region conservation to identify Spo0B. Comparison of Spo0F residues involved in interaction with Spo0B (though not the exact set of specificity residues defined by Skerker et al. [2008] and Capra et al. [2010]) reveal similarities to *B. subtilis* specificity residues and suggest interaction. The identification of Spo0F and Spo0B homologs in Class Clostridia by these studies suggest that the phosphorelay architecture may not be isolated to Class Bacilli.

2.3.3. Evolution of the Spo0 pathway architectures

Considering that there are two different signal transduction architectures that both orchestrate the initiation of sporulation through the phosphorylation of an orthologous output RR, they likely arose from a common ancestral pathway. How then, did different signaling architectures evolve in present day species? Given the two distinct architectures observed, remodeling of the signal

transduction architecture that results in the phosphorylation of Spo0A must have occurred. Understanding of the nature of this remodeling between the Spo0 pathway is incomplete. The prevailing view is that the ancestral Spo0 pathway was a direct phosphorylation architecture and the more complicated phosphorelay arose in the Bacillar lineage [Durre, 2011; Stephenson and Hoch, 2002; Talukdar et al., 2015].

2.3.3.1. The prevailing view of the evolution of sporulation initiation

Predictions that the original Spo0 pathway, encoded by the ancestor of all Firmicutes, had a direct phosphorylation architecture are fueled by its simplicity and the similarly anaerobic lifestyles of the ancestral Firmicute and present-day Clostridia [Talukdar et al., 2015]. Experimental characterization of the Spo0 pathway in *C. acetobutylicum* wherein Spo0A is directly phosphorylated [Steiner et al., 2011], along with the apparent lack of Spo0F and Spo0B proteins in sequenced Clostridium genomes [Stephenson and Lewis, 2005] gave more evidence to this ancestral direct phosphorylation architecture hypothesis [Durre, 2011]. The expansion into a phosphorelay within the Bacilli is supported by the observation that this architecture offers more potential points of control that may have allowed early Bacillar species to adapt to rising oxygen levels during the Great Oxidation Event [Durre, 2014].

Chapter 3

Prediction of Spo0 Proteins in Firmicutes Genomes

In order to study the evolutionary processes and forces that shaped the Spo0 pathway, I first investigated the evolutionary history of this pathway, which remains poorly understood. The hypothesis that the ancestral Spo0 pathway was a two-component, direct phosphorylation architecture was proposed in the when only a few Firmicutes genomes had been sequenced, primarily genomes within the evolutionarily distant genera, *Bacillus* and *Clostridia*. In the intervening years, hundreds of sequences within the Firmicutes have become available, but the architecture of the Spo0 pathway encoded in these genomes has received little attention. The first step for such an analysis is the prediction of homologs of the known Spo0 proteins across the Firmicutes phylum.

In this chapter, I present a survey of Spo0 proteins in a representative set of Firmicutes, broadly sampled to cover the phylum's diversity. Prediction of sporulation kinases, Spo0F, and Spo0B proved to be non-trivial, requiring specific criteria to distinguish these proteins from other histidine-aspartate phosphotransfer signal transduction proteins. These criteria revealed putative homologs of Spo0 proteins, including the intermediate phosphorelay proteins, Spo0F and Spo0B, in many spore-formers belonging to both the Bacillar and Clostridial lineages. In spore-former genomes where homologs of all four of the Spo0 phosphorelay components have been identified, it is likely that the Spo0 pathway has a phosphorelay architecture. Other spore-formers, primarily within the Clostridial lineage, were found not to encode either Spo0F or Spo0B, suggesting phosphorylation of Spo0A directly by a sporulation kinase, as observed in *C. acetobutylicum*.

Prior to this work, there have been sparse reports of proteins that may be homologous to Spo0F and Spo0B within Clostridial genomes [Dalla Vecchia et al., 2014; Galperin, 2006; Mattoo et al.,

2008]. However, due to lack of characterization, these reports have not resulted in re-definition of the Spo0 pathway architectures in Class Clostridia. Thus, in addition to presenting the most comprehensive set of homologs of Spo0F and Spo0B in Class Clostridia genomes, I provide bioinformatic and experimental evidence that all Spo0 component homologs that I have identified are functional. Specifically, I compare the specificity residues to experimentally verified examples and demonstrate that a Clostridial phosphorelay is functional through *in vitro* phosphotransfer experiments.

3.1. Representative Set of Firmicutes Genomes

To facilitate a survey of present day Spo0 architectures, I assembled a set of 84 whole genome sequences that are representative of the two major sporogenous Firmicutes Classes: the Clostridia and the Bacilli. These genomes, listed in Supplementary Table A1, were used for a comparative genomic analysis and construction of a species phylogeny.

I seeded my initial set based on the genomes presented in a recently published phylogeny [Yutin and Galperin, 2013]. The genomes for the Yutin and Galperin [2013] tree were sampled to assure coverage of the class Clostridia clusters presented by Collins et al. [1994]. I noted the absence of three prominent genera presented in Bergey's Manual of Systematic Bacteriology [Vos et al., 2009], and added one genome from each of these — *Mahella australiensis*, *Acetivibrio cellulolyticus*, and *Dorea formicigenerans*. Further, this set only contained 8 species of class Bacilli and, notably, several major Bacillar lineages were not represented. Thus, I added several genomes to obtain broader coverage in this class, guided by a prior phylogenetic study [Wei Wang, 2009], based on 16S rRNA, that split the Bacilli into nine groups. Bacilli genomes were selected such that there was at least one representative from each group.

This process resulted in a set of 94 Firmicutes genomes and two outgroup genomes to root the tree. Several of these genomes, subsequently, proved to be topologically unstable during phylogeny reconstruction. Several of these genomes had very long branches, increasing the risk of long branch attraction artifacts. Most were associated with species believed to be asporogenous [Galperin et al., 2012] and are therefore not central to my analyses. These were removed from the analysis, resulting in the final set of 84 Firmicutes genomes and one outgroup genome.

3.1.1. Species Tree

I constructed a maximum likelihood phylogeny for these 84 representative genomes from a concatenated alignment of 50 ribosomal proteins. Ribosomal proteins have largely congruent phyloge-

netic signal in the Firmicutes [Antunes et al., 2016], and phylogenies constructed on concatenated ribosomal protein sequences provide robust relationships in this phylum .

3.1.1.1. Data Collection

The set of aligned ribosomal protein sequences on which the Yutin and Galperin [2013] tree was constructed was generously provided by Dr. Michael Galperin and Dr. Natalya Yutin. To collect the ribosomal protein sequences for the species not found in the Yutin and Galperin [2013] tree, I identified the ribosomal protein sequences using HMMER3 [Eddy, 2008]. HMMER takes an multiple sequence alignment as input and generates a profile hidden Markov model, which is then used to search the sequence database. Searching with a model that reflects the properties of many sequences from a homologous family, instead of a single query sequence, increases the sensitivity of the search. [Eddy, 2008].

Profiles were constructed from the multiple sequence alignments provided by Yutin and Galperin [2013] for each ribosomal protein. The proteins associated with the additional genomes not represented in their data set were downloaded from the NCBI protein database in FASTA format and concatenated for use as a database. As typically only one copy of each ribosomal protein family will be encoded in each genome, I took the highest E-value hit for each species as the ribosomal protein of interest. The extraction of samples was done using custom Java code to search for the gene of the top hit in the original NCBI protein FASTA files.

3.1.1.2. Phylogeny Reconstruction

A multiple sequence alignment for each ribosomal protein family was constructed separately using GUIDANCE2 [Sela et al., 2015] with MAFFT [Katoh et al., 2005] to construct the underlying multiple sequence alignment. Columns possessing at least 50% gaps or a GUIDANCE alignment score below 92% were trimmed from the alignment. Next, the 50 trimmed multiple sequence alignments were concatenated into a supergene alignment. After concatenation, TIGER was used to eliminate uninformative sites [Cummins and McInerney, 2011]. TIGER analysis was performed to group sites into ten bins that are predicted to be evolving at similar rates. The most rapidly evolving sites (Bin_10) were removed, along with columns that were less informative than a randomized site (ptp test with defaults, Bin_Disagreement). The maximum likelihood species tree (Fig 3.1) was built from the resulting alignment using RaxML version 8.2 [Stamatakis, 2014] with the RAxML implementation of the CAT model [Stamatakis, 2006], which accounts for site-specific heterogeneity, and bootstrapped with 100 replicates (branch labels in Fig 3.1).

3.1.1.3. Results

The resulting phylogeny (Fig 3.1) supports early divergence of Classes Bacilli and Clostridia and the divergence of orders and families that is consistent with other protein-based phylogenies [Antunes et al., 2016; Yutin and Galperin, 2013]. The position of several taxa conflicts with accepted phylogenies. *B. subtilis* is placed basal to the Lactobacillaceae in our tree. The bootstrap support for this configuration is 30. *B. subtilis* is placed deep within the *Bacillus sensu stricto* clade in the Antunes et al. [2016] tree and most other accepted phylogenies [Wei Wang, 2009]. The placement of *D. acetoxidans* in the Antunes tree and in our tree differ slightly. Our topology supports the placement of *D. acetoxidans* as a sister taxon to *Pelatomaculum thermopropionicum*. This pair form a sister taxon to the clade consisting of *Desulfotomaculum reducens* and *Desulfotomaculum ruminis*. In the Antunes tree, *D. acetoxidans* is basal to the clade consisting of *P. thermopropionicum*, *D. reducens*, and *D. ruminis*.

My tree supports the placement of the Erysipelatrichaceae within the Bacillales. This placement is consistent with the Yutin tree, which places these species as a divergent clade from other class Bacilli species. However, my set includes additional early branching genomes in Class Bacilli and suggests the divergence of the Erysipelotrichaceae is internal to class Bacilli. It is unclear if this result is due to phylogenetic error as this is well-supported in my tree, and I was unable to identify any phylogenies including at least one genome sampled from each of the Paenibacillaceae, Bacillaceae, and Erysipelatrichaceae.

3.2. Identification of Novel Phosphorelay Homologs

In this section, I present the identification of Spo0 components in a representative set of Firmicutes in order to characterize the evolutionary history of the Spo0 pathway. In each case, the criteria I present for homology assignment are based on a comparative genomics approach with known examples of these proteins. While Spo0A is readily identifiable by its conserved output domain, standard homology identification did not fair well with the other Spo0 components. Sequence similarity methods were not sensitive enough to distinguish Spo0F and Spo0B orthologs from other protein families with similar domain architecture. This issue stems from the commonality of the interaction domains observed throughout histidine-aspartate phosphotransfer signal transduction architectures. As a solution to this issue, I developed a comparative genomics approach that leverages similarities between the genome neighborhoods of known examples of Spo0 proteins to identify potential orthologs in other genomes. This approach readily identified candidate orthologs to Spo0F and Spo0B. Unexpectedly, these proteins were identified in both major lineages of the

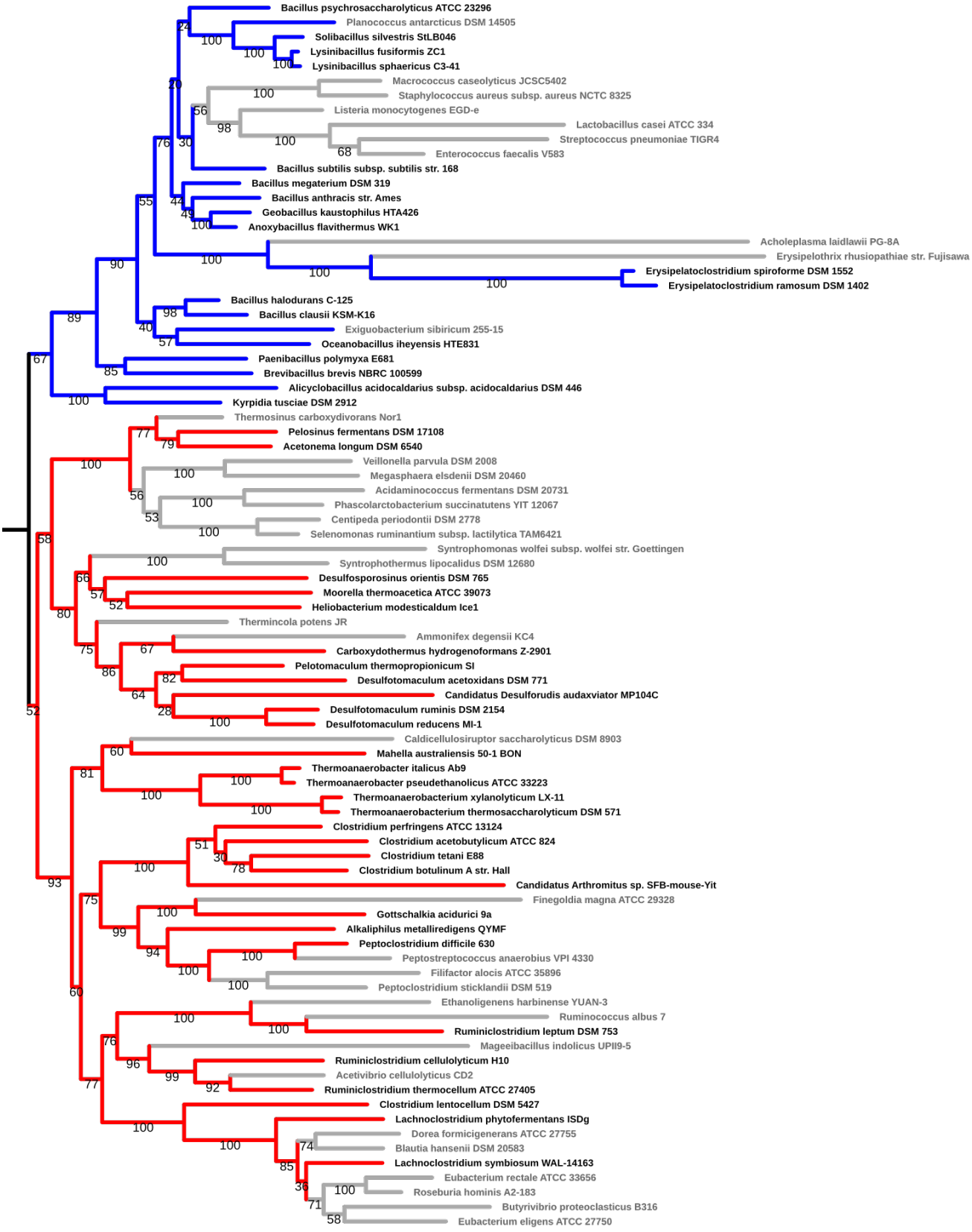

Bacillus psychrosaccharolyticus ATCC 23296
Planococcus antarcticus DSM 14505
Solibacillus silvestris StLB046
Lysinibacillus fusiformis ZC1
Lysinibacillus sphaericus C3-41
Macrococcus caseolyticus JCSC5402
Staphylococcus aureus subsp. aureus NCTC 8325
Listeria monocytogenes EGD-e
Lactobacillus casei ATCC 334
Streptococcus pneumoniae TIGR4
Enterococcus faecalis V583
Bacillus subtilis subsp. subtilis str. 168
Bacillus megaterium DSM 319
Bacillus anthracis str. Ames
Geobacillus kaustophilus HTA426
Anoxybacillus flavithermus WK1
Acholeplasma laidlawii PG-8A
Erysipelothrix rhusiopathiae str. Fujisawa
Erysipelatoclostridium spiroforme DSM 1552
Erysipelatoclostridium ramosum DSM 1402
Bacillus halodurans C-125
Bacillus clausii KSM-K16
Exiguobacterium sibiricum 255-15
Oceanobacillus iheyensis HTE831
Paenibacillus polymyxa E681
Brevibacillus brevis NBRC 100599
Alicyclobacillus acidocaldarius subsp. acidocaldarius DSM 446
Kyrpidia tusciae DSM 2912
Thermosinus carboxydivorans Nor1
Pelosinus fermentans DSM 17108
Acetonema longum DSM 6540
Veillonella parvula DSM 2008
Megasphaera elsdenii DSM 20460
Acidaminococcus fermentans DSM 20731
Phascolarctobacterium succinatutens YIT 12067
Centipeda periodontii DSM 2778
Selenomonas ruminantium subsp. lactilytica TAM6421
Syntrophomonas wolfei subsp. wolfei str. Goettingen
Syntrophothermus lipocalidus DSM 12680
Desulfosporosinus orientis DSM 765
Moorella thermoacetica ATCC 39073
Heliobacterium modesticaldum Ice1
Thermincola potens JR
Ammonifex degensii KC4
Carboxydothermus hydrogenoformans Z-2901
Pelotomaculum thermopropionicum SI
Desulfotomaculum acetoxidans DSM 771
Candidatus Desulforudis audaxviator MP104C
Desulfotomaculum ruminis DSM 2154
Desulfotomaculum reducens MI-1
Caldicellulosiruptor saccharolyticus DSM 8903
Mahella australiensis 50-1 BON
Thermoanaerobacter italicus Ab9
Thermoanaerobacter pseudethanolicus ATCC 33223
Thermoanaerobacterium xylanolyticum LX-11
Thermoanaerobacterium thermosaccharolyticum DSM 571
Clostridium perfringens ATCC 13124
Clostridium acetobutylicum ATCC 824
Clostridium tetani E88
Clostridium botulinum A str. Hall
Candidatus Arthromitus sp. SFB-mouse-Yit
Finegoldia magna ATCC 29328
Gottschalkia acidurici 9a
Alkaliphilus metalliredigens QYMF
Peptoclostridium difficile 630
Peptostreptococcus anaerobius VPI 4330
Filifactor alocis ATCC 35896
Peptoclostridium sticklandii DSM 519
Ethanoligenens harbinense YUAN-3
Ruminococcus albus 7
Ruminiclostridium leptum DSM 753
Mageeibacillus indolicus UPII9-5
Ruminiclostridium cellulolyticum H10
Acetivibrio cellulolyticus CD2
Ruminiclostridium thermocellum ATCC 27405
Clostridium lentocellum DSM 5427
Lachnoclostridium phytofermentans ISDg
Dorea formicigenerans ATCC 27755
Blautia hansenii DSM 20583
Lachnoclostridium symbiosum WAL-14163
Eubacterium rectale ATCC 33656
Roseburia hominis A2-183
Butyrivibrio proteoclasticus B316
Eubacterium eligens ATCC 27750

Figure 3.1: Phylogeny of 84 representative Firmicutes genomes constructed for this study, represented as a phylogram, outgroup rooted with *Leptotrichia buccalis*. Tree constructed from the concatenated alignment of 50 ribosomal protein sequences using RaxML, as described in above. Colored branches indicate known sporulating species of Class Bacilli (blue) and Class Clostridia (red).

Firmicutes Phylum.

In order to provide a complete characterization of the evolution of the Spo0 pathway architecture, prediction of sporulation kinases is also required. I treat these separately.

Data Collection Methods

Protein information for this survey was obtained from MiST 2.2 [Ulrich and Zhulin, 2010] through their website GUI, found at www.mistdb.com. This database was created to catalogue the signal transduction proteins of prokaryotic organisms including one- and two-component signaling systems, as well as hybrid histidine kinases and response regulators. Much of this information is accessible using URL-based queries. Specially formatted URLs allow direct input to the search function and genome, gene and protein summary pages. Although it is not a publicized feature of MistDB, all ORFs in all of the included genomes are accessible in the database and are annotated with locus name, RefSeq information [O'Leary et al., 2016], domain content, and external cross-reference identifiers (including Uniprot, PDB, and NCBI accession numbers). Domain content in MistDB is predicted using the domain models from the PFAM database (version 26 [Punta et al., 2012]) and a custom HMM-based domain classifier that specializes in signal transduction domains (Agfam version 1 [Alexander and Zhulin, 2007]). Adjacent genes are organized by consecutive MistDB protein IDs. I have created custom Java code to access the search or adjacent gene functions of the URL-based query services of MistDB, download the HTML page source, and parse for genome names, sequence identifiers (gene locus), domain content, and amino acid sequences. It should be noted that this database was created as a snapshot, and annotations are based on the information available in 2012 when the website updated to MiST 2.2.

3.2.1. Identification of Spo0A by Domain Content

All known Spo0A proteins, spanning the entire phylum, encode a conserved C-terminal output domain called Spo0A_C. This domain is a typical DNA-binding helix-turn-helix output domain, but has a unique sequence compared to other output domains of these families, permitting the

development of a PFAM domain model. Thus, Spo0A proteins can be reliably identified by domain content, specifically those proteins that contain only a REC domain and a Spo0A_C domain.

Within the representative set, I identified 68 genomes, broadly distributed across the Firmicutes phylum, that encode an apparent Spo0A ortholog (Fig 3.3, green dots). Of these, 53 are reported to form spores (Table A1, red leaves in Class Clostridia, blue leaves in Class Bacilli Fig 3.1). The presence of Spo0A in 15 reported non-spore-formers (Table A1) could be due to a recent loss of sporulation or an alternate functional role for Spo0A in those species. It is also possible that these species are sporogenous, but spore formation has not been observed under the conditions tested [Browne et al., 2016; Galperin et al., 2012] (see also Section 2.2.2.2).

3.2.2. Identification of Spo0F and Spo0B

The intermediate phosphorelay proteins, Spo0F and Spo0B, along with a sporulation kinase, are necessary to phosphorylate Spo0A and initiate sporulation in *Bacillus subtilis* [Burbulys et al., 1991] and likely also in other Bacillar pathways [Bick et al., 2009; Brunsing et al., 2005; Park et al., 2012] (see also Section 2.3.2). The only reported functional roles for Spo0F and Spo0B are as signal transduction intermediates in the Spo0 pathway [reviewed in Galperin et al., 2012; Mattoo et al., 2008]. Thus, the presence of homologs of both Spo0F and Spo0B is strong evidence of a Spo0 pathway with a phosphorelay architecture. However, prediction of Spo0F and Spo0B homologs via sequence similarity methods has proven challenging due to the specific characteristics of Spo0F and Spo0B.

3.2.2.1. Challenges

Homologs are frequently identified based on significant sequence similarity or shared domain content. Unfortunately, due to the expansive and modular nature of histidine-aspartate phosphotransfer signal transduction systems results in many proteins in each genome that encode the TCS interaction domains, REC and HisKA [Galperin, 2006]. As a result, genes that encode histidine-aspartate signaling proteins are easily recognized, based on the presence of conserved interaction domains, but specific signaling families are difficult to distinguish from each other. Typical cognate two-component signaling system proteins are encoded adjacent to each other providing addition information for homology identification. This, for example, provides more sequence to compare between potential homologs. Further, if the TCS system has specific sensing and output domains, the pairing of such domains may be indicative of homology.

However, this strategy does not work for the Spo0 pathway, as all experimentally verified or annotated examples of Spo0F and Spo0B are encoded in dispersed regions of the genome compared

to Spo0A or any known sporulation kinases. For example, the distribution of Spo0 proteins of *B. subtilis*, depicted in Figure 4.3, are disperesed throughout the genome. Further, although Spo0F and Spo0B have different domain content than typical TCS proteins, this domain content is not sufficiently distinctive that it can be used as an identifier. In the following paragraphs, I detail the specific characteristics of Spo0F and Spo0B that make them difficult to identify using sequence or domain content similarity.

Spo0F contains only a REC domain (Response_reg, PFAM:PF00072), but lacks an output domain. This characteristic domain content does not distinguish Spo0F homologs from homologs of other stand-alone response regulators, such as CheY. Proteins matching this criteria are readily alignable, but in such alignments it is not clear which sequences are homologs of which protein as these sequences have similar characteristics.

Spo0B proteins, generally, have similar characteristics to histidine kinases [Mattoo et al., 2008]. Spo0B sequences lack strong sequence conservation, even within the same genus. Over broader evolutionary distances, sequence comparison cannot distinguish between Spo0B proteins and histidine kinases unambiguously [Mattoo et al., 2008]. For example, BLAST [Altschul et al., 1997] searches with *B. subtilis* Spo0B as a query retrieve both histidine kinase and Spo0B homologs. The retrieved sequences cannot be separated into two separate groups based on E-value or bit score. Clear Spo0B homologs are only identifiable in closely related species by this method.

In the current release of PFAM (version 31), two domains are assigned to *B. subtilis* Spo0B and its homologs: one corresponding to the N-terminal alpha-helical region (SPOB_a, PF:14689) and another corresponding to the C-terminal alpha-beta region (SPOB_ab, PF:14682). The SPOB_a domain is very similar to the HisKA domain and also contains a conserved, active histidine residue (at position 21 in the PFAM HMM profile). The SPOB_a domain is found throughout the Firmicutes, Proteobacteria, and Actinobacteria. Since endosporulation is a phenotype uniquely found in the Firmicutes, the presence of sequences annotated with SPOB_a in the latter two phyla suggests that this domain is not specific to Spo0B proteins. Further, within the firmicutes, protein sequences of histidine kinases not resembling the *B. subtilis* Spo0B are also assigned this domain by the PFAM HMM. These histidine kinases are likely the citrate/malate kinases, which are also identified in BLAST searches using Spo0B sequence queries [Mattoo et al., 2008]. The second domain, SPOB_ab, is only found in Class Bacilli Spo0B homologs. It is unclear whether this is due to Spo0B proteins only being present in Bacillar genomes or because the model is too specific to find Spo0B sequences in Clostridial genomes. Thus, PFAM domains associated with Spo0B are either too specific or too general to allow for definitive identication of homologs.

Earlier PFAM releases contained no domain models corresponding to Spo0B-like sequences.

PFAM version 26 was used to assess ORF domain content in MistDB version 2.2. Thus, in the data used in this study, lack of domain annotatation was a characteristic of Spo0B-like sequences.

3.2.2.2. Genome Neighborhood Conservation

As sequence similarity and domain content do not provide sufficient information to identify phosphorelay protein orthologs, I devised an alternative method for the identification of orthologs of Spo0F and Spo0B, based on genome context. As a basis for a comparative genomics approach, we considered the several dozen proteins from strains of *B. subtilis*, and their closest relatives that are annotated as Spo0F or Spo0B in the RefSeq database [O'Leary et al., 2016]. This guide set includes three experimentally verified Spo0F proteins [Bick et al., 2009; Burbulys et al., 1991; Malvar et al., 1994] and two experimentally verified Spo0B proteins [Burbulys et al., 1991; Mattoo et al., 2008]. These annotated Spo0F and Spo0B proteins were obtained through the MistDB website as described earlier in this section. We considered this set of "known" orthologs to Spo0F and Spo0B for use in identifying distantly related orthologs. Unfortunately, in both cases, the set of orthologs were not sufficient to create sequence-based classifiers for these proteins.

I hypothesized that the genome neighborhoods of a pair of homologs would be more likely to contain homologous genes than the genome neighborhoods of a pair of unrelated genes. To analyze the genome context of the guide set proteins, I first collected the four genes up- and downstream from each annotated protein *i.e.* the "genome neighborhood" of the annotated protein. Comparison of these genes for similarity was non-trivial, as this is preceded by a matching problem (i.e. which genes should be aligned between neighborhoods). If genes were annotated with protein family identifiers (e.g., RefSeq gene names), then these could be used to determine whether the same family was represented in both neighborhoods. However, many genes in bacterial genomes lack RefSeq identifiers. Another approach is to compare the domain content of the two neighborhoods. Domain content annotation can be performed using readily available tools, making it an ideal method to characterize a large sets of unknown proteins. Note that for the purpose of neighborhood comparison, it is not necessary to identify specific pairs of homologous flanking genes. General evidence of shared homology is sufficient.

Thus, to identify potential gene markers I analyzed the domain content of genes within the neighborhood of the proteins. For each of the two guide sets, I manually selected three genes, identified by domain content, that were frequently observed within the neighborhood of the annotated Spo0F or Spo0B, and were not abundant outside of the genome. I also favored genes encoded close to the annotated protein as this increases the likelihood that the protein will remain in that neighborhood, even in more distantly related species. Each of these genes (Table 3.1) had a

characteristic domain which could be used in searching with the MistDB interface.

Table 3.1: Spo0F Genome Neighborhood Marker Genes

Gene Name	Search Domain	PFAM ID	Typical Location
Fructose Bisphosphate Aldolase	F_bp_aldolase	PF01116	1 gene downstream
Transaldolase	Transaldolase	PF00923	2 genes downstream
CTP Synthase	CTP_Synth_N	PF06418	1 gene upstream

I next used these marker genes to search all genomes in the representative set. For each genome, the MiSTDB interface was used to search for marker genes using the characteristic domain (Table 3.1, Table A6). The genome neighborhood for all genes encoding the characteristic domains were collected. If two or more marker genes were within the same genome neighborhood they were merged. Each of these potential Spo0F neighborhoods was searched for proteins matching the Spo0F criteria: a protein encoding only a single REC domain, taking up 90% or more of the total protein (as measured by amino acid coverage).

This procedure revealed that all but two spore-forming Class Bacilli genomes investigated contain a putative Spo0F (Fig 3.2, light blue octagons). I also identified candidate Spo0F genes in 18 spore-forming genomes within Class Clostridia in the representative set. No genome encodes more than one predicted Spo0F. In contrast, 13 spore-forming Class Clostridia genomes do not encode a Spo0F-like gene in the vicinity of any of the Spo0F neighborhood markers; nor do they encode any two of the neighborhood markers in close proximity to each other. In particular, no candidate Spo0F proteins were found in species in which a direct phosphorylation architecture has been experimentally verified (*C. acetobutylicum* [Steiner et al., 2011], *R. thermocellum* [Mearls and Lynd, 2014], and *C. difficile* [Underwood et al., 2009]). These results suggest that Spo0F homologs can be identified by conserved genome neighborhoods and that they are found not only in Class Bacilli, but also in many taxa in Class Clostridia.

A similar procedure was used to identify candidate Spo0B orthologs. Analysis of the neighborhoods of the Spo0B guide set proteins resulted in the selection of three marker genes for the Spo0B neighborhood (Table 3.2, A7). Again, I used these marker genes to search all genomes in the representative set for the presence of marker genes. Marker gene genome neighborhoods were collected and analyzed for the presence of Spo0B-like proteins. For retrieval from MistDB version 2.2 (which uses PFAM version 26), candidate Spo0B proteins fit the criteria of no assigned PFAM domains and sequence similarity to guide set Spo0B proteins within the the first 50 amino acids. This is the most conserved region in Spo0B-like proteins. The Spo0B marker genes were typically singleton in all guide set genomes and were encoded in the same window of four ORFs. All annotated Spo0Bs were encoded between the ribosomal protein L21 and the ObgE-like GTPase.

Table 3.2: Spo0B Genome Neighborhood Marker Genes

Gene Name	Search Domain	PFAM ID	Typical Location
ObgE-like GTPase	GTP1_OBG	PF01018	1 gene upstreamstream
Ribosomal protein L21	Ribosomal_L21p	PF00829	1 genes downstream
Ribosomal protein L27	Ribosomal_L27	PF1016	3 gene upstream

In our set of representative Firmicutes, 75 genomes encode the trio of Spo0B marker genes in close proximity (Fig 3.2, Table A7). All but two spore-formers in Class Bacilli were found to encode a Spo0B-like protein within a five gene window that includes all three marker genes. Genomes of 18 spore-formers within the Class Clostridia also had a region containing the three marker genes and a candidate Spo0B ortholog. With the release of the latest version of PFAM (version 31), Spo0B can also be identified as proteins that encode SPOB_a and SPOB_ab or SPOB_a and no other domains, and all identified Spo0Bs meet this criteria. The remaining Class Clostridia genomes encoded the three Spo0B neighborhood markers in close proximity, but did not encode a protein meeting the criteria of Spo0B in that vicinity. No Class Clostridia species that has been experimentally verified to directly phosphorylate Spo0A encodes a candidate Spo0B. Thus, Spo0B homologs can also be identified by conservation of genome neighborhood and are found in almost all genomes in which a Spo0F homolog was identified.

3.2.2.3. Discussion of Spo0F and Spo0B predictions

The genome neighborhood conservation method allowed the successful prediction of several unknown orthologs to both Spo0F and Spo0B, especially within Class Clostridia. These results conflict with the standing hypothesis, which suggests that phosphorelays, and therefore phosphorelay proteins, are restricted to Class Bacilli.

In nearly all cases, regardless of whether a genome was a spore-former, if Spo0F is predicted then Spo0B is also predicted, and vice versa. This striking result is further suggestive of the hypothesis that these proteins participate in the same function, as observed in *B. subtilis*, in all genomes where both were identified.

There are three cases where Spo0F and Spo0B do not co-occur:

1. *Thermincola potens* is closely related to a number of spore-formers which encode both Spo0F and Spo0B and has been observed to sporulate in some cases [Sokolova et al., 2004]. Further, the identified Spo0F is encoded in the canonical Spo0F neighborhood containing all three marker domains and shares strong sequence similarity with other Spo0F proteins identified in closely related genomes. Thus, the absence of Spo0B in this genome may be a

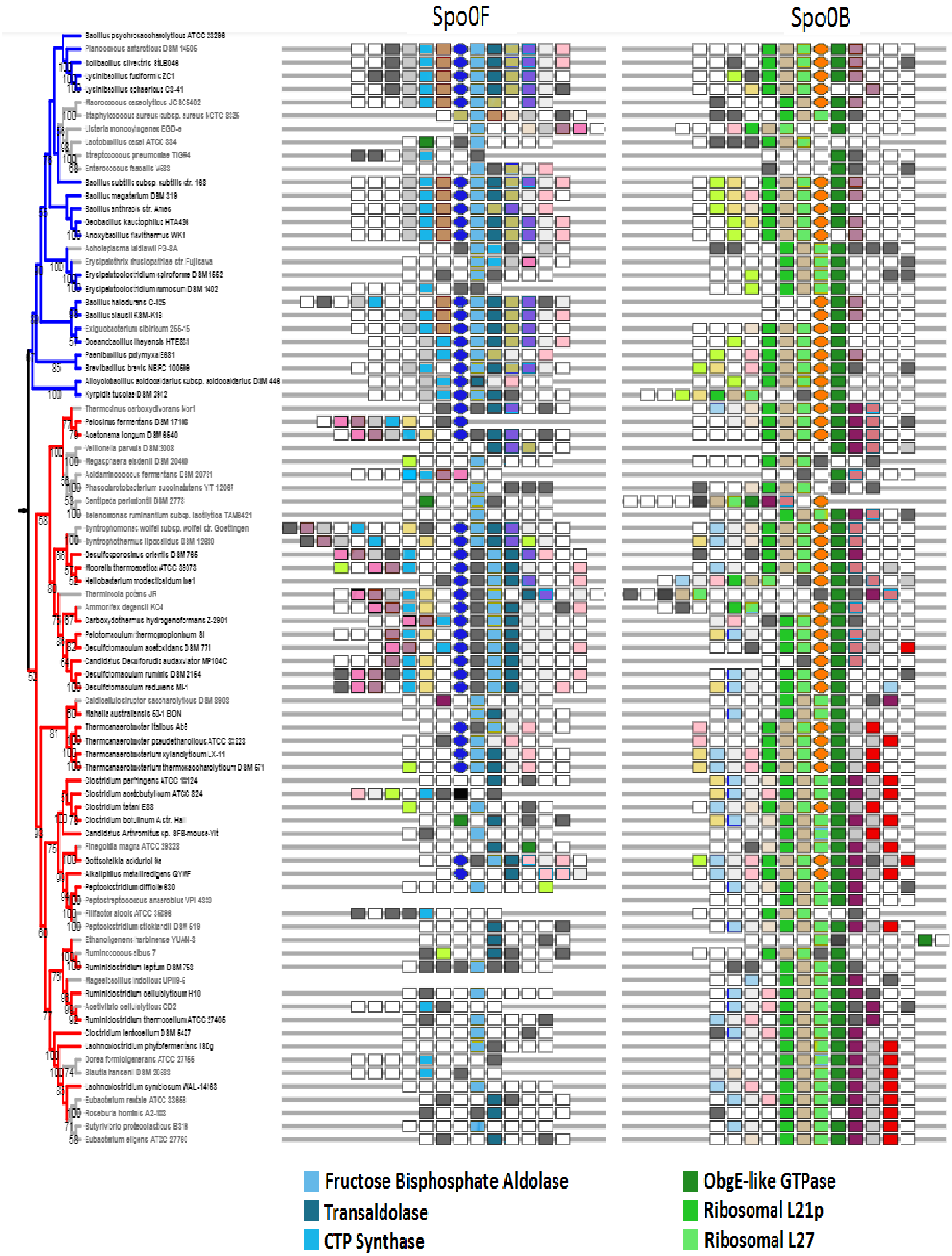

Spo0F
Spo0B
Fructose Bisphosphate Aldolase
Transaldolase
CTP Synthase
ObgE-like GTPase
Ribosomal L21p
Ribosomal L27

Figure 3.2: Extended genome neighborhoods prepared as detailed above. Top 20 most frequent genes (as determined by domain content) shown in color. Genes lacking domains are represented by light gray boxes; all other genes are represented by white boxes. General and lineage-specific conservation is extends beyond the marker genes in the neighborhoods of both phosphorelay proteins. Predicted Spo0F proteins are centered, blue circles in the first column. Predicted Spo0B proteins are centered, orange circles in the second column. Strikingly, the Spo0B neighborhood is conserved even in species that do not encode Spo0B.

sequencing error or a variation restricted to the strain that was sequenced. (The region that typically encodes Spo0B, between the ObgE-like and L27 ribosomal protein gene is occupied by several spore germination proteins including GerA. This may be the result of transfer or genome re-arrangement and warrants further investigation.

2. The predicted Spo0B in *Centipeda periodontii* was identified within the genome neighborhood of a gene harboring a GTP1_OBG domain and aligns well to the active site of annotated Spo0Bs. However, unlike other predicted Spo0B proteins, it is not encoded between the ObgE-like protein and the L27 ribosomal protein. Further, this protein is not annotated by PFAM as SPOB_A. I hypothesize that this protein may be a degenerate Spo0B-coding region or coincidentally encodes a region that is alignable with Spo0B.

3. The strain of *Clostridium tetanus* used in this study is a non-sporulating variant and was derived in a labratory setting [Galperin et al., 2012]. These factors suggest that the presence of Spo0B may be specific to this strain and may be not be typical of the *C. tetanus* genomes.

3.2.3. Identification of Potential Sporulation Kinases based on Genome Context

A different approach was required to identify potential sporulation kinases, as analysis of the regions flanking known sporulation kinases did not reveal any conservation of the genomic neighborhood. Further, experimentally verified Spo0 kinases possess few shared sequence features that definitively distinguish sporulation kinases from other sensor histidine kinases. Sequence similarity and strict domaint content methods also present mixed results, as the HisKA and HK_CA domains of known sporulation kinases are not markedly more similar to each other than to those encoded by other sensor kinases, and the N-terminal sensor regions of *bonafide* sporulation kinases vary substantially in both sequence and domain content.

However, all experimentally verified sporulation kinases (Table A2) are orphans, i.e., are not

found adjacent to genes encoding any other two-component signaling pathway proteins two-component signaling pathway proteins. Moreover, N-terminal PAS domains are observed more frequently in Spo0 kinases than in the set of all kinases in the same species. For example, three out of the five experimentally verified kinases of *B. subtilis* and one of the three in *C. acetobutylicum* contain at least one PAS domain. This is compared to 24% and 11% overall, respectively, including all kinases in each genome. Thus, orphan status, combined with the presence of an N-terminal PAS domain, furnishes a signature for predicting candidate Spo0 kinases. All 48 spore-formers in our data set (with the exception of one *Erysipeloclostridium*) were found to encode at least one orphan kinase and 41 of those have at least one orphan kinase with an N-terminal PAS domain.

3.2.4. Results of Survey and Distribution of Spo0 Components

Thus far in this chapter, I have presented the identification of putative homologs of Spo0 pathway components of *B. subtilis* and other experimentally verified Spo0 pathways in the representative set of Firmicutes. Specifically, I predicted Spo0A in all spore-formers, consistent with the conserved role of Spo0A as the master regulator of sporulation. Further, I predicted 33 pairs of Spo0F and Spo0B orthologs in spore-formers, including most Class Bacilli genomes and a distributed set of Class Clostridia genomes. The remaining 15 spore-former genomes do not encode either Spo0F or Spo0B. As a guide to the results of this survey, I have prepared Figure 3.3, depicting Spo0 component and sporulation kinase predictions on the species tree.

Considering the sequence similarity and similarity of domain content to the experimentally verified examples, it is unlikely that these putative Spo0 proteins are performing any alternative function. Thus, as the only known function of Spo0F and Spo0B proteins is phosphotransfer within the Spo0 phosphorelay, this is also the likely function of the predicted phosphorelay proteins. This is corroborated by the consistency of the predictions across the complete data set: almost all species either encode both Spo0F and Spo0B or encode neither. Further, orphan kinases with potential to be involved in the Spo0 pathway were identified in all, but one, spore-former.

Therefore, I hypothesize that the presence of Spo0F and Spo0B, in conjunction with at least one potential sporulation kinase and a putative Spo0A, is indicative of a functional phosphorelay architecture; the lack of predicted Spo0F and Spo0B proteins in a spore-former genome is indicative of an architecture where Spo0A is directly phosphorylated. Alternative hypotheses include the scenario where these proteins have been maintained but act as phosphorelay intermediates in a different pathway, are non-functional, or are not the only means of phosphorylating Spo0A. In these cases, some set of spore-former genomes would encode Spo0F and Spo0B, but still signal the initiation of sporulation via direct phosphorylation of Spo0A, or be capable of signaling sporu-

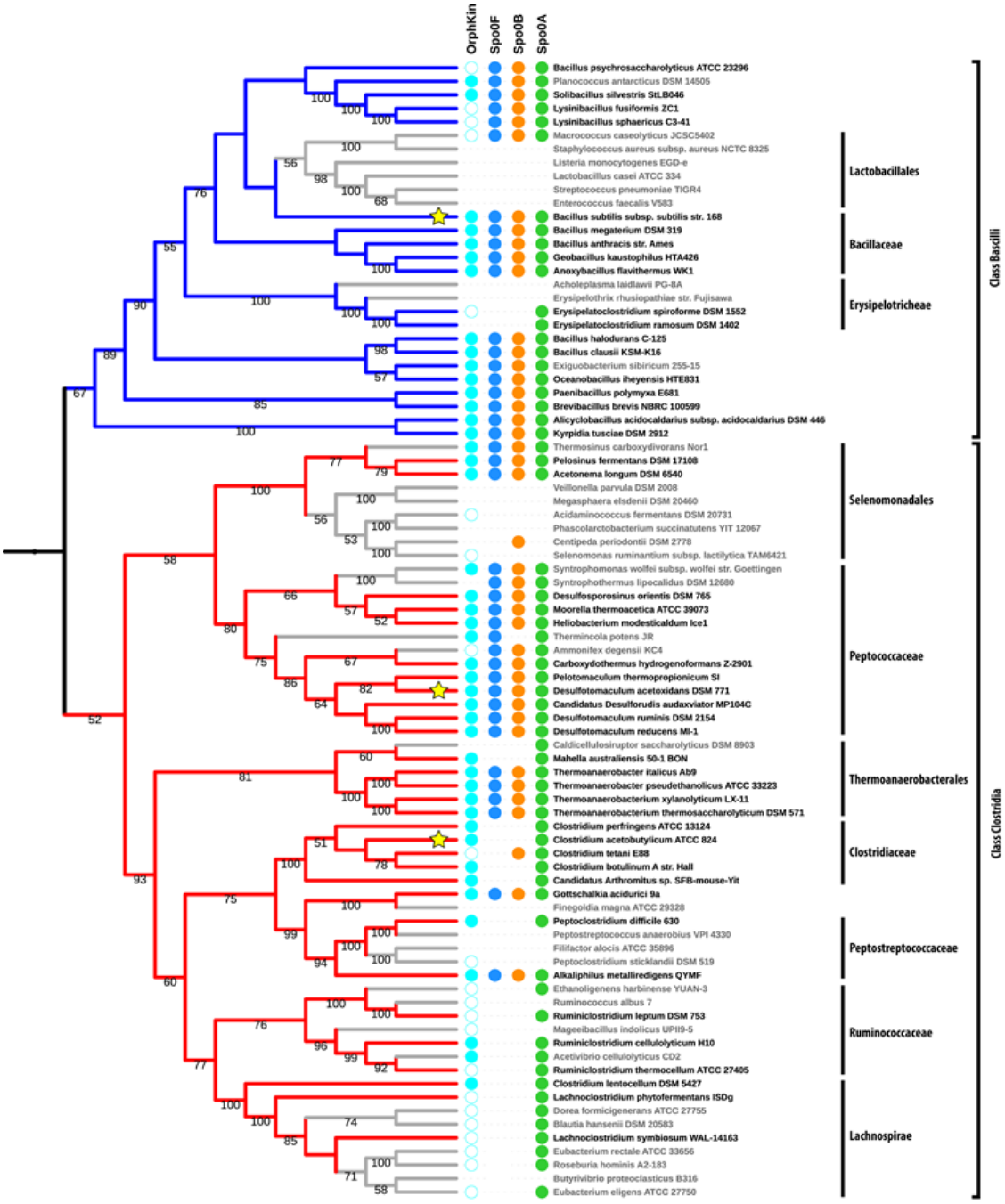

OrphKin
Spo0F
Spo0B
Spo0A
Bacillus psychrosaccharolyticus ATCC 23296
Planococcus antarcticus DSM 14505
Solibacillus silvestris StLB046
Lysinibacillus fusiformis ZC1
Lysinibacillus sphaericus C3-41
Macrococcus caseolyticus JCSC5402
Staphylococcus aureus subsp. aureus NCTC 8325
Listeria monocytogenes EGD-e
Lactobacillus casei ATCC 334
Streptococcus pneumoniae TIGR4
Enterococcus faecalis V583
Bacillus subtilis subsp. subtilis str. 168
Bacillus megaterium DSM 319
Bacillus anthracis str. Ames
Geobacillus kaustophilus HTA426
Anoxybacillus flavithermus WK1
Acholeplasma laidlawii PG-8A
Erysipelothrix rhusiopathiae str. Fujisawa
Erysipelatoclostridium spiroforme DSM 1552
Erysipelatoclostridium ramosum DSM 1402
Bacillus halodurans C-125
Bacillus clausii KSM-K16
Exiguobacterium sibiricum 255-15
Oceanobacillus iheyensis HTE831
Paenibacillus polymyxa E681
Brevibacillus brevis NBRC 100599
Alicyclobacillus acidocaldarius subsp. acidocaldarius DSM 446
Kyrpidia tusciae DSM 2912
Thermosinus carboxydivorans Nor1
Pelosinus fermentans DSM 17108
Acetonema longum DSM 6540
Veillonella parvula DSM 2008
Megasphaera elsdenii DSM 20460
Acidaminococcus fermentans DSM 20731
Phascolarctobacterium succinatutens YIT 12067
Centipeda periodontii DSM 2778
Selenomonas ruminantium subsp. lactilytica TAM6421
Syntrophomonas wolfei subsp. wolfei str. Goettingen
Syntrophothermus lipocalidus DSM 12680
Desulfosporosinus orientis DSM 765
Moorella thermoacetica ATCC 39073
Heliobacterium modesticaldum Ice1
Thermincola potens JR
Ammonifex degensii KC4
Carboxydothermus hydrogenoformans Z-2901
Pelotomaculum thermopropionicum SI
Desulfotomaculum acetoxidans DSM 771
Candidatus Desulforudis audaxviator MP104C
Desulfotomaculum ruminis DSM 2154
Desulfotomaculum reductens MI-1
Caldicellulosiruptor saccharolyticus DSM 8903
Mahella australiensis 50-1 BON
Thermoanaerobacter italicus Ab9
Thermoanaerobacter pseudethanolicus ATCC 33223
Thermoanaerobacterium xylanolyticum LX-11
Thermoanaerobacterium thermosaccharolyticum DSM 571
Clostridium perfringens ATCC 13124
Clostridium acetobutylicum ATCC 824
Clostridium tetani E88
Clostridium botulinum A str. Hall
Candidatus Arthromitus sp. SFB-mouse-Yit
Gottschalkia acidurici 9a
Finegoldia magna ATCC 29328
Peptoclostridium difficile 630
Peptostreptococcus anaerobius VPI 4330
Filifactor alocis ATCC 35896
Peptoclostridium sticklandii DSM 519
Alkaliphilus metalliredigens QYMF
Ethanoligenens harbinense YUAN-3
Ruminococcus albus 7
Ruminiclostridium leptum DSM 753
Mageeibacillus indolicus UPII9-5
Ruminiclostridium cellulolyticum H10
Acetivibrio cellulolyticus CD2
Ruminiclostridium thermocellum ATCC 27405
Clostridium lentocellum DSM 5427
Lachnoclostridium phytofermentans ISDg
Dorea formicigenerans ATCC 27755
Blautia hansenii DSM 20583
Lachnoclostridium symbiosum WAL-14163
Eubacterium rectale ATCC 33656
Roseburia hominis A2-183
Butyrivibrio proteoclasticus B316
Eubacterium eligens ATCC 27750
Lactobacillales
Bacillaceae
Erysipelotricheae
Selenomonadales
Peptococcaceae
Thermoanaerobacterales
Clostridiaceae
Peptostreptococcaceae
Ruminococcaceae
Lachnospirae
Class Bacilli
Class Clostridia

Figure 3.3: Cladogram of representative Firmicutes constructed with 50 concatenated ribosomal protein sequences using RAxML with the CAT model and 100 bootstrap replicates (bootstrap support greater than 50 labeled on branches). Rooted by outgroup, not shown. Colored branches indicate known sporulating species of Class Bacilli (blue) and Class Clostridia (red). Species where no spores have been reported are shown in grey. Colored dots indicate Spo0 pathway components predicted in this study: one or more orphan HKs (cyan); Spo0F (blue); Spo0B (orange); Spo0A (green). Filled cyan dot indicates at least one orphan HK encodes a PAS domain. For number of orphan kinases in each genome, see Table A9. Genomic origin of proteins used in *in vitro* phosphotransfer assays (Figs 3.6, 4.2, 4.9) are starred.

lation directly as well as through the phosphorelay. The following section (Section 3.3) provides evidence towards these hypotheses using experimental and *in silico* approaches.

Finally, I analyze the present-day distribution of predicted pathway architectures that will facilitate the development of an evolutionary history of the Spo0 pathway (Section 3.4).

3.3. Predicted Spo0 Components are likely functional

In this section, I present results examining the hypothesis that the presence of Spo0F and Spo0B, with Spo0A and a potential sporulation kinase, indicates sporulation signaling via a phosphorelay and the absence of Spo0F and Spo0B indicates sporulation signaling via direct phosphorylation of Spo0A.

Here, I present two strategies that provide evidence for the interaction pattern of the predicted Spo0 components. First, I examine specificity residues for similarity to the set of experimentally verified examples of Spo0 proteins (Section 3.3.1). Next, noting that there have been no experimentally verified Spo0 phosphorelays outside of Class Bacilli, I performed a phosphotransfer profiling experiment using the predicted Spo0 proteins of a member of Class Clostridia (Section 3.3.2).

3.3.1. Specificity Residue Signatures are Similar to Experimentally Verified Spo0 Components

Inspection of the specificity residues encoded in candidate Spo0 components supports the accuracy of our predictions. Specificity residues were inferred for all predicted Spo0 proteins and all orphan kinases encoded by spore-formers within the representative set (Tables A8 and A9). Specificity residues for candidate Spo0B homologs and orphan kinases were identified by manual alignment of relevant sites within SPOB_a or HisKA domains to the HisKA domains of proteins with known

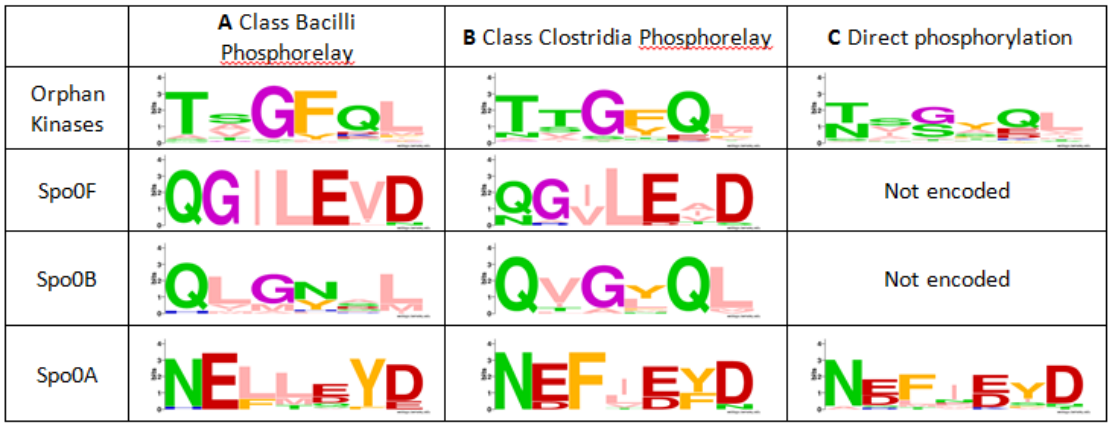
specificity residues [Skerker et al., 2008]. Kinases and candidate Spo0B were aligned by the conserved phosphorylated Histidine residues of EnvZ, RstB, CpxA, and TM0853. The columns aligned with those identified as HK-RR covarying residues [Capra et al., 2010] were predicted to be the specificity residues for the kinases and candidate Spo0 components. The specificity residues of the SPOB_a domain are likely equivalent to those of HisKA because the Spo0F-Spo0B interaction was instrumental in uncovering interfacial contact residues [Zapf et al., 2000] and have since typified the residues that co-vary between REC and HisKA domains [Skerker et al., 2008] (see also PDB:1F51).

Similarly, specificity residues for candidate Spo0F and Spo0A were determined by manual alignment to REC domains of proteins with known specificity residues [Capra et al., 2010]. The REC domains of the candidate Spo0Fs and Spo0As were aligned with the RR domain of OmpR, RstA, CpxR, and TM0468. Again, the columns aligned with those identified as HK-RR covarying residues [Capra et al., 2010] were predicted to be the specificity residues for the kinases and candidate Spo0 components (Table A8).

To permit further analysis of these specificity residues, I created specificity residue logos for each set of predicted Spo0 components, further subdividing these sets by predicted architecture and, within the phosphorelay architecture, by taxonomic class (Fig 3.4). The specificity signatures of predicted phosphorelay proteins (Fig 3.4A and 3.4B, Table A8 and A9) are markedly similar to specificity residues in experimentally verified phosphorelay proteins (Figure 3.4D and Tables A2—A5). Given that similarity in specificity residues correlates with phosphotransfer capability in vitro [Capra et al., 2010; Podgornaia and Laub, 2015; Skerker et al., 2008], this similarity further supports the hypothesis that the predicted phosphorelays interact similarly to canonical phosphorelays.

3.3.2. Phosphotransfer Profiling of a Clostridial Phosphorelay, *Desulfotomaculum acetoxidans*

The presence of putative Spo0F and Spo0B proteins in some spore-forming Class Clostridia species suggests that these organisms may, like those in Class Bacilli, signal the initiation of sporulation through a phosphorelay architecture. To determine whether the Spo0 proteins predicted by our method do, in fact, participate in a phosphorelay, I sought to test the phosphotransfer properties [Laub and Goulian, 2007b] of the putative phosphorelay proteins from *Desulfotomaculum acetoxidans* DSM771 (Class Clostridia, starred in Fig 3.3), a spore-forming species in Peptococcaceae [Widdel and Pfennig, 1977]. The predicted homologs of Spo0F and Spo0B in this genome have conserved genomic neighborhoods. Comparison of the predicted Spo0 proteins in *D. acetoxidans*



D

Phosphorelay	Direct phosphorylation
T S G F Q L	N V S A Q V
T S G F K L	N I S A Q L
T V G F Q L	S V G L Q L
A S G F Q L	N L S S Q V
T A G F Q L	N V S S Q L
T G G F Q L	A S G L Q L
T A G F Q L	N I S T Q I
T S G F Q I	S V G A Q L
T V G F Q L	
T A G F Q L	
Q V G F Q L	
T A G F Q L	
T S G F Q L	

Figure 3.4: A, B, C) Sequence logos for predicted specificity residues of orphan kinases, Spo0F, Spo0B, and Spo0A. Phosphorelay architecture subdivided by taxonomic class. All but two predicted direct phosphorylation architecture are encoded in Class Clostridia. Created using WebLogo [Crooks et al., 2004]. (The Spo0F sequence in Solibacillus is truncated and was not included.) D) Specificity residues of experimentally verified sporulation kinases separated by architecture (see also Tables A2—A5).

with their *B. subtilis* counterparts indicated a high degree of similarity in their respective specificity residues (Table 3.3).

For phosphotransfer profiling, I also chose a potential sporulation kinase to test. For the purposes of this experiment, I aimed to select one that was the most likely to interact in the phosphorelay. *D. acetoxidans* has seven orphan kinases, six of which encode a PAS domain. Each has putative specificity residues (Table 3.3) similar to verified Bacillus sporulation kinases suggesting that they may target Spo0F. Of these six kinases, Dtox_1918 was chosen for the phosphotransfer experiments as it has specificity residues differing from *B. subtilis* KinA at only one position.

Table 3.3: Specificity residues of *D. acetoxidans* predicted Spo0 components

Locus	PAS	Orphan Kinase	Spo0F	Spo0B	Spo0A
Dtox_0091	Yes	TTGFQM			
Dtox_1564		TAAFEL			
Dtox_1918	Yes	TTGFQL			
Dtox_2569	Yes	TTGFQM	QGILEVD	QVGLQL	NEFLDFD
Dtox_3081	Yes	TTGFQF			
Dtox_3426	Yes	TTGFQL			
Dtox_3834	Yes	TTGFQM			

3.3.2.1. Phosphotransfer Profiling Method

Phosphotransfer profiling reactions were based on the protocols outlined in [Laub and Goulian, 2007a].

Autophosphorylation was performed at a final concentration of approximately 5uM kinase in HKEDG buffer (10 mM HEPES-KOH, pH8.0, 50mM KCl, 10% glycerol, 0.1 mM EDTA, 2mM DTT) supplemented with 5 mM MgCl2, 500 uM ATP, and 0.5 uCi [32P]-ATP from a stock at 6000Ci/mmol (Perkin Elmer). The results from the autophosphorylation experiments 3.5 demonstrated that 15 minutes was an appropriate amount of time for Dtox_1918 to autophosphorylate to peak levels. Dt_Spo0B was also subjected to these conditions but no autophosphorylation was observed.

For phosphotransfer profiling of the candidate phosphorelay, Dtox_1918 was allowed to autophosphorylate for 15 minutes before addition of another putative phosphorelay component. Each additional component was added at a low volume and high concentration to maintain the concentration of kinase and other components which had been added. In Fig 3, a solution containing 2uM autophosphorylated Dtox_1918 was split to accommodate different sequences of component addition. Sequence one was addition of Spo0F, Spo0B, and Spo0A at 4 minute intervals. Sequence two was addition of Spo0B and Spo0A at 4 minute intervals. Sequence three was addition of Spo0A. After the addition of Spo0A in each sequence a sample was taken after 10 minutes to observe the equilibrium phosphorylation state for each sequence. Spo0F and Spo0B were added at a concentration of 6 uM while Spo0A was added at 10 uM. Samples were taken 3 minutes after the addition of each additional component and after the 10 minute incubation post-addition of Spo0A. The reaction for each sample was stopped by the addition of 2X Novex LDS Loading buffer (Life Technologies) and analyzed by 12% SDS-Page gel and phosphorimaging.

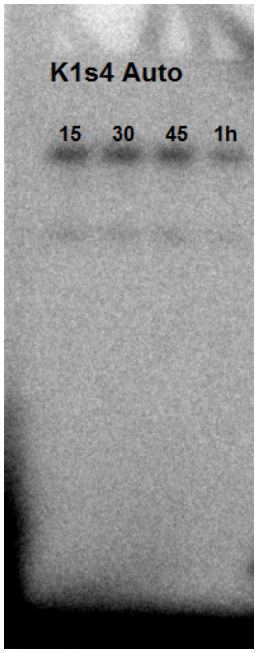


Figure 3.5: Dtox_1918 was incubated with 10 mM HEPES-KOH, pH8.0, 50mM KCl, 10% glycerol, 0.1 mM EDTA, 2mM DTT, 5 mM MgCl2, 500 uM ATP, and 0.5 uCi [32P]-ATP from a stock at 6000Ci/mmol for 1 hour. Peak autophosphorylation observed after 15 minutes.

3.3.2.2. Predicted Spo0 components of *D. acetoxidans* act as a phosphorelay

The similarity of the *D. acetoxidans* Spo0 components to known Spo0 phosphorelays suggested that phosphotransfer to its Spo0A will only be observed in the presence of a sporulation kinase, Spo0F, and Spo0B. To test this hypothesis, I purified His6-tagged variants of each of the four components. The kinase, Dtox_1918, was prepared without the N-terminal sensor region, and Spo0A (Dtox_2041) without the Spo0A_C output domain; Full length D. acetoxidans Spo0F (Dtox_0055) and Spo0B (Dtox_3313) were used.

The kinase, Dtox_1918, was first incubated alone with radiolabeled ATP for 15 min and then examined by SDS-PAGE and autoradiography (Fig 3.6A, lane 1). Two bands likely representing autophosphorylated Dtox_1918 were seen; the faster migrating band likely corresponds to monomeric Dtox_1918 with the slower migrating band reflecting dimeric Dtox_1918 that was not fully dissociated during SDS-PAGE analysis.

Inclusion of Spo0F in the reaction (Fig 3.6A, lane 2) produced an additional band indicating that Spo0F can be directly phosphorylated by Dtox_1918. Similarly, the addition of Spo0F and Spo0B to autophosphorylated Dtox_1918 produced bands for each protein (Fig 3.6A, lane 3). Importantly, the phosphorylation of Spo0B depended on the presence of Spo0F (Fig 3.6A, lane 6).

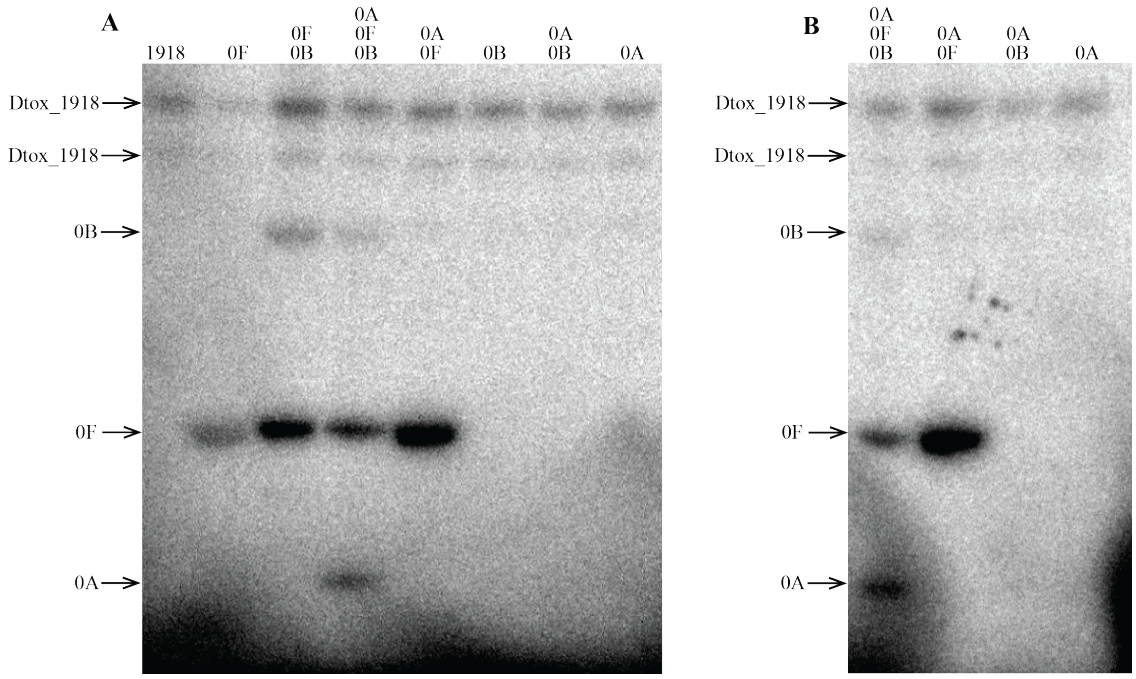


Figure 3.6: Phosphotransfer analysis of putative *D. acetoxidans* Spo0 Components A) Reactions contained 5uM Dtox_1918 and 10uM each of the proteins listed above each lane in the phosphorimage. Incubation for 3 minutes after adding final component. B) Reactions with Spo0A as in A, allowed to incubate for 10 minutes after adding Spo0A. Abbreviations: 0F, Dtox_0055; 0B, Dtox_3313; 0A, Dtox_2041.

Addition of Spo0F, Spo0B, and Spo0A to autophosphorylated Dtox_1918 produced bands corresponding to each protein (Fig 3.6A, lane 4). The phosphorylation of Spo0B and Spo0A depended on the presence of SpoF (Fig 3.6A, lanes 7) and the phosphorylation of Spo0A also depended on the inclusion of Spo0B (Fig 3.6A, lane 5).

Finally, we confirmed that Dtox_1918 cannot, under these conditions or at longer time point (Fig 3.6B), directly phosphorylate Spo0A (Fig 3.6A, lane 8). Collectively, these results indicate that Dtox_1918, along with the candidate Spo0F, Spo0B, and Spo0A homologs, comprise a *bonafide* phosphorelay, similar in architecture to that first characterized in *B. subtilis*. The *D. acetoxidans* phosphorelay is, to my knowledge, the first experimentally verified Spo0 phosphorelay outside of Class Bacilli.

3.4. Distribution of Spo0 Pathway Architectures

Examination of the phylogenetic distribution of the predicted architectures (Fig 3.3, summarized in Fig 3.7) reveals patchiness within both taxonomic classes of the Firmicutes phylum. That is, I

find that neither predicted architecture is monophyletic within either class. Within Class Bacilli, almost all spore-formers are predicted to encode a phosphorelay architecture with the exception of two genomes of the genus *Erysipelatoclostridium*. The distribution of architectures within Class Clostridia is noticeably patchier than Class Bacilli. All but one spore-former (*Mahella australiensis*) in the Peptococcaceae and Thermoanaerobacterales is predicted to encode a phosphorelay. The remaining Clostridia are predicted to encode DPAs with the exception of two relatively recently branching species within the Clostridiaceae, *Gottschalkia acidurici* and *Alkaliphilus metalliredigens*.

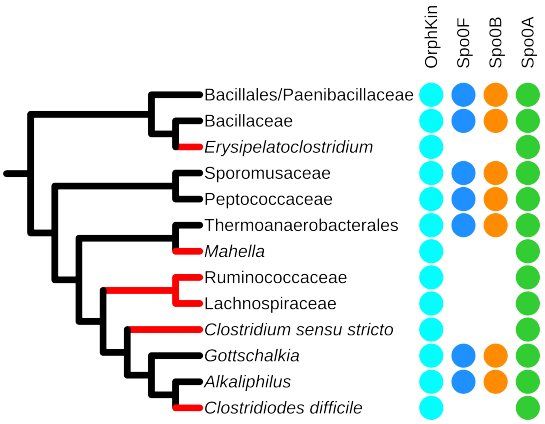


Figure 3.7: Summary of Spo0 architectures in spore-forming taxa; Clades selected such that all members encode the same predicted Spo0 components; the branch labelled Paenibacillaceae/Bacillales consists of four paraphyletic clades.

Collectively, the 48 spore-formers form 15 clades with the same architecture; five of these groups are predicted to encode a direct phosphorylation architecture, while the remaining ten are predicted to encode a phosphorelay. Parsimony suggests that it is likely that the common ancestor of each of these monophyletic clades also encode the same architecture as its constituents. This interleaved distribution of predicted Spo0 architectures is consistent with multiple changes in pathway architecture over the course of evolution.

To verify that the patchy distribution observed in our species tree is not an artifact of the selection of genomes in our study or the phylogenetic methods used, I took advantage of two recently published, alternative Firmicutes phylogenies [Antunes et al., 2016; Yutin and Galperin, 2013], one of which is based on a much larger set of Firmicutes genomes [Antunes et al., 2016].

3.4.1. Distribution of Spo0 Components in Alternative Phylogenies

I scanned genomes in the Antunes et al. [2016] and Yutin and Galperin [2013] data sets for orthologs of Spo0 pathway components, using the same procedure described in Chapter 3. The Spo0 architectures predicted in spore-forming genomes in both sets have patchy phylogenetic distributions (Figs 3.8 and 3.9), similar to that observed in our original analysis (Fig 3.3). In particular, in both data sets, I found phosphorelay architectures in both major taxonomic classes. The patchiness observed in all three trees is consistent with multiple changes in pathway architecture over the course of evolution.

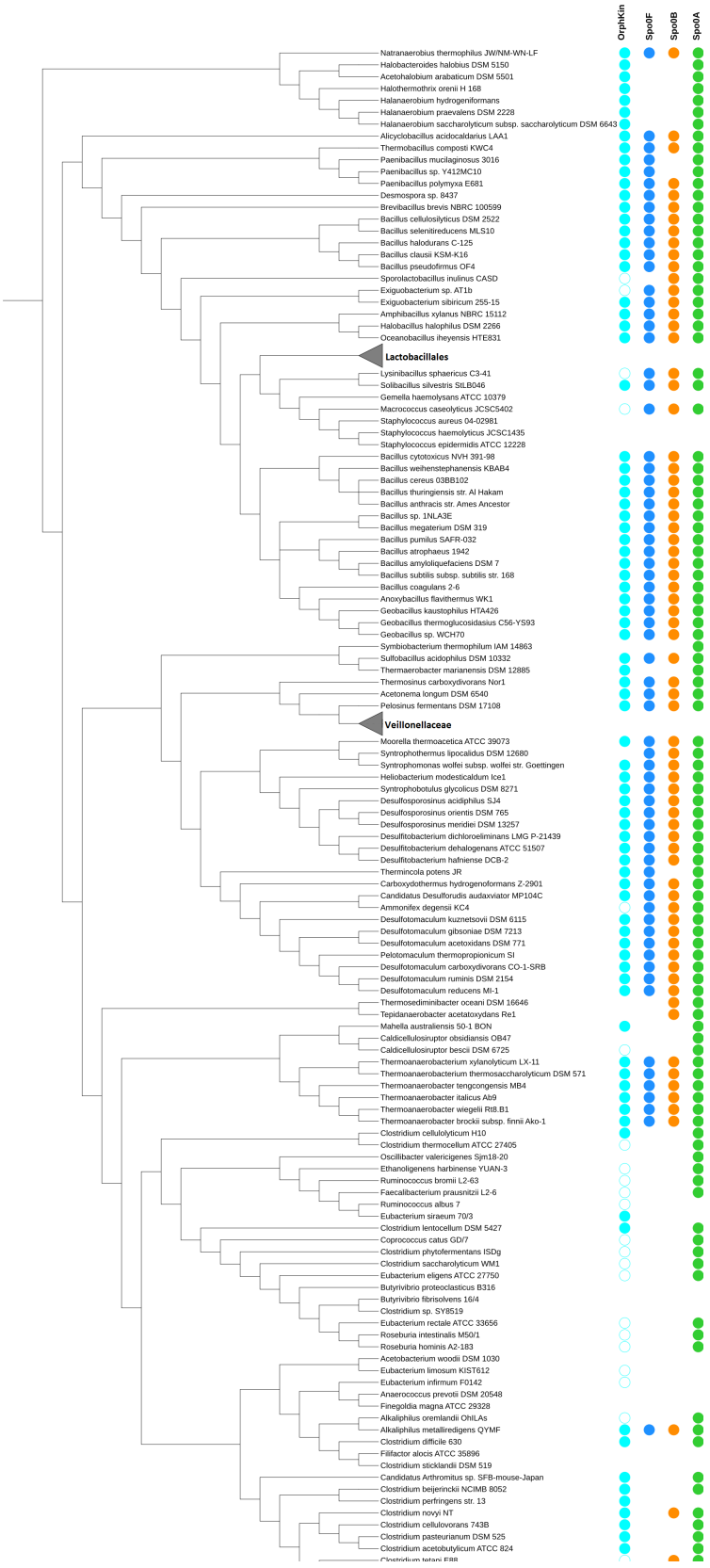
At the highest taxonomic ranks, the Yutin and Antunes tree topologies are similar to each other and to our tree (Fig 3.3), despite differences in pre-processing, tree reconstruction methodology, and taxon sampling. There are minor differences in the inferred relationships lower in the taxonomy, especially with respect to branches with weak sequence support. The remainder of this section examines the positions of subgroups that differ across the three trees. In each case, I discuss the implications of those differences for our hypotheses regarding the ancestral pathway or patchy distribution.

3.4.1.1. Halanaerobiales and Natranaerobiales

The Halanaerobiales and Natranaerobiales are anaerobic, halophilic extremophiles [Mesbah et al., 2007; Roush et al., 2014]. The Antunes data set includes all currently available, fully sequenced genomes from these taxa; i.e., the genomes of *Natranaerobius thermophilus* and five members of the Halanaerobiales. The Yutin data set includes *N. thermophilus* and one Halanaerobiales genome. No genomes from these taxa were included in our data set.

Analysis of these genomes for Spo0 components revealed putative Spo0B and Spo0F orthologs in the only available Natranaerobiales genome (*Natranaerobius thermophilus*), which is consistent with a phosphorelay, although this species is a reported non-spore former [Mesbah et al., 2007]. One spore-forming member of the Natranaerobiales order has been reported, *Natranaerobaculum magdiense* [Zavarzina et al., 2013], but a whole genome sequence for this species has not been published.

All Halanaerobiales analyzed encode Spo0A and at least one PAS-containing histidine kinase, but neither Spo0F, nor Spo0B. Again, all of the Halanaerobiales species included in either the Antunes or Yutin tree are considered to be asporogenous [Mavromatis et al., 2009; Oren et al., 1991; Sikorski et al., 2010; Vos et al., 2009; Zhilina et al., 1996]. However, several species within the order Halanaerobiales have been reported to form spores, including *Halonatronum saccharophilum*


OrphKin
Spo0F
Spo0B
Spo0A
Natranaerobius thermophilus JW/NM-WN-LF
Halobacteroides halobius DSM 5150
Acetohalobium arabaticum DSM 5501
Halothermothrix orenii H 168
Halanaerobium hydrogeniformans
Halanaerobium praevalens DSM 2228
Halanaerobium saccharolyticum subsp. saccharolyticum DSM 6643
Alicyclobacillus acidocaldarius LAA1
Thermobacillus composti KWC4
Paenibacillus mucilaginosus 3016
Paenibacillus sp. Y412MC10
Paenibacillus polymyxa E681
Desmospora sp. 8437
Brevibacillus brevis NBRC 100599
Bacillus cellulosilyticus DSM 2522
Bacillus selenitireducens MLS10
Bacillus halodurans C-125
Bacillus clausii KSM-K16
Bacillus pseudofirmus OF4
Sporolactobacillus inulinus CASD
Exiguobacterium sp. AT1b
Exiguobacterium sibiricum 255-15
Amphibacillus xylanus NBRC 15112
Halobacillus halophilus DSM 2266
Oceanobacillus iheyensis HTE831
Lactobacillales
Lysinibacillus sphaericus C3-41
Solibacillus silvestris StLB046
Gemella haemolysans ATCC 10379
Macrococcus caseolyticus JCSC5402
Staphylococcus aureus 04-02981
Staphylococcus haemolyticus JCSC1435
Staphylococcus epidermidis ATCC 12228
Bacillus cytotoxicus NVH 391-98
Bacillus weihenstephanensis KBAB4
Bacillus cereus 03BB102
Bacillus thuringiensis str. Al Hakam
Bacillus anthracis str. Ames Ancestor
Bacillus sp. 1NLA3E
Bacillus megaterium DSM 319
Bacillus pumilus SAFR-032
Bacillus atrophaeus 1942
Bacillus amyloliquefaciens DSM 7
Bacillus subtilis subsp. subtilis str. 168
Bacillus coagulans 2-6
Anoxybacillus flavithermus WK1
Geobacillus kaustophilus HTA426
Geobacillus thermoglucosidasius C56-YS93
Geobacillus sp. WCH70
Symbiobacterium thermophilum IAM 14863
Sulfobacillus acidophilus DSM 10332
Thermaerobacter marianensis DSM 12885
Thermosinus carboxydivorans Nor1
Acetonema longum DSM 6540
Pelosinus fermentans DSM 17108
Veillonellaceae
Moorella thermoacetica ATCC 39073
Syntrophothermus lipocalidus DSM 12680
Syntrophomonas wolfei subsp. wolfei str. Goettingen
Heliobacterium modesticaldum Ice1
Syntrophobotulus glycolicus DSM 8271
Desulfosporosinus acidiphilus SJ4
Desulfosporosinus orientis DSM 765
Desulfosporosinus meridiei DSM 13257
Desulfitobacterium dichloroeliminans LMG P-21439
Desulfitobacterium dehalogenans ATCC 51507
Desulfitobacterium hafniense DCB-2
Thermincola potens JR
Carboxydothermus hydrogenoformans Z-2901
Candidatus Desulforudis audaxviator MP104C
Ammonifex degensii KC4
Desulfotomaculum kuznetsovii DSM 6115
Desulfotomaculum gibsoniae DSM 7213
Desulfotomaculum acetoxidans DSM 771
Pelotomaculum thermopropionicum SI
Desulfotomaculum carboxydivorans CO-1-SRB
Desulfotomaculum ruminis DSM 2154
Desulfotomaculum reducens MI-1
Thermosediminibacter oceani DSM 16646
Tepidanaerobacter acetatoxydans Re1
Mahella australiensis 50-1 BON
Caldicellulosiruptor obsidiansis OB47
Caldicellulosiruptor bescii DSM 6725
Thermoanaerobacterium xylanolyticum LX-11
Thermoanaerobacterium thermosaccharolyticum DSM 571
Thermoanaerobacter tengcongensis MB4
Thermoanaerobacter italicus Ab9
Thermoanaerobacter wiegelii Rt8.B1
Thermoanaerobacter brockii subsp. finnii Ako-1
Clostridium cellulolyticum H10
Clostridium thermocellum ATCC 27405
Oscillibacter valericigenes Sjm18-20
Ethanoligenens harbinense YUAN-3
Ruminococcus bromii L2-63
Faecalibacterium prausnitzii L2-6
Ruminococcus albus 7
Eubacterium siraeum 70/3
Clostridium lentocellum DSM 5427
Coprococcus catus GD/7
Clostridium phytofermentans ISDg
Clostridium saccharolyticum WM1
Eubacterium eligens ATCC 27750
Butyrivibrio proteoclasticus B316
Butyrivibrio fibrisolvens 16/4
Clostridium sp. SY8519
Eubacterium rectale ATCC 33656
Roseburia intestinalis M50/1
Roseburia hominis A2-183
Acetobacterium woodii DSM 1030
Eubacterium limosum KIST612
Eubacterium infirmum F0142
Anaerococcus prevotii DSM 20548
Finegoldia magna ATCC 29328
Alkaliphilus oremlandii OhILAs
Alkaliphilus metalliredigens QYMF
Clostridium difficile 630
Filifactor alocis ATCC 35896
Clostridium sticklandii DSM 519
Candidatus Arthromitus sp. SFB-mouse-Japan
Clostridium beijerinckii NCIMB 8052
Clostridium perfringens str. 13
Clostridium novyi NT
Clostridium cellulovorans 743B
Clostridium pasteurianum DSM 525
Clostridium acetobutylicum ATCC 824

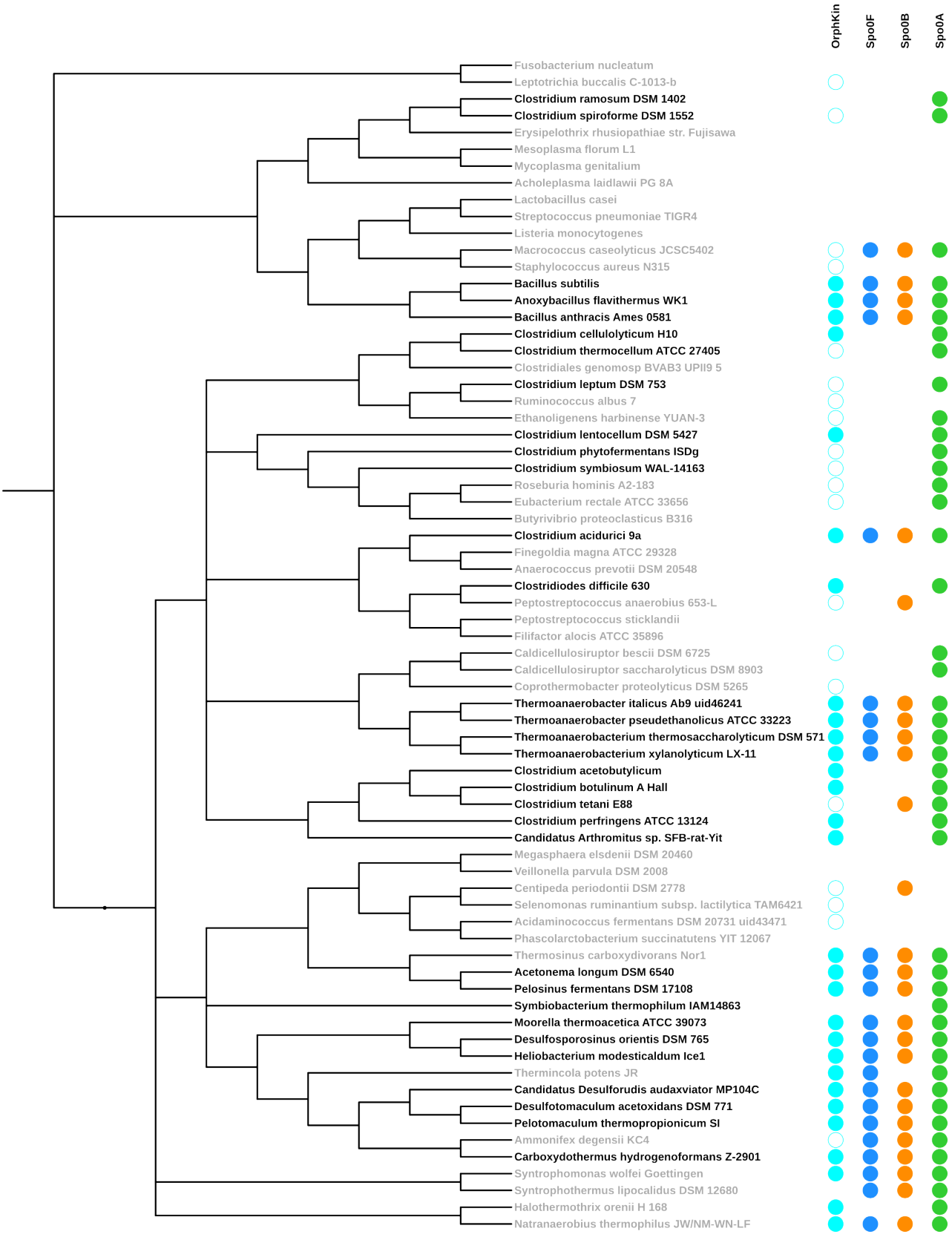

OrphKin
Spo0F
Spo0B
Spo0A
Fusobacterium nucleatum
Leptotrichia buccalis C-1013-b
Clostridium ramosum DSM 1402
Clostridium spiroforme DSM 1552
Erysipelothrix rhusiopathiae str. Fujisawa
Mesoplasma florum L1
Mycoplasma genitalium
Acholeplasma laidlawii PG 8A
Lactobacillus casei
Streptococcus pneumoniae TIGR4
Listeria monocytogenes
Macrococcus caseolyticus JCSC5402
Staphylococcus aureus N315
Bacillus subtilis
Anoxybacillus flavithermus WK1
Bacillus anthracis Ames 0581
Clostridium cellulolyticum H10
Clostridium thermocellum ATCC 27405
Clostridiales genomosp BVAB3 UPII9 5
Clostridium leptum DSM 753
Ruminococcus albus 7
Ethanoligenens harbinense YUAN-3
Clostridium lentocellum DSM 5427
Clostridium phytofermentans ISDg
Clostridium symbiosum WAL-14163
Roseburia hominis A2-183
Eubacterium rectale ATCC 33656
Butyrivibrio proteoclasticus B316
Clostridium acidurici 9a
Finegoldia magna ATCC 29328
Anaerococcus prevotii DSM 20548
Clostridiodes difficile 630
Peptostreptococcus anaerobius 653-L
Peptostreptococcus sticklandii
Filifactor alocis ATCC 35896
Caldicellulosiruptor bescii DSM 6725
Caldicellulosiruptor saccharolyticus DSM 8903
Coprothermobacter proteolyticus DSM 5265
Thermoanaerobacter italicus Ab9 uid46241
Thermoanaerobacter pseudethanolicus ATCC 33223
Thermoanaerobacterium thermosaccharolyticum DSM 571
Thermoanaerobacterium xylanolyticum LX-11
Clostridium acetobutylicum
Clostridium botulinum A Hall
Clostridium tetani E88
Clostridium perfringens ATCC 13124
Candidatus Arthromitus sp. SFB-rat-Yit
Megasphaera elsdenii DSM 20460
Veillonella parvula DSM 2008
Centipeda periodontii DSM 2778
Selenomonas ruminantium subsp. lactilytica TAM6421
Acidaminococcus fermentans DSM 20731 uid43471
Phascolarctobacterium succinatutens YIT 12067
Thermosinus carboxydivorans Nor1
Acetonema longum DSM 6540
Pelosinus fermentans DSM 17108
Symbiobacterium thermophilum IAM14863
Moorella thermoacetica ATCC 39073
Desulfosporosinus orientis DSM 765
Heliobacterium modesticaldum Ice1
Thermincola potens JR
Candidatus Desulforudis audaxviator MP104C
Desulfotomaculum acetoxidans DSM 771
Pelotomaculum thermopropionicum SI
Ammonifex degensii KC4
Carboxydothermus hydrogenoformans Z-2901
Syntrophomonas wolfei Goettingen
Syntrophothermus lipocalidus DSM 12680
Halothermothrix orenii H 168
Natranaerobius thermophilus JW/NM-WN-LF

Figure 3.8: Antunes tree (Page 60): Cladogram of 205 Firmicutes genomes rooted by 13 outgroup species, adapted from Figure 2 in [Antunes et al., 2016]. This tree was constructed using Bayesian tree inference on a concatenated alignment of 47 ribosomal protein sequences. Collapsed clades (Veillonellaceae and Lactobacillaceae) represent species that do not harbor Spo0A (or Spo0F or Spo0B) [this work] and are considered to be non-sporulators [Ludwig et al., 2015; Rogosa, 1971]. Some species encode orphan kinases, but, of those, none encode a PAS domain. Colored dots indicate Spo0 pathway components predicted by the methods used in this study: one or more orphan HK (cyan); Spo0F (blue); Spo0B (orange); Spo0A (green). Filled cyan dot indicates at least one orphan HK encodes a PAS domain.

Figure 3.9: Yutin tree (Page 61): Cladogram of 68 Firmicutes genomes, outgroup rooted using Leptotrichia buccalis and Fusibacterium nucleatum. Tree constructed from the concatenated alignment of 50 ribosomal protein sequences using FastTree [Price et al., 2010] and Treefinder [Jobb et al., 2004], as described in Yutin and Galperin [2013]. Species known to sporulate labeled in black. Colored dots indicate Spo0 pathway components predicted by the methods used in this study: one or more orphan HKs (cyan); Spo0F (blue); Spo0B (orange); Spo0A (green). Filled cyan dot indicates at least one orphan HK encodes a PAS domain.

[Zhilina et al., 2012], *Fuchsiella ferrireducens* [Zhilina et al., 2015], and *Natroniella acetigena* [Zhilina et al., 1996]. The genus *Sporohalobacter* was initially reported to form spores [Oren et al., 1991], but subsequent characterization called this result into question as no growth was obtained following heat treatment [Ben Abdallah et al., 2015]. Swollen end cells were characterized as pre-spore-like structures, but no spores were observed via microscopy. These experimental differences could be due to the conditions tested or different spore definitions. Of spore-forming Halanaerobiales, the only available genome is *Halonatronum saccharophilum*, which does not encode Spo0F or Spo0B orthologs by our genome neighborhood conservation method.

Genomes from Halanaerobiales and Natranaerobiales form a clade in both recently published trees, but the location of that clade, in relation to other Firmicutes differs. In the Antunes tree, these taxa are basal to the divergence of Classes Bacilli and Clostridia. Since this placement makes them the earliest branching clade within the Firmicutes, the presence of an apparent phosphorelay architecture in *N. thermophilus* supports the hypothesis that the emergence of the Spo0 phosphorelay predates the divergence of the Clostridia and Bacilli.

In the Yutin tree, the clade (which includes one representative of each order, *Natranaerobius thermophilus* and *Halothermothrix oreni*) is one of several descendants of a polytomy at the base of Class Clostridia. This placement does not change our prediction that the common ancestor of the Clostridia and Bacilli likely encoded a phosphorelay Spo0 pathway.

Regardless of their phylogenetic placement in the context of other species, the presence of both

architectures in these sister taxa adds to the patchy distribution of Spo0 architectures observed throughout the Phylum, requiring an additional remodeling event to explain the present-day phylogenetic distribution. The placement of these sister taxa in the two recently published trees does not contradict the hypothesis that the phosphorelay architecture was present in the common ancestor of the Bacilli or Clostridia; moreover, the evidence from one of those studies suggests that it predates that common ancestor.

3.4.1.2. *Alkaliphilus*, *Gottschalkia*, and Sporulating Peptostreptococcaceae

All three trees predict a clade consisting of species from the families Clostridiaceae (only genus *Alkaliphilus*), Eubacteriaceae (genera *Acetobacterium*, *Eubacterium*), Peptoniphilaceae (genera *Finegoldia*, *Anaerococcus*), Incertae sedis XI (*Gottschalkia*), and Peptostreptococcaceae (genera *Clostridiodes*, *Peptoclostridium*, *Filifactor*). In all three trees, this clade is a sister taxon to the *Clostridium sensu strictu*. Many genera within this clade are reportedly asporogenous [Galperin et al., 2012]. The key exceptions are four species in the genera *Alkaliphilus* [Cao et al., 2003], *Gottschalkia* [Yutin and Galperin, 2013], and *Clostridioides* [Lawson et al., 2016]. Each tree contains a different subset of these four spore-formers: all contain *Clostridioides difficile*, the Antunes et al. [2016] tree and my tree contain *Alkaliphilus metalliredigens*, the Yutin and Galperin [2013] tree and my tree contain *Gottschalkia acidurici*, Antunes additionally encodes *Alkaliphilus oremlandii*.

All trees agree on the relationships between these taxa: *Alkaliphilus* is a sister taxon to *Clostridiodes difficile* and other Peptostreptococcaceae, when *Gottschalkia* is not present [Antunes et al., 2016], and vice versa [Yutin and Galperin, 2013]. When both are present (as in my tree), *Gottschalkia* is basal to a clade that includes both *Alkaliphilus* and *Clostridiodes*. Notably, there are asporogenous species interleaved between these taxa in all three trees.

These species are particularly interesting because, although closely related, they have different predicted Spo0 architectures. *A. metalliredigens* and *G. acidurici* were found to encode homologs of Spo0F and Spo0B and are therefore likely to initiate sporulation via a phosphorelay. No phosphorelay homologs were observed in either of the sporulating species *A. oremlandii* or *C. difficile*, suggesting that they have a direct phosphorylation Spo0 pathway. The presence of the phosphorelay homologs in *G. acidurici* and *A. metalliredigens* suggests that the phosphorelay has persisted despite repeated losses of Spo0F and Spo0B and/or sporulation within closely related taxa. This mixed distribution implies multiple transitions from phosphorelay to direct phosphorylation architecture within the Clostridiales, one at the base of each divergent group. This inference is supported by all three trees.

3.4.1.3. Predicted architectures within Class Bacilli

Each tree has a different set of species from Class Bacilli, but the results of these differences are not at variance with the observations made here. Homologs of Spo0F and Spo0B were detected by our methods in all genomes in spore-forming Bacilli represented in the three data sets, with the exception of two *Erysipelaclostridium* genomes, two *Paenibacillus* genomes and the genome of *Sporolactobacillus inulinus*. Each of these exceptions is treated below.

***Erysipelatoclostridium*:** The genomes of *Erysipelatoclostridium ramosum DSM 1402* and *Erysipelatoclostridium spiroforme DSM 1552* both encode Spo0A; *E. spiroforme* additionally encodes an orphan kinase. Spore formation in these species has been described as "rare or absent" [Kaneuchi et al., 1979; Lavigne et al., 2003; Yutin and Galperin, 2013]. Based on the set of proteins present in the sequenced strains used in this study, if they do form spores, it is likely signaled by direct phosphorylation of Spo0A. No Erysipelatrichiaceae species were included in the Antunes et al. [2016] data set. In the Yutin and Galperin [2013] tree, these genomes are basal to all other Bacilli. In this study, these genomes are placed within the Bacillaceae clade. However, since the Yutin data set does not include any early branching genomes in class Bacilli (e.g. Paenibacillaceae), the two placements are consistent.

***Paenibacillus*:** In my tree, both Paenibacillaceae genomes, *Paenibacillus polymyxa* and *Brevibacillus brevis*, included in our representative set possess all phosphorelay components. These genomes are monophyletic and are phylogenetically placed basal to the Bacillaceae. These taxa are not represented in the Yutin and Galperin [2013] tree. The Antunes et al. [2016] tree includes the genomes of six members of the Paenibacillaceae, including genomes from the genera *Desmospora*, *Brevibacillus*, *Paenibacillus*, and *Thermobacillus*. These species are phylogenetically placed basal to the Bacillaceae, though paraphyletically.

Spo0B was not identified in two of these species, *Paenibacillus mucilaginosa* and *Paenibacillus sp. Y412MC10*. In both cases, inspection of the genome neighborhood of these two species reveals a hypothetical protein with the similar in sequence and domain content to Spo0B, though it appears to be missing a stop codon. This could be a loss of function mutation or due to an error in sequencing or assembly. If Spo0B is truly a pseudogene in these species, this could indicate either the loss of sporulation or gain of the ability to sporulate via direct phosphorylation of Spo0A in these individuals. Interestingly, the *Paenibacillus polymyxa* kinase PP_1077 has been reported to directly phosphorylate *P. polymyxa* Spo0A when heterologously expressed in a *B. subtilis* mutant lacking Spo0B but not lacking any of the *B. subtilis* sporulation kinases [Park et al., 2012]

(See Section 2.3.2.1.3). Spo0 architectures in these species warrant further investigation. This was accepted as sufficient evidence for the presence of Spo0B in our analysis.

***Sporolactobacillus*:** *S. inulinus* is present in the Antunes et al. [2016] data set, but was not included in either my tree or the Yutin tree. The Antunes tree places this species basal to the Bacillaceae, diverging after the Paenibacillaceae. Spo0F was not identified by conserved genome neighborhood in *S. inulinus*, although *S. inulinus* is reported to produce endospores [Kitahara and Lai, 1967]. However, inspection of predicted RRs lacking an output domain in that species did reveal a possible candidate. This protein (SINU_10335) aligns well to known and predicted Spo0F sequences and has specificity residues typical of Spo0F (QGILEVD), although it is not encoded in the proximity of any of the Spo0F neighborhood markers. All other single-domain RRs in this species had less sequence similarity to Spo0Fs and specificity residues that did not reflect the Spo0F signature (Fig 4). This was accepted as sufficient evidence for the presence of Spo0F in our analysis.

3.4.1.4. Ruminococcaceae and Lachnospiraceae

Species from the Ruminococcaceae and Lachnospiraceae are well-sampled in all three trees, however the relationship of these two families with respect to the Clostridiaceae varies slightly. In our tree and the Antunes et al. [2016] tree, they are sister taxa to the Clostridiaceae, while the relationship between these three clades is not resolved in the Yutin and Galperin [2013] tree. All sporogenous members of the Ruminococcaceae and Lachnospiraceae families are predicted to encode a direct phosphorylation architecture. If these two clades represent distinct lineages that are not sister taxa, as possible in the Yutin and Galperin [2013] tree, then an additional transition between phosphorelay and direct phosphorylation architecture is required to explain the phylogenetic distribution of Spo0 pathway arechitectures in these species.

3.4.1.5. Spore-formers with symbiotic or syntrophic life styles

Several spore-forming species in the Antunes et al. [2016] tree have symbiotic or syntrophic lifestyles, including members of the genera *Symbiobacterium* [Ohno et al., 2000], *Thermaerobacter* [Han et al., 2010], *Tepidanaerobacter* [Westerholm et al., 2011], and *Thermosediminibacter* [Pitluck et al., 2010]. Only *Symbiobacterium thermophilus* was included in the Yutin and Galperin [2013] tree and none were included in my tree.

Candidate Spo0F and Spo0B orthologs were not found in the *Symbiobacterium* and *Thermaerobacter* genomes, but were found in the closely related species, *Sulfobacillus acidophilus DSM*

10332, which has a typical free-living lifestyle [Norris et al., 1996]. The sister taxa *Tepidanaerobacter* and *Thermosediminibacter* encode orthologs of Spo0B, but not Spo0F, and lack orphan kinases. Of these species, all but *Thermosediminibacter oceani* have been observed to produce spores.

Interpreting the Spo0 pathway architectures in these species is complicated by their symbiotic nature. For example, *Symbiobacterium thermophilum* displays marked growth dependence on microbial commensalism with *Bacillus* sp. Strain S [Ueda et al., 2004] similarly, sporulation increases from 0.1% to 20% when cultured in a dialysis flask with a constant influx of media used by *Bacillus* sp. Strain S [Ueda et al., 2004]. This apparent reliance on external factors to promote sporulation initiation could be mediated as a signal or potentially as access to proteins that facilitate the phosphorylation of Spo0A. Further work on the growth factors that these syntrophic and symbiotic bacteria rely on may reveal the mechanism of initiation of sporulation in these species. Taking a conservative stance, I do not interpret the absence of genes encoding Spo0 pathway proteins to be evidence of an alternative Spo0 pathway architecture in symbiotic or syntrophic strains.

3.5. Summary

In this chapter, I have presented evidence for the existence of functional phosphorelays in many Clostridia and most Bacilli surveyed. This is in contrast to the current prevailing hypothesis for Spo0 pathway evolution, which predicts that that Spo0 pathways with a phosphorelay architecture only be observed in class Bacilli.

Further, examination of the phylogenetic distribution of the predicted architectures in the Clostridia and the Bacilli reveals that neither predicted architecture is monophyletic within either taxonomic class. To ensure that this patchiness is not a byproduct of taxon sampling or phylogeny reconstruction artifacts, I repeated the computational analysis with two other Firmicutes phylogenies [Antunes et al., 2016; Yutin and Galperin, 2013], one of which includes a much larger set of genomes. Both analyses revealed similar patchiness.

The presence of Clostridial phosphorelays and the patchy distribution of both architectures calls for the reconsideration of the evolutionary history of the sporulation initiation pathway in the Firmicutes.

Chapter 4

Evolutionary History of the Spo0 Pathway

A proposed evolutionary history of the Spo0 pathway must provide an explanation for the phylogenetic distribution of present-day Spo0 architectures and answer the following questions: What is the architecture of the common ancestral pathway? In which lineage(s) did the phosphorelay arise? Did the phosphorelay arise more than once? In which lineages did remodeling occur?

The prevailing hypothesis (reviewed in Section 2.3.3.1) is that the ancestral Spo0 pathway was a conventional two-component pathway with the emergence of the Bacillar phosphorelay following the separation of the classes Bacilli and Clostridia. This model predicts that phosphorelays will only be observed in Class Bacilli and direct phosphorylation architecture pathways only in Class Clostridia, which conflicts with the evidence reported in Chapter 3.

The common ancestral pathway could still have a direct phosphorylation architecture, assuming that the phosphorelay arose following the divergence of classes Bacilli and Clostridia. However, if the phosphorelay only arose once, this scenario requires multiple transfers resulting in remodeling from a direct phosphorylation architecture to result in the present-day distribution. Alternatively, each clade encoding a phosphorelay could have invented it independently, remodeling from the ancestral direct phosphorylation architecture to a phosphorelay.

If the common ancestral pathway was a phosphorelay, only a single genesis of the phosphorelay is required and all phosphorelay architectures could have been vertically inherited. In this scenario, each direct phosphorylation architecture could have been invented independently, or a kinase that directly phosphorylates Spo0A could have been transferred between direct phosphorylation architecture encoding clades.

Here, I present evidence based on the comparison of specificity residues and specific interactions, that all phosphorelays are related to the same common ancestral phosphorelay, consistent with the scenarios in which the phosphorelay only arose once. Analysis of the genomic context of Spo0F and Spo0B proteins suggests an unlikely series of transfer events would be required to obtain the proteins necessary to remodel from a direct phosphorylation architecture to a phosphorelay. These results are corroborated by phylogenetic analysis of both Spo0F and Spo0B that are consistent with vertical descent. Thus, my results support an ancestral phosphorelay hypothesis where the direct phosphorylation architecture is the derived state through remodeling.

4.1. Specificity Residue Similarity

In order to assess change and conservation in the genetic determinants of specificity in the course of Spo0 pathway evolution, we compared the specificity residues in Spo0 pathway proteins between pathway architectures (phosphorelay versus direct phosphorylation) and, for the phosphorelay, between taxonomic groups (Bacilli versus Clostridia). Qualitative comparisons of the specificity logos for the phosphorelays from both classes (Fig 3.4) suggest that specificity residues are relatively similar to each other. If the phosphorelays from both classes share common ancestry, I hypothesize that specificity will have been maintained, which may be evident from similarities in specificity residue logos. The converse (where specificity has not been maintained between the phosphorelay in the two classes) could suggest multiple independent inventions of the phosphorelay or could be the result of common ancestry with perturbation of specificity by another pathway.

To provide a more rigorous comparison of these logos, I created a method to quantify the differences in specificity residues for two sets of orthologous proteins, as follows. The specificity residues for a set of n orthologous proteins can be represented as an n by l matrix, where l is the number of specificity sites (i.e., $l = 6$ for kinases and $l = 7$ for response receivers). At each site, I focus on differences in residue frequencies that distinguish the two sets by considering only those residues for which I can reject the null hypothesis that the frequencies in set 1 and set 2 represent the same underlying distribution. Specifically, I consider the frequencies of amino acid a at site j in the two sets to be significantly different, if I can reject this null hypothesis at the α level.

Given specificity residues from two sets of orthologs, I calculate a score $S_{1,2}(j)$ for site j:

$$S_{1,2}(j) = \sum_{a} |f_1(a, j) - f_2(a, j)|,$$

where the summation is over the 20 amino acids and $f_m(a, j)$ is defined to be the frequency of

amino acid a at site j in set $m \in \{1, 2\}$, if the frequencies of a in the two sets are significantly different. If they are not significantly different, $f_m(a, j) = 0$. An overall score, $S_{1,2}$, for the difference in two sets is obtained by averaging $S_{1,2}(j)$ over all l sites.

I used the program TwoSampleLogo [Vacic et al., 2006] to obtain the values of $f_m(a, j)$ required to calculate $S_{1,2}$. This program includes a statistical framework for assessing whether residue frequencies are significantly different and offers a selection of statistical tests. I used the t-test with significance threshold $\alpha = 0.01$. Given two sets of amino acids and a significance threshold, α, as input, `TwoSampleLogo -F TXT` reports the frequencies, filtered by p-value, of residue a at each site in the two sets. I also used TwoSampleLogo [Vacic et al., 2006] to obtain visual comparison of two sets of specificity residues.

Using this technique, I compared the logos built on Bacillar and Clostridial phosphorelay Spo0 components (Fig 3.4A and B) using filtering with a p-value of α=0.01 and no p-value filter for comparison (Table 4.1, Figure 4.1).

For all components, the average difference per site, using p-value filtering where α=0.01, suggests very little significant change is observed between these logos. These results suggest that the specificity residues of orphan kinases of the Bacilli and Clostridia phosphorelays are, in fact, similar to each other. Spo0B and Spo0F are also similar between Bacillar and Clostridial phosphorelays and most differences are conservative (e.g. sites three and six in Spo0F and site two in Spo0B are hydrophobic residues in both sets, f_1 and f_2).

Table 4.1: Comparison of Spo0 Phosphorelay Logos

Taxonomy Set		p-value Filtering			
f_1	f_2	α=0.01		No Filtering	
Bacilli	Clostridia	Orphan Kinases	Spo0A	Orphan Kinases	Spo0A
		0.18	0.29	0.45	0.69
Bacilli	Clostridia	Spo0F	Spo0B	Spo0F	Spo0B
		0.28	0.44	0.46	0.87

4.2. Preservation of interaction between *D. acetoxidans* and *B. subtilis* Phosphorelays

To test the hypothesis that specificity has been preserved between Bacillar and Clostridial phosphorelays, as suggested by the comparison of specificity residues, I employed a cross-species com-

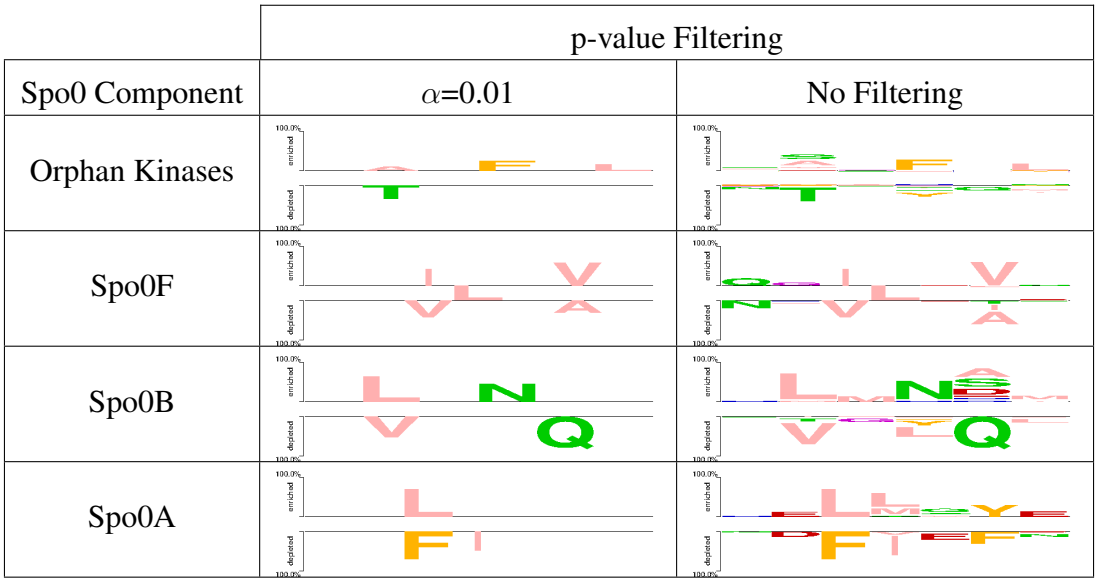


Figure 4.1: Visual comparison of Spo0 phosphorelay logos. In each row, taxonomic set f_1 is Bacilli and set f_2 is Clostridia.

plementation assay. I asked whether phosphorelay proteins in *D. acetoxidans* could recapitulate the function of the corresponding proteins in *B. subtilis*, *in vitro*. Specifically, I examined phosphotransfer in the *B. subtilis* phosphorelay, as described for *D. acetoxidans* in Section 3.3.2, systematically replacing each *B. subtilis* protein with its *D. acetoxidans* counterpart (Fig 4.2, lanes 1-6). For each step in the pathway, a band corresponding to the replacement *D. acetoxidans* protein was observed, demonstrating that each *D. acetoxidans* protein was capable of accepting a phosphoryl group from the upstream *B. subtilis* Spo0 pathway component (Fig 4.2, lanes 2, 4, and 6). Moreover, bands were observed for downstream components of the *B. subtilis* phosphorelay, where included, indicating that the *D. acetoxidans* replacement was also capable of transferring a phosphoryl group to the downstream component in the *B. subtilis* phosphorelay (Fig 4.2, lanes 1, 3, 5). Additionally, the phosphorylation of *D. acetoxidans* Spo0A by the *B. subtilis* phosphorelay required the presence of both *B. subtilis* Spo0F and Spo0B proteins (Fig 4.2, lanes 7-9), indicating that *D. acetoxidans* Spo0A cannot be directly phosphorylated by *B. subtilis* KinA.

Taken together, our results demonstrate that, in all cases, the *D. acetoxidans* proteins were able to recapitulate the function of their counterparts in *B. subtilis*. Thus, not only do these two pathways both independently interact as phosphorelays, *in vitro* [Burbulys et al., 1991, and this work], they also encode sufficiently similar phosphotransfer specificity to render them functionally interchangeable, at least *in vitro*. Therefore, either these two phosphorelays arose independently with interchangeable genetic determinants of specificity, which I deem highly unlikely, or these

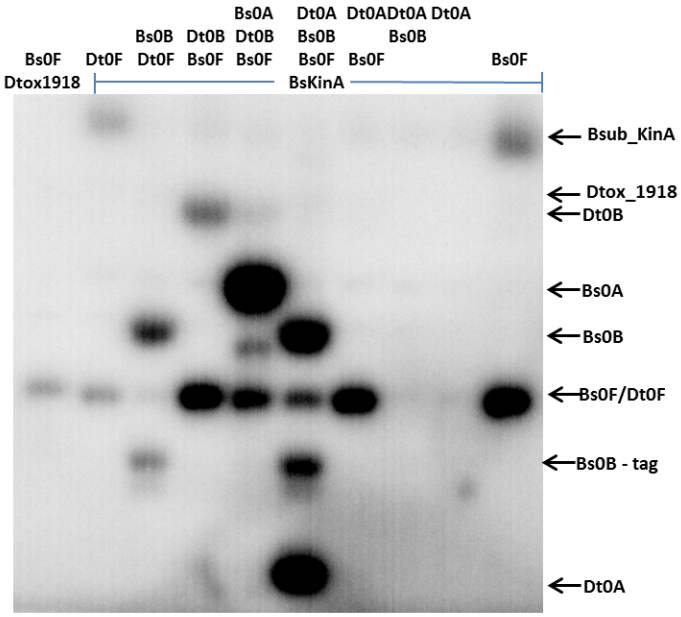


Figure 4.2: Phosphotransfer profiling of interchanging Spo0 phosphorelay components from *D. acetoxidans* (Dt) with *B. subtilis* (Bs) Spo0 phosphorelay components. Reactions contained 5uM Dtox_1918 or Bsub_KinA and 10uM each of the proteins listed above each lane in the phosphorimage. Dtox_1918 and Bsub_KinA were autophosphorylated for 10 minutes. All components were incubated together for 5 minutes.

pathways are the descendants of a common ancestral pathway.

4.3. Ancestral Phosphorelay Hypothesis

Based on the degree of specificity similarity between Bacillar and Clostridial phosphorelays, as measured by specificity residues and *in vitro* interchangeability, I hypothesize that a Spo0 phosphorelay arose only once in evolutionary history. I further hypothesize that this sole genesis of the phosphorelay occurred prior to the divergence between Class Bacilli and Class Clostridia. In this scenario, the present-day direct phosphorylation architectures arose through multiple, independent episodes of pathway remodeling, resulting in a patchy distribution of pathway architectures. Further, given the evidence that I have presented for the single genesis of a phosphorelay (Sections 4.1 and 4.2), this is the most parsimonious evolutionary history of the Spo0 pathway compared to the alternative hypotheses considered above.

Whereas multiple independent inventions of a phosphorelay would also result in a patchy distribution, the complexity of the pathway, coupled with the dramatic similarities between predicted

phosphorelay proteins in both classes, render two or more independent inventions of the sporulation initiation phosphorelay unlikely. Based on the similarities observed, it is unlikely that these two pathways encode similar specificity by chance; rather, specificity in the Spo0 phosphorelay has been preserved over 2.7 billion years of independent evolution.

Alternatively, if the ancestral Firmicute encoded a direct, two-component pathway and the phosphorelay arose in one of the two major lineages after their divergence, horizontal transfer of the phosphorelay to taxa in the other class would be required. However, acquisition of the phosphorelay through horizontal gene transfer entails an improbable series of events. Regardless of where the phosphorelay arose, multiple independent acquisitions through transfer would be required to produce the present-day distribution, because genomes that harbor a phosphorelay are not monophyletic in either taxonomic class. Moreover, each acquisition of the full phosphorelay would likely require multiple, independent horizontal transfer events, because the genes encoding Spo0 components are dispersed throughout the genome (Figure 4.3). Spo0F is the only Spo0 protein that is consistently encoded in a specific genomic region. It is almost always located near the origin of replication (i.e., between 0 and 36 degrees or between 270 and 360 degrees). The remaining Spo0 components are dispersed throughout the genome, although they are rarely encoded near the origin of replication. Further, because the genomic neighborhoods of Spo0F and Spo0B are conserved, each transfer event would have to result in the insertion of Spo0F or Spo0B into the same neighborhood or the transfer of the entire neighborhood.

Thus, based on the combined evidence, which suggests that there was a single genesis of the phosphorelay and that transfer is unlikely, I conclude that the phosphorelay was most likely present in the ancestor of all Firmicutes and all present-day phosphorelays are derived from it by vertical descent.

4.3.1. Evolution of kinase and Spo0A proteins

Having established strong evidence of common ancestry and conserved specificity in the Spo0 phosphorelay, we next considered the evolutionary history of the direct phosphorylation Spo0 pathway. According to this ancestral phosphorelay hypothesis, present-day direct phosphorylation pathways are a result of multiple independent transitions away from a phosphorelay architecture, wherein Spo0F and Spo0B were lost and direct phosphorylation of Spo0A was gained. Our experimental and computational results suggest that the Spo0A specificity spectrum has remained relatively constant and that the transition was, therefore, primarily mediated by changes in the sporulation kinases. In this section, I compare characteristics of kinases and regulators in phosphorelay and direct phosphorylation to gain insight into the changes thatmay have resulted in such

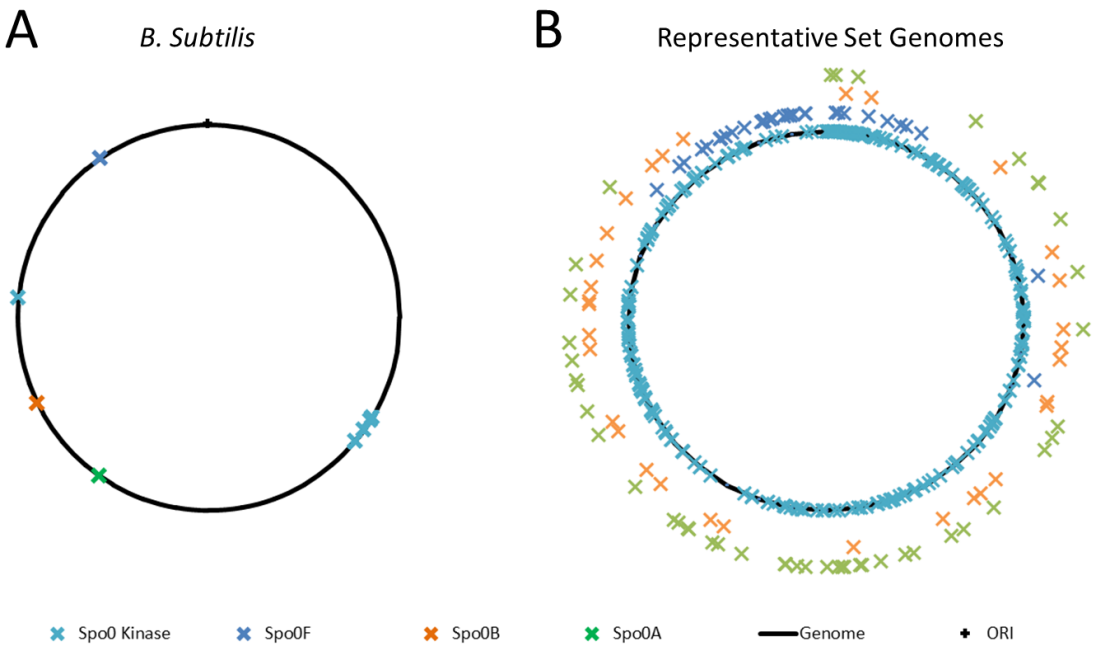


Figure 4.3: Genomic distribution of Spo0 components in (A) *B. subtilis* and (B) all members of the Firmicutes representative set. Protein location calculated based on the position of the start codon divided by total number of bases in the genome. Orphan Histidine kinases, Spo0B, and Spo0A (shown in teal, orange, and green, respectively) are found throughout the genomes; Spo0F (blue) is commonly near the origin of replication. Separation by Firmicutes class or predicted Spo0 architecture (not shown) does not change the observed broad distribution of orphan kinases, Spo0B, or Spo0A; nor does it affect the regional localization of Spo0F.

a remodeling.

4.3.1.1. Specificity residues

To obtain insight into the remodeling events that resulted in the present-day distribution of Spo0 architectures, I compared the specificity logos of orphan kinases and Spo0A from phosphorelays and direct phosphorylation architectures. This comparison revealed that the similarity across architectures is greater for Spo0A proteins than for orphan kinases. When Clostridial and Bacillar phosphorelays are considered separately, Spo0A specificity residues are more similar within the same taxonomic class, than within the same pathway type (Figs 4.4). The opposite is true for candidate sporulation kinases. The specificity residues of candidate phosphorelay kinases from both the Clostridia and the Bacilli differ markedly from those of kinases predicted to phosphorylate Spo0A directly. This difference is even more dramatic when experimentally verified sporulation kinases associated with the two architectures are compared (Fig 3.4E). These results suggest that

architectural remodeling was driven primarily by changes in kinase, rather than Spo0A, specificity.

Table 4.2: Comparison of Spo0 phosphorelay and direct phosphorylation logos

		p-value Filtering			
		α=0.01		No Filtering	
Sets		Orphan Kinases	Spo0A	Orphan Kinases	Spo0A
f_1	f_2				
Bacilli Phosphorelay	Direct Phosphorylation	0.50	0.40	0.80	0.80
Clostridia Phosphorelay	Direct Phosphorylation	0.48	0.00	0.73	0.44

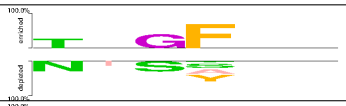
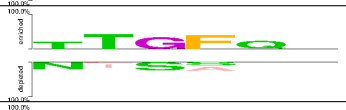
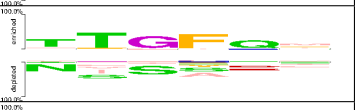
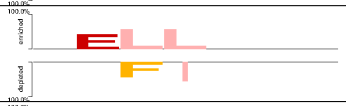
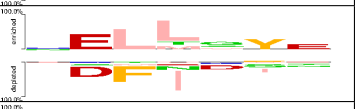
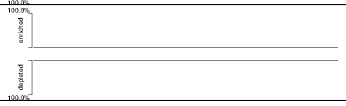
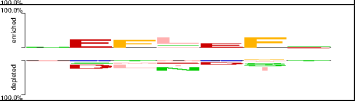
	Sets	p-value Filtering	
Spo0 Component	f_1 f_2	α=0.01	No Filtering
Orphan Kinases	Bacilli Phosphorelay Direct Phosphorylation		
Orphan Kinases	Clostridia Phosphorelay Direct Phosphorylation		
Spo0A	Bacillus Phosphorelay Direct Phosphorylation		
Spo0A	Clostridia Phosphorelay Direct Phosphorylation		

Figure 4.4: Comparison of Spo0 Phosphorelay and Direct Phosphorylation Logos

4.3.1.2. Catalytic Domains differ by predicted architecture

The differences between the two sets of orphan kinases associated with the two pathway architectures extend beyond specificity residues: the histidine kinase catalytic domains found in the two sets also differ. Kinase catalytic domains have been categorized into 23 subfamilies based on HMM models of multiple sequence alignments [Alexander and Zhulin, 2007]. While many of these catalytic domain subtypes are present in kinases of the Firmicutes, only two, HK_CA:2 and HK_CA:3, are commonly observed in orphan kinases. In genomes with a predicted direct phosphorylation Spo0 pathway, 51 out of 53 orphan kinases (96%) possess a HK_CA:2 type catalytic domain. In genomes with a predicted Spo0 phosphorelay, 130 out of 155 orphan kinases

(86%) possess a HK_CA:3 type catalytic domain. The same dichotomy is observed for experimentally verified kinases: direct phosphorylation architecture Spo0 kinases in *C. acetobutylicum*, *R. thermocellum*, and *C. difficile* encode an HK_CA:2 type catalytic domain [Mearls and Lynd, 2014; Steiner et al., 2011; Underwood et al., 2009] and the phosphorelay sporulation kinases in *B. subtilis*, *D. acetoxidans*, and other experimentally verified Spo0 pathways harbor a HK_CA:3 type catalytic domain (see Table A9 for a complete list). There is one exception in the experimentally verified set; Ca_C3319 of *C. acetobutylicum* encodes an HK_CA:3 domain, despite being involved in a direct phosphorylation architecture.

4.3.1.3. Gene trees

To better understand the evolutionary history of the individual gene families represented in the Spo0 pathway, I constructed gene trees for each one.

Spo0F and Spo0A: Here, I present a phylogenetic analysis of the Spo0 component REC domain sequences from both the Spo0F (69 sequences) and the Spo0A (121 sequences) proteins.

These sequences obtained from the Spo0 components predicted in the Antunes et al. [2016] data set, as this provides the the most comprehensive taxon sampling of the three phylogenetic data sets considered in this thesis. In addition, six LytTR REC domain sequences from the genomes of *B. subtilis*, *C. acetobutylicum*, and *D. acetoxidans* were included as outgroups, in order to root the phylogeny. The LytTR output domain-associated REC domains were selected, because LytTR has been observed in other bacterial phyla, suggesting that it pre-dates the advent of the Spo0F and Spo0A REC domains and, hence, is a suitable outgroup.

The resulting 196 sequences were aligned using MAFFT [Katoh et al., 2005] and GUIDANCE [Sela et al., 2015], with alignment quality filtering set at 92%. The phylogeny was constructed using RAxML [Stamatakis, 2014] with the CATPROTGAMMA model [Stamatakis, 2006] and bootstrapped 100 times.

The resulting phylogeny (Figs 4.5-4.7) presents a clear partition of Spo0F (blue branches) and Spo0A (green branches) REC domains into separate clades. The backbone of this tree is not very well-supported, as measured by bootstrap analysis except at the root of the Spo0A subtree. This is likely due to the short length of these sequences (the alignment comprises 121 columns) and high level of sequence conservation among REC domains, both of which reduce the phylogenetic information available to resolve these subclades. However, the support for the separation of the Spo0A clade from the Spo0F clade is quite strong (71).

The overall topology of the tree is consistent with a single genesis of the phosphorelay architecture. Not only do Spo0F and Spo0A form separate clades, but the topology of the Spo0F and

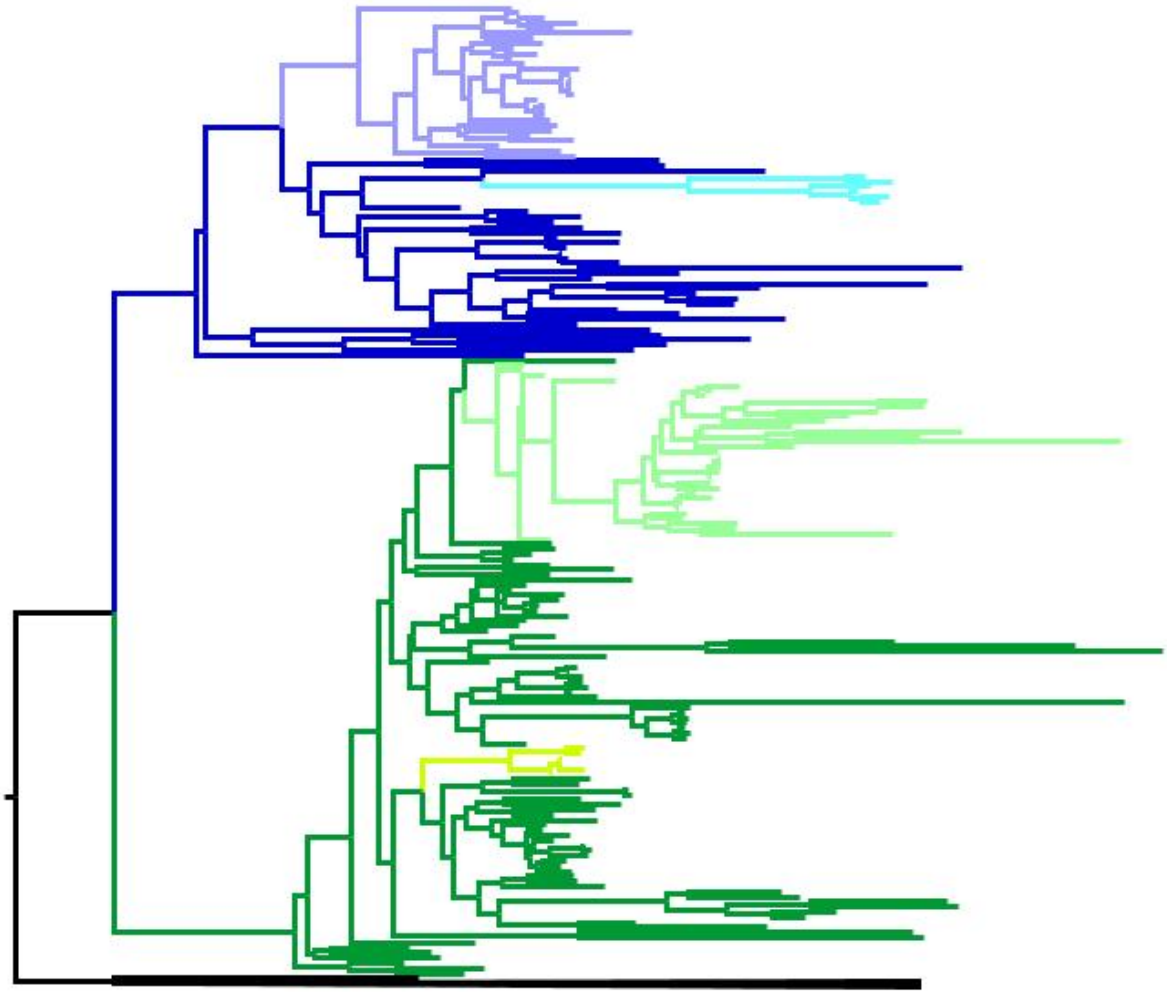


Figure 4.5: Phylogenetic relationships between Spo0F (blue) and Spo0A (green) REC domains in Firmicutes genomes. Spo0F and Spo0A sequences form separate clades. Within each subfamily, major species groups are clearly separated. Bacillar 0F/A sequences shown in light blue/light green; Thermoanaerobacterales 0F/0A sequences shown in turquoise/lime; 0F/0A sequences in all other Firmicutes shown in dark blue/green; outgroup sequences shown in black.

Spo0A subtrees is roughly in accordance with that of the Antunes species phylogeny. Further, this topology is consistent with an ancient duplication of the REC domain, at the base of, or possibly prior to, the emergence of the Firmicutes phylum, followed by vertical inheritance of both Spo0 response regulators.

The position of one clade near the base of the Spo0F subtree deviates from the Antunes tree. In this clade, a group of closely related, early branching Bacillales species (*Oceanobacillus iheyensis*, *Halobacillus halophilus*, and *Amphibacillus xylanus*), which should be located deep in the tree, are clustered with the early branching extremophile, *Natranaerobius thermophilus* (Natranaerobiales,

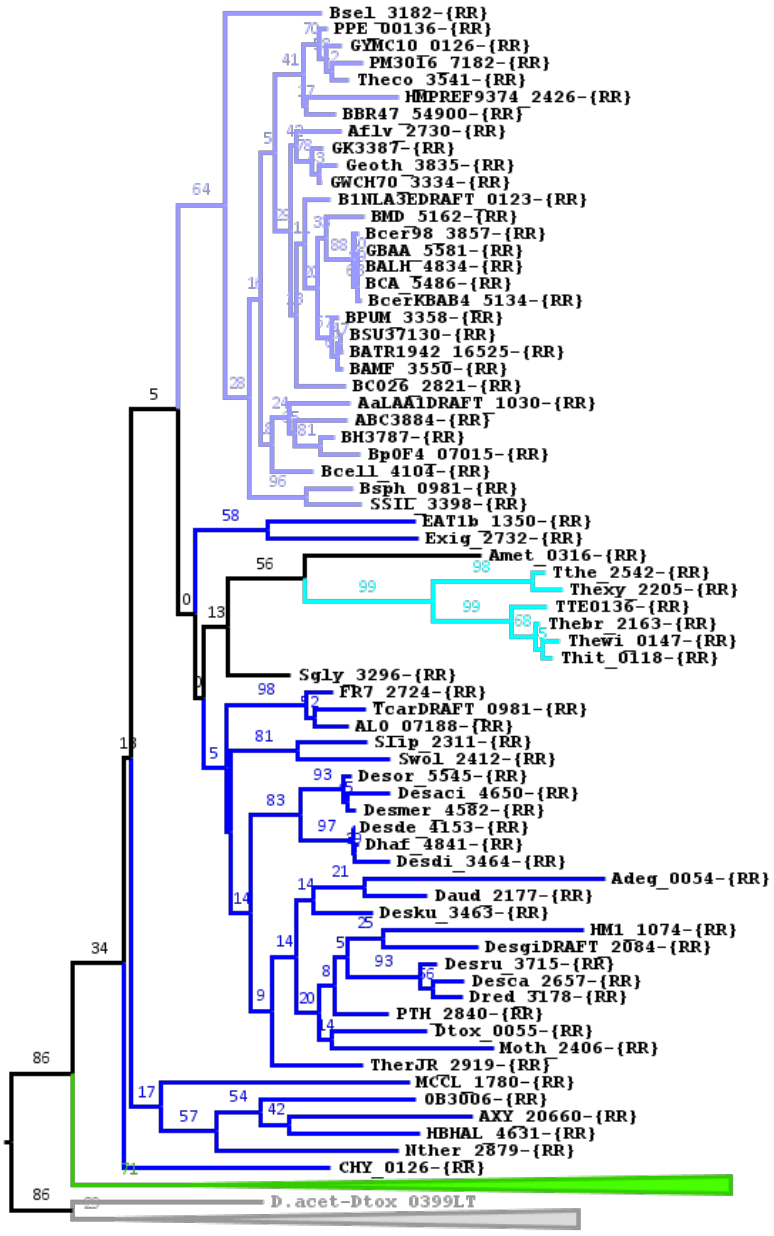


Figure 4.6: .
Detailed view of the Spo0F clade in the gene tree shown in Fig 4.5. Bacillar: light blue; Thermoanaerobacterales sequences: turquoise; all other Firmicutes: dark blue; outgroup: black. Spo0A sequences: green triangle.

See Section 3.4.1.1). This clade is relatively well supported (57). Since the three Bacillales species are all halotolerant, alkalitolerant, or both [Chee and Takami, 2011; Jahns, 1996; Roessler and Müller, 2002], this grouping may be a phylogeny reconstruction artifact.

A number of sequences, including the Spo0A REC sequences of FP2_08180 , RBR_16700, and Ethha_0717 and the Spo0F sequences, Adeg_0054 and HM1_1074, have very long branches and appear to be changing much more rapidly than their neighbors. Interestingly, most of these are reported to be asporogenous, so this may be evidence of degeneration or pseudogenization of these sequences.

Orphan kinases: The number of orphan kinases in the Antunes et al. [2016] data set greatly exceeds the number of Spo0A/F sequences, since most genomes have at most one copy of Spo0F and Spo0A, but several orphan kinases. Phylogenetic noise tends to increase with the number of input sequences. Further, only the interaction and catalytic sequences can be used for molecular phylogenetics, due to the variation in N-terminal sensor domains. Instead of constructing a tree for the full set of kinases, I used a reduced data set consisting of the orphan kinases from four Clostridium species and three Bacillus species, resulting in 33 sequences in total.

A multiple alignment was constructed of the C-terminal amino acid sequences of the kinases, comprising the HisKA and HK_CA domain, using MAFFT [Katoh et al., 2005] with default parameters. Columns possessing gaps in greater than 50% of sequences were trimmed manually. From this alignment, a maximum likelihood phylogeny reconstructed using PhyML [Guindon et al., 2010].

The resulting tree (Fig. 4.8) has two well supported clades, consisting, with one exception, of the Clostridial and Bacillar kinases, respectively. One *C. acetobutylicum* kinase (CA_C3319) clusters with the Bacillus group. In contrast to the REC domain phylogeny, the branching order among the Bacillus kinases is not consistent with evolution by vertical descent from the Firmicutes common ancestor, but rather with a pattern of repeated duplication, followed by specialization, especially in *Bacillus cereus* and *Bacillus anthracis*.

Figure 4.7: Detailed view of the Spo0A clade in the gene tree shown in Fig 4.5. Bacillar: light green; Thermoanaerobacterales sequences: lime ;all other Firmicutes: dark green; outgroup: black. Spo0A sequences: blue triangle.

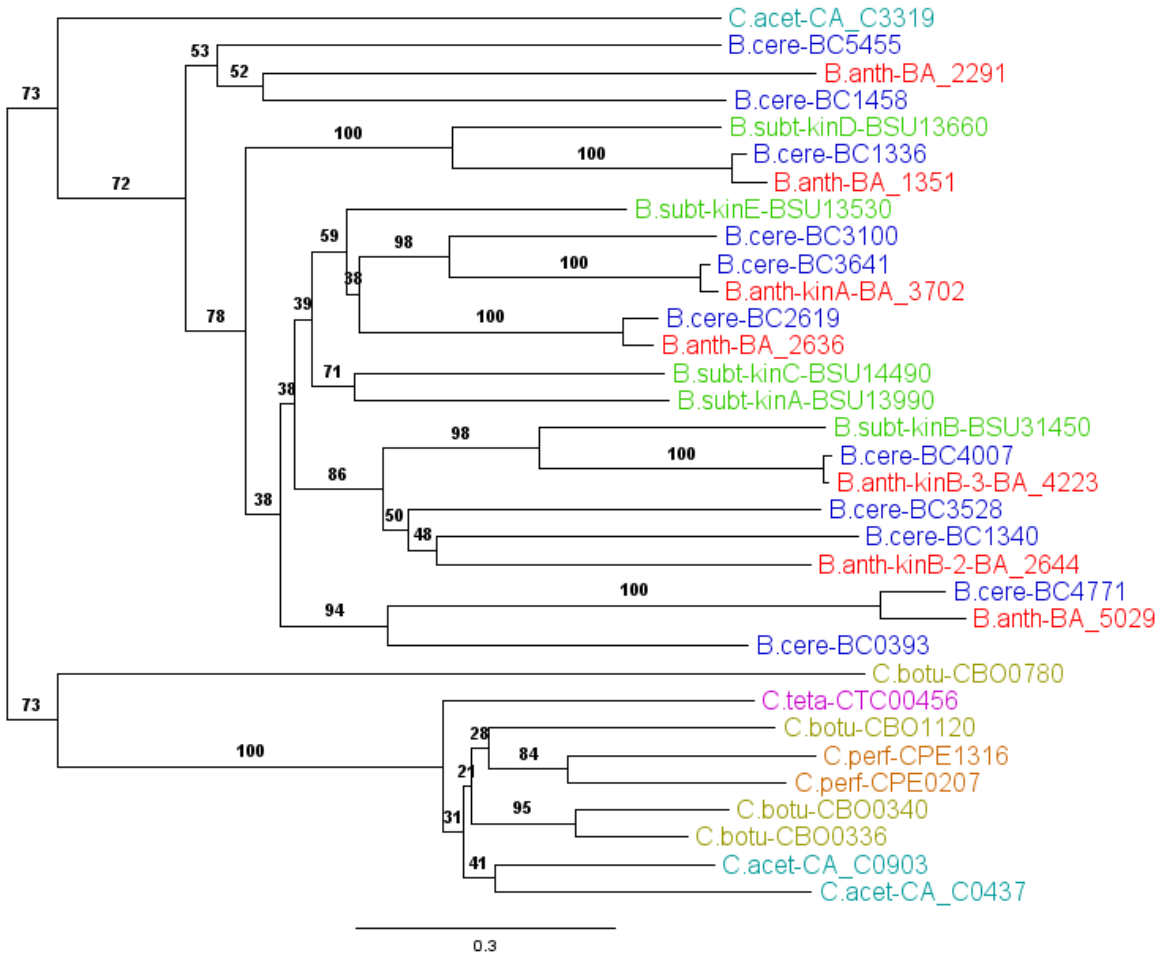


Figure 4.8: Gene tree showing phylogenetic relationships between orphan kinases in *C. anthracis* (B.anth), *B. cereus* (B.cere), *B. subtilis* (B.subt), *C. acetobutylicum* (C.acet), *C. botulinum* (C.botu), *C. perfringens* (C.perf), and *C. tetani* (C.teta). All Clostridium kinases shown, except CA_C3319, harbor a HK_CA:2 catalytic domain. All Bacillar kinases shown and CA_C3319, harbor a HK_CA:3 catalytic domain.

4.3.2. Cross-species experiments

To further investigate the contributions of flexibility and constraint to Spo0 pathway evolution, I sought to characterize the heterologous interactions between Spo0 proteins in a Bacillar phosphorelay (*B. subtilis*), a Clostridial phosphorelay (*D. acetoxidans*), and a Clostridial direct phosphorylation pathway (*C. acetobutylicum*). Each kinase was autophosphorylated and then incubated with each receiver protein (two Spo0F and three Spo0A proteins) in separate reactions. Here, I refer to reactions between components from the same species as autologous and reactions between components across species as heterologous. To test Spo0B interaction connectivity to each Spo0A, Spo0B was incubated with each Spo0A protein in the presence of its autologous kinase and Spo0F. The resulting interactions were characterized as strong or weak (Fig 4.9, summarized in Table 4.3).

The strong interactions observed for all proteins were consistent with their native interaction pattern. Strong phosphorylation of heterologous, as well as autologous, Spo0F proteins was observed for phosphorelay kinases (Fig 4.9C, Lanes 4, 5, 9, and 10), but not for direct phosphorylation kinases (Fig 4.9F). Similarly, we observed strong phosphorylation of heterologous, as well as autologous Spo0A proteins by Spo0B (Fig 4.9A and 6B) and by direct phosphorylation kinases (CA_C0903, Fig 4.9D and CA_C3319, Fig 4.9E), but not by kinases associated with phosphorelays (Bs_KinA: Fig 4.9C, Lanes 1-3; Dtox_1918: Fig 4.9C, Lanes 6-8). These strong heterologous interactions provide evidence that the interaction specificities of Spo0A and Spo0F have remained generally conserved.

Some proteins were also capable of weak heterologous interactions. The phosphorelay kinase, Dtox_1918, although unable to phosphorylate the Spo0A encoded in the same genome (Fig 4.9C, Lane 7), exhibited weak phosphotransfer to both heterologous Spo0As (Fig 4.9C, Lanes 6 and 8). These interactions suggest that there may have been minor shifts in the specificity of Spo0A. In addition, CA_C3319, which directly phosphorylates Spo0A in its native environment, is capable of weak phosphotransfer to both Spo0Fs (Fig 4.9F, Lane 3, 4). The interactions of CA_C3319 demonstrate that there is some degeneracy in the specificity residues that permit phosphotransfer. In the case of each of these two kinases, heterologous combinations break the constraint of the native interaction pattern. I hypothesize that the existence of such interactions presents a mechanism that allows repeated remodeling in the course of evolution of the Spo0 pathway.

	Bs_Spo0F	Dt_Spo0F	Bs_Spo0A	Dt_Spo0A	Ca_Spo0A
Bs_KinA	+	+	-	-	-
Bs_Spo0B	Not tested	Not tested	+	+	+
Dtox_1918	+	+	Weak	-	Weak
Dt_Spo0B	Not tested	Not tested	+	+	+
Ca_C0903	-	-	+	+	+
Ca_C3319	Weak	Weak	+	+	+

Table 4.3: (See also Fig 4.9) involving phosphotransfer from sporulation kinases and phosphotransferases to Spo0F and Spo0A from each architecture. Legend: "+" - interaction,"Weak" weak interaction, "-" no interaction observed. Phosphotransfer was categorized as weak if bands were faint compared to other reactions with the same kinase or Spo0B.

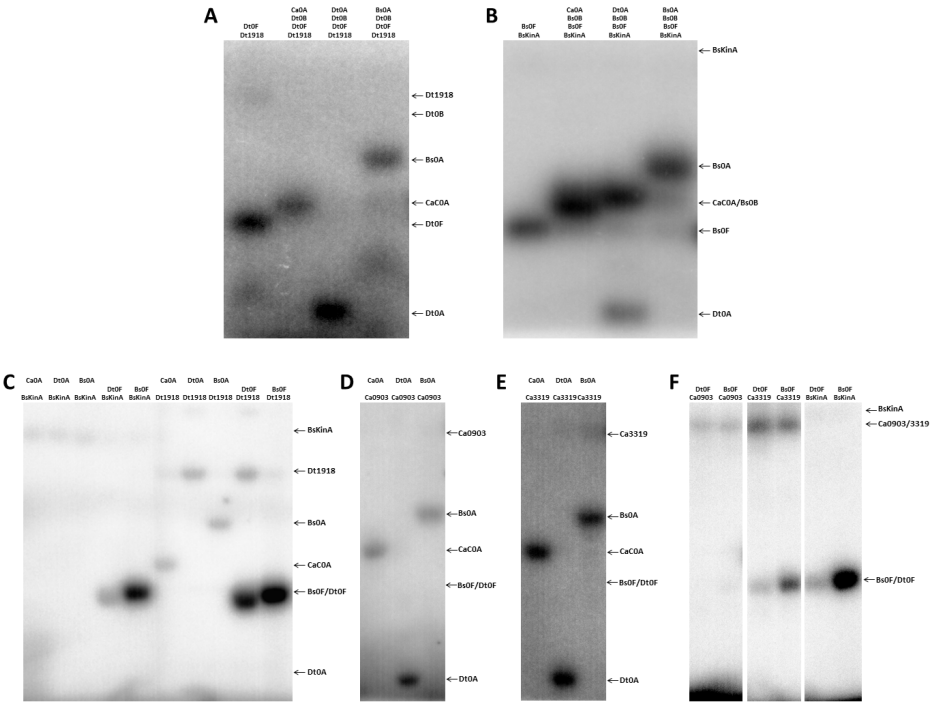


Figure 4.9: Phosphotransfer profiling of Spo0F and Spo0A with phosphorelay and direct phosphorylation architecture components. Reactions contained 5uM sporulation kinase and 10uM of each of the proteins listed above each lane in the phosphorimage. Sporulation kinases autophosphorylated for 10 minutes. All components incubated together for 5 minutes. Kinases include A D. acetoxidans Spo0B B B. subtilis Spo0B C *B. subtilis* KinA and *D. acetoxidans* Dtox_1918 D *C. acetobutylicum* Ca_C0903 E *C. acetobutylicum* Ca_C3319 F Ca_C0903 and Ca_C3319 to Spo0F proteins.

Chapter 5

Discussion

The status of the Spo0 pathway in Class Clostridia has intrigued scientists since the first whole genome sequences of *Clostridium* species became available. Unlike their Bacillar counterparts, there were no apparent homologs to Spo0F and Spo0B, suggesting that the Spo0 pathway had a different architecture. This was later confirmed in a handful of species [Mearls and Lynd, 2014; Steiner et al., 2011; Underwood et al., 2009], in which multiple sensor kinases were found to be specific for Spo0A. Considering that this pathway is simpler and the *Clostridium* have an anaerobic lifestyle more similar to the ancestral Firmicute compared to the species of Class Bacilli, a prevailing hypothesis emerged for the evolution of the Spo0 pathway. Specifically, this prevailing hypothesis suggests that the common ancestral pathway of both the Bacillar phosphorelay and the Clostridial direct phosphorylation architecture also had a direct phosphorylation or two-component like architecture [Durre, 2011, 2014; Talukdar et al., 2015]. Under this hypothesis, it was proposed that the phosphorelay developed within the Bacillar lineage and all phosphorelays are descended from that ancestral Bacillar phosphoerlay [Durre, 2011; Paredes et al., 2005; Stephenson and Lewis, 2005]. Further, it was suggested that the Bacillar phosphorelay was potentially the result of elaboration of the direct phosphorylation architecture [Stephenson and Lewis, 2005] and that this architecture emerged to permit more nuanced control of sporulation during adaptation to oxygenic environments [Durre, 2014; Paredes et al., 2005; Talukdar et al., 2015]. The prevailing hypothesis, therefore, suggested a partitioned Spo0 architecture distribution where all Clostridial spore-formers encoded a Spo0 pathway with a direct phosphorylation architecture and all Bacillar spore-formers encoded a Spo0 pathway with a phosphorelay architecture.

My results challenge this hypothesis. I have identified and presented evidence for the functionality of homologs of phosphorelay Spo0 components in both Classes Bacilli and Clostridia. The

presence of these homologs is indicative of a phosphorelay architecture, thus suggesting that there are Clostridial phosphorelays. This goes against the partitioned architecture distribution predicted by the prevailing hypothesis. Instead, my results suggest a novel hypothesis for the evolution of the Spo0 pathways in the Firmicutes phylum. Specifically, the common ancestral Spo0 pathway had a phosphorelay architecture and the direct phosphorylation architectures are the result of multiple, independent transitions. The experiments that I have performed also provide insight into the potential mechanisms that resulted in these remodeling events. The scenarios I discuss are also indicative of the forces that govern the evolution of pathways with complex architecture that maintain signal fidelity through molecular recognition.

5.1. Repeated, independent remodeling of the Spo0 Pathway

The ancestral phosphorelay hypothesis suggests an alternative evolutionary history of the Spo0 pathway compared to the prevailing theory. Specifically, based on the distribution of Spo0 architectures, this novel hypothesis suggests that there have been multiple, independent remodeling events where a sporulation kinase has become specific for Spo0A and the other phosphorelay proteins were lost. These changes could be mediated through changes in either the sporulation kinases or Spo0A specificity. Here, I analyze the comparison of interaction specificity, which suggest that direct phosphorylation sporulation kinases, which are specific for Spo0A, have altered specificity compared to their phosphorelay counterparts which are specific for Spo0F. Next, I discuss several mechanisms that would result in such a remodeling event mediated by change in sporulation kinase specificity.

5.1.1. Remodeling events were mediated by changes in Spo0 Kinase specificity

According to the ancestral phosphorelay hypothesis, present-day direct phosphorylation pathways are a result of multiple, independent transitions wherein Spo0F and Spo0B were lost and direct phosphorylation of Spo0A was gained. My results support a scenario wherein these transitions arose through changes in or replacement of the kinases. Similarities in the genetic determinants of Spo0A specificity reflect shared taxonomic relationships, not shared pathway architecture, consistent with conservation of Spo0A specificity spectra throughout the phylum. Kinase specificity residues, in contrast, are most similar within the same architecture, consistent with the hypothesis

that changes to sensor kinase, and not Spo0A, specificity are responsible for the change in pathway architecture. This is further demonstrated in the comparison of specificity logos (Fig 4.4).

This observation is strengthened by the results of heterologous phosphotransfer assays (Figs 4.2 and 4.9), summarized in Fig 5.1. If Spo0A specificity is largely unchanged, then phosphodonors that phosphorylate Spo0A in their native environment should also phosphorylate Spo0A proteins from heterologous environments. This is what I observe. When paired with heterologous Spo0 proteins, phosphorelay kinases have a strong preference for Spo0F. In contrast, direct phosphorylation kinases display a strong affinity for Spo0A from both architectures and not for heterologous Spo0F. Spo0B proteins are also able to phosphorylate Spo0A from both architectures. The strong interactions I observe between both autologous and heterologous pairs are consistent with major differences in the specificity of the phosphorelay and direct phosphorylation architecture kinases tested. They also suggest that little change in Spo0A specificity has occurred. However, weak heterologous interactions involving the *D. acetoxidans* kinase, Dtox_1918, indicate minor and lineage specific differences in the specificity of the *D. acetoxidans* Spo0A compared to the other two proteins tested.

5.1.2. Possible mechanisms of histidine kinase-based signaling architecture remodeling

My results suggest that each transition to a direct phosphorylation architecture was mediated by specificity changes in the set of sporulation kinases. Here, I highlight several scenarios (Fig 5.2) in which changes in sporulation kinase specificity could result in acquisition of direct phosphorylation of Spo0A. One possibility is that substitutions in an autologous sporulation kinase resulted in a loss of specificity for Spo0F and gain of specificity for Spo0A (Fig 5.2A). Given the requirement that Spo0F and Spo0A specificity spectra must overlap, only a few substitutions may be required. Alternatively, an autologous hybrid histidine kinase could encode a HisKA domain with pre-existing specificity for Spo0A (Fig 5.2B), as the REC domain of a hybrid kinase insulates it from interaction with non-cognate receivers [Capra et al., 2012a]. Loss of the REC domain would result in direct phosphorylation of Spo0A.

Acquisition, via horizontal transfer, of a novel kinase already possessing specificity for Spo0A would result in immediate remodeling to a direct phosphorylation architecture (Fig 5.2C and D). This scenario requires that a kinase encoded in a different species be able to phosphorylate the local Spo0A. This could occur if the donor were a non-sporulating species, in which the interaction spectra associated with Spo0A in spore-formers were occupied by the receiver from an unrelated pathway (Fig 5.2C). This could also occur if the donor were a spore-former, but due to minor shifts

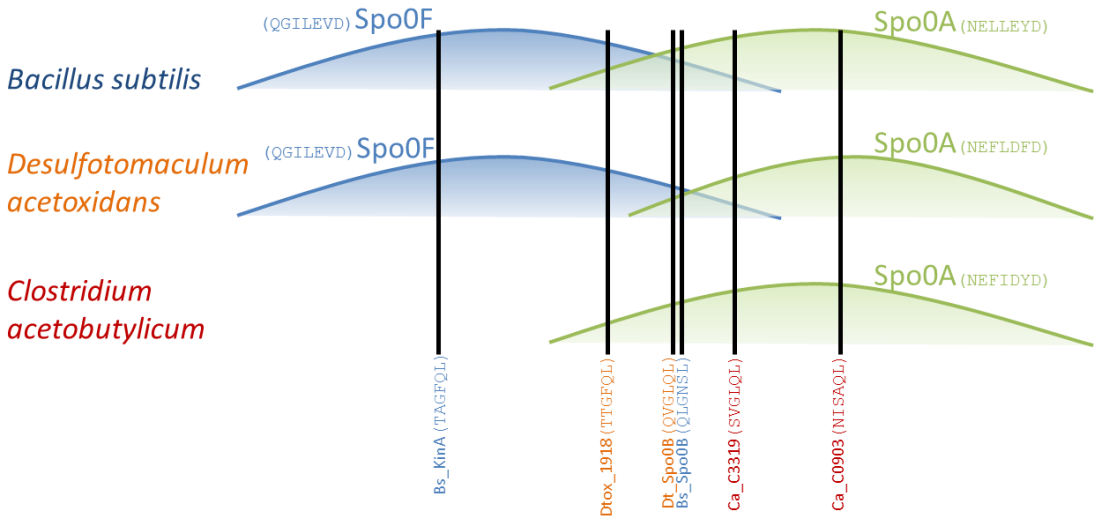


Figure 5.1: Specificity spectra schematic summarizing autologous and heterologous interactions between proteins from three Spo0 pathways were tested: *B. subtilis* phosphorelay, *D. acetoxidans* phosphorelay, and *C. acetobutylicum* direct phosphorylation architecture (Figs 4.2 and 4.9). In this qualitative representation, each phosphodonor corresponds to one point on the x-axis; x-axis is shared between cells. Each line corresponds to the point in specificity space a phosphodonor occupies. The interactions observed, therefore, provide partial information about the location of receiver specificity spectra for Spo0F (blue spectra) and Spo0A (green spectra) in each genome. Both Spo0B exhibited strong interactions with all receivers tested, confirming that autologous Spo0F and Spo0A spectra overlap. The heterologous interactions of Spo0B and these receivers further suggests that this overlap is in the same area of specificity spectra in each genome and that *C. acetobutylicum* Spo0A also shares part of its specificity spectrum with this overlap. The interactions of Bs_KinA and Ca_C0903 suggest that the Spo0F and Spo0A spectra do not overlap completely. As Bs_kinA interacts with all Spo0F and no Spo0A proteins, there must exist a region of each Spo0F spectrum that does not overlap with that of Spo0A, and vice versa for the interactions observed with Ca_C0903. Ca_C3319 interacts strongly with all Spo0A and weakly with all Spo0F. Dtox_1918 interacts strongly with each Spo0F and weakly with heterologous Spo0A proteins, suggesting that the *D. acetoxidans* Spo0A specificity spectrum has shifted compared to that of *B. subtilis* and *C. acetobutylicum*. These weak interactions are not targeted by selection as they are heterologous, though they provide insight into the relative locations of the Spo0F and Spo0A specificity spectra.

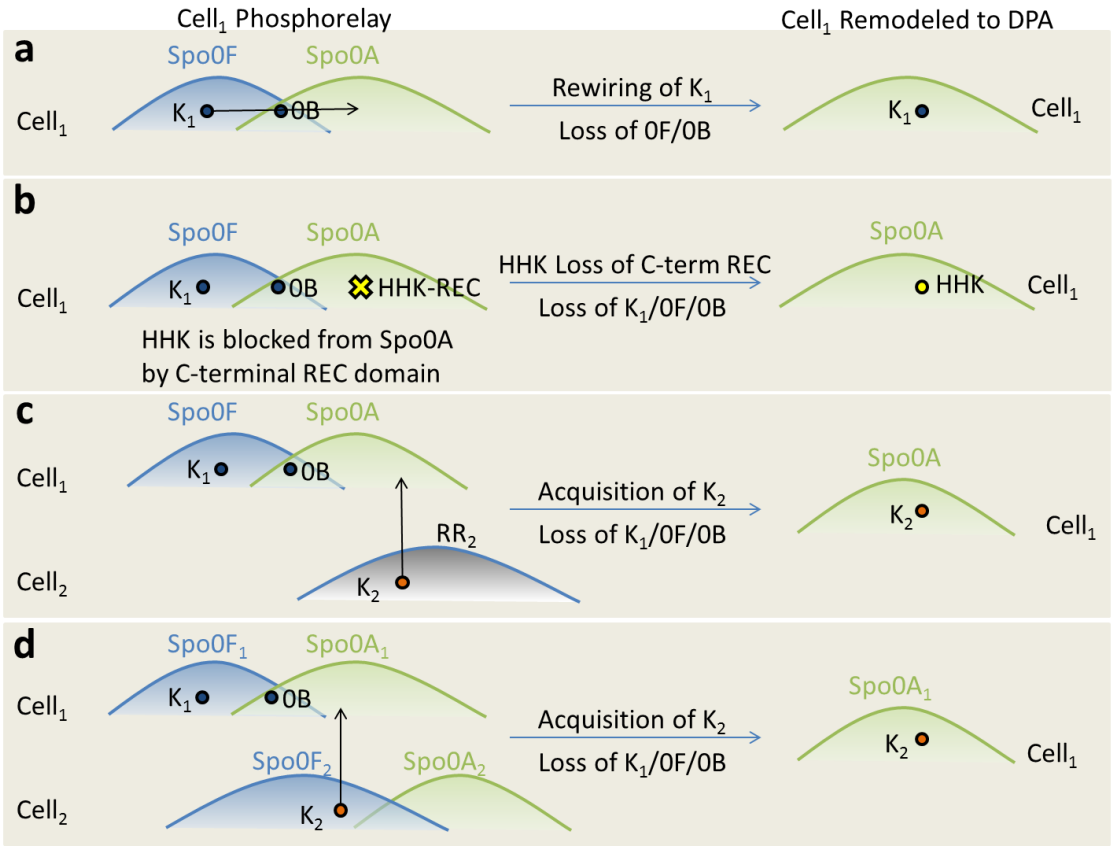


Figure 5.2: Specificity spectra for scenarios leading to remodeling of the Cell1 phosphorelay to a direct phosphorylation architecture. (A) Sporulation kinase K1 accrues specificity residue substitutions resulting in a change of specificity from Spo0F to Spo0A. Subsequently, Spo0F and Spo0B are lost. (B) A hybrid histidine kinase (HHK-REC) is insulated from interaction with Spo0A by spatial tethering. Subsequently, the REC domain of the hybrid histidine kinase is lost and interaction with Spo0A is gained; K1, Spo0F, and Spo0B are also lost. (C) Sensor kinase K2 has specificity for the Spo0A of Cell1. Acquisition of K2 by horizontal gene transfer results in a direct phosphorylation architecture. Subsequently, K1, Spo0F, and Spo0B are lost. (D) The specificity spectra for the Spo0 phosphorelay of Cell2 is shifted compared to that of Cell1 such that sporulation kinase K2 has specificity for Cell1 Spo0A. Acquisition of K2 by horizontal gene transfer results in a direct phosphorylation architecture. Subsequently, K1, Spo0F, and Spo0B are lost.

in the specificity spectra, the transferred kinase was insulated from Spo0A in its own cell, but within the Spo0A specificity spectrum of the recipient (Fig 5.2D). The weak interactions observed between *D. acetoxidans* Dtox_1918 and heterologous Spo0A proteins suggest that such shifts in the specificity spectra of Spo0F and Spo0A do occur.

Following the first remodeling event that resulted in a kinase that phosphorylates Spo0A di-

rectly, it is possible that subsequent remodeling events occurred through the acquisition of that kinase by horizontal transfer. The consistency with which I observe HK_CA:2 catalytic domains in direct phosphorylation architectures suggests that transfer may be a method of propagation of the direct phosphorylation architectures. However, varied specificity residue signatures and the differences between clades suggests that the kinases that mediated each transition are not all related. The relationship of these kinases remains an intriguing question.

The *C. acetobutylicum* sporulation kinase, CA_C3319, which exhibited weak affinity for heterologous Spo0F proteins (Fig 4.9), may be an example of this last scenario (Fig 5.2D). CA_C3319 harbors a HK_CA:3 type catalytic domain, which is commonly observed in phosphorelay, but not two-component, sporulation kinases. Further, it possesses unusual specificity residues (SVGLQL) that do not match the typical signatures of either architecture. These distinct characteristics suggest that CA_C3319 could be a recently acquired phosphorelay kinase that was specific for Spo0F in the donor cell. Upon acquisition, it may have interacted weakly with Spo0A initially, as there is no Spo0F present in *C. acetobutylicum*, and subsequently evolved a stronger preference for Spo0A.

5.2. Propensity for remodeling

Our results suggest an evolutionary history wherein remodeling of an ancient phosphorelay resulted in a simpler, two-component signal transduction pathway. This is consistent with recent theories of reductive genome evolution, which posit that present-day species with streamlined genomes evolve from gene-rich ancestors via a process of specialization [Csuros and Miklos, 2009; Wolf and Koonin, 2013]. The observation of repeated, independent episodes of pathway remodeling may indicate that the Spo0 pathway has a particular susceptibility for this type of reorganization.

The propensity for pathway remodeling may result from juxtaposition of the particular interaction requirements of the Spo0 phosphorelay and the ecological role of the phenotype that it controls. The specificity of the spectra of Spo0F and Spo0A must intersect to some extent, since both interact with Spo0B. Given their proximity in interaction space, the mutational trajectories required to lose interaction with Spo0F and gain direct interaction with Spo0A may be short. Further, since sporulation is only essential in survival conditions, selection acting on these mutational trajectories may be relatively permissive. Thus, pathway remodeling via substitutions that change interaction specificity may arise easily.

A second mechanism of pathway remodeling, by acquisition of a foreign kinase with specificity for Spo0A, may be a byproduct of adaptation to changing environments, since acquisition of novel

sensor kinases is a source of novel signal recognition capabilities. The diversity of environmental conditions that induce sporulation in various taxa [Auchtung and Grossman, 2008], as well as the diversity of lineage specific sporulation kinase repertoires [Stephenson and Lewis, 2005] (see also Tables A2 and A9, Fig 4.8), are both consistent with a process of ongoing, lineage-specific turnover of sporulation kinases. Pathway remodeling via acquisition of novel kinases could also be linked to the loss and recovery of the spore formation phenotype. Sporulation is a metabolically expensive process and is lost frequently in stable conditions [Maughan et al., 2009]. Loss of Spo0F or Spo0B is one scenario that would result in loss of sporulation. If environmental conditions subsequently became less favorable, acquisition of a kinase with specificity to Spo0A would restore sporulation, albeit with a direct phosphorylation architecture. Indeed, several *Clostridium sensu strictu* species, which likely encode a direct phosphorylation pathway, nevertheless possess Spo0B (see Section 3.4) as might be expected in this scenario. Further, we observe that clades harboring direct phosphorylation architectures tend to encode a mix of spore-formers and non-spore-formers, which is consistent with the hypothesis that Spo0 pathway remodeling is linked to loss of sporulation.

5.2.0.1. Signal transduction via molecular recognition versus spatial tethering

What we have learned about the Spo0 phosphorelay provides some general design principles for histidine-aspartate pathways, in which a single protein must interact with multiple partners and specificity is enforced by molecular recognition. It also provides a perspective on the properties that distinguish the Spo0 phosphorelay from other phosphorelays. Sporulation is initiated by multi-input pathways in which each step in the cascade is encoded in a separate protein, requiring that interaction specificity be controlled entirely by molecular recognition. In most of the phosphorelays that have been studied, two or more of the four interaction domains are encoded in the same protein, such that interaction specificity is controlled by spatial tethering [Capra et al., 2012a]. Signal transduction based on hybrid kinases may be more robust, but less easily reconfigured or expanded. The differences between spatial tethering and molecular recognition in a phosphorelay could represent different trade-offs between flexibility and constraint

5.3. Future Work

5.3.1. On the origins of the phosphorelay

In the context of the ancestral direct phosphorylation hypothesis, it was proposed that the acquisition of a phosphorelay architecture in the emerging Bacilli during the Great Oxidation Event

(GOE) was an "adaptation to an aerobic lifestyle required enhanced sensing and regulating capabilities, which probably resulted in development of the phosphorelay" [Durre, 2014]. The ancestral phosphorelay hypothesis predicts an earlier origin for the Spo0 phosphorelay, suggesting that it did not originate as an adaptation to the rise of oxygen during the GOE. More nuanced regulation of sporulation may, indeed, have been a fitness benefit of the original Spo0 phosphorelay, but the selective pressures in that early environment remain unclear. It is also possible that more nuanced regulation offered by the additional control points in the phosphorelay architecture may be beneficial in fluctuating oxygen levels in present-day environments, such as biofilms [Paredes et al., 2005].

Further, the scope of the taxon sampling in this study only permits the hypothesis that the phosphorelay was present in the ancestor of all Firmicutes. Where did this pathway actually develop? Further examination of the presence of conserved neighborhoods of Spo0F and Spo0B may contain evidence for the development of these intermediate phosphorelay components even in their absence. Even in the absence of ORFs resembling ancestral Spo0F and Spo0B in non-firmicutes, there may be evidence of loss of such proteins in the non-coding regions of these conserved neighborhoods. However, this also poses the question, is there a fitness advantage that selects for Spo0F and Spo0B to remain in the conserved neighborhoods we identified?

Additional computational study may also reveal the history of Spo0B. The similarities between the interaction domains of citrate and malate sensing histidine kinases and Spo0B is intriguing. A comprehensive phylogenetic analysis involving such kinases from non-firmicutes may reveal a detailed evolutionary history of Spo0B developed. However, the degenerated nature of Spo0B compared to those kinases may have eroded the phylogenetic information required to develop such a hypothesis.

Finally, I note that it has been suggested as part of the prevailing views of the evolutionary history of the Spo0 pathway that the phosphorelay is the result of the duplication of a TCS pair, followed by modification of domain content and inclusion into the phosphorelay. Based on the similarities of Spo0F and Spo0A REC domains, this hypothesis further suggests that the duplicated TCS pair may have involved Spo0A. If this is true, Spo0F homologs should form a monophyletic subtree within the Spo0A REC domain tree. However, the Spo0F and Spo0A REC domains form two separate, well-supported clades, refuting this hypothesis.

5.3.2. Spo0 Interaction Specificity

A rational rewiring approach, similar to that reported by Podgornaia and Laub [2015], could be used to characterize the global specificity spectra of Spo0 pathways. This experiment would pro-

vide strong evidence of which specificity residue signatures facilitate phosphorelay or direct phosphorylation of Spo0A and allow for greater accuracy in architecture prediction. The information obtained from such a study would be valuable to a refinement of the hypotheses presented and tested here.

Another intriguing aspect not considered in this study, is the possible involvement of hybrid histidine kinases and hybrid response regulators in Spo0 pathways. I did not consider hybrid histidine kinases as potential sporulation kinases in this study as none had been implicated in any Spo0 pathway in prior work. However, the identification of a hybrid response regulator as an essential protein for sporulation in *Ruminiclostridium thermocellum* suggests that there may be other hybrid histidine-aspartate phosphotransfer proteins involved in sporulation.

Finally, some evolutionary studies reveal transition states between two phenotypes [Tsong et al., 2003]. I envision that, if such a transition state exists for this pathway, it would likely take the form of a species that encodes Spo0F and Spo0B but is either additionally or only able to signal the initiation by phosphorylation through direct phosphotransfer of Spo0A. Outliers in the comparison of specificity residues of candidate sporulation kinases suggest some candidate organisms that may be in the process of remodeling from a phosphorelay to a direct phosphorylation architecture, including *Gottschalkia acidurici*, *Alkaliphilus metalliredigens*, and species from the Paenibacillaceae family.

Bibliography

A. Abecasis, M. Serrano, R. Alves, L. Quintais, J. Pereira-Leal, and A. Henriques. A genomic signature and the identification of new sporulation genes. *J Bacteriol*, 195:2101–2115, May 2013.

I. Ahmed, A. Yokota, A. Yamazoe, and T. Fujiwara. Proposal of *Lysinibacillus boronitolerans* gen. nov. sp. nov., and transfer of *Bacillus fusiformis* to *Lysinibacillus fusiformis* comb. nov. and *Bacillus sphaericus* to *Lysinibacillus sphaericus* comb. nov. *Int J Syst Evol Microbiol*, 57: 1117–1125, May 2007.

G. Ahnert-Hilger and H. Bigalke. Molecular aspects of tetanus and botulinum neurotoxin poisoning. *Prog Neurobiol*, 46:83–96, May 1995.

M. Al-Hinai, S. Jones, and E. Papoutsakis. The *Clostridium* sporulation programs: diversity and preservation of endospore differentiation. *Microbiol Mol Biol Rev*, 79:19–37, Mar 2015.

R. P. Alexander and I. B. Zhulin. Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors. *Proc Natl Acad Sci U S A*, 104(8): 2885–90, Feb 20 2007.

S. D. Allen, C. L. Emery, and D. M. Lyerly. *Clostridium in Manual of Clinical Microbiology*, pages 835–856. American Society for Microbiology, Washington D.C., 8th edition, 2003.

E. Alm, K. Huang, and A. Arkin. The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLoS Comput Biol*, 2:e143, Nov 2006.

S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.

C. Antoniewski, B. Savelli, and P. Stragier. The spoIIJ gene, which regulates early developmental steps in *Bacillus subtilis*, belongs to a class of environmentally responsive genes. *J Bacteriol*, 172:86–93, Jan 1990.

L. C. Antunes, D. Poppleton, A. Klingl, A. Criscuolo, B. Dupuy, C. Brochier-Armanet, C. Beloin, and S. Gribaldo. Phylogenomic analysis supports the ancestral presence of LPS-outer membranes in the Firmicutes. *Elife*, 5, Aug 31 2016.

M. Asayama, A. Yamamoto, and Y. Kobayashi. Dimer form of phosphorylated Spo0A, a transcriptional regulator, stimulates the spo0F transcription at the initiation of sporulation in *Bacillus subtilis*. *J Mol Biol*, 250:11–23, Jun 1995.

J. Auchtung and A. Grossman. Extracellular peptide signaling and quorum responses in development, self-recognition, and horizontal gene transfer in *Bacillus subtilis*. In Stephen Winans and Bonnie Bassler, editors, *Chemical communication among bacteria*, chapter 2, pages 13–30. ASM Press, 2008.

M. Austin, L. Rabe, S. Srinivasan, D. Fredricks, H. Wiesenfeld, and S. Hillier. *Mageeibacillus indolicus* gen. nov., sp. nov.: a novel bacterium isolated from the female genital tract. *Anaerobe*, 32:37–42, Apr 2015.

E. Baril, L. Coroller, O. Couvert, M. El Jabri, I. Leguerinel, F. Postollec, C. Boulais, F. Carlin, and P. Mafart. Sporulation boundaries and spore formation kinetics of *Bacillus* spp. as a function of temperature, pH and a(w). *Food Microbiol*, 32:79–86, Oct 2012.

F. U. Battistuzzi, A. Feijao, and S. B. Hedges. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol*, 4:44, Nov 9 2004.

M. Ben Abdallah, F. Karray, N. Mhiri, J. L. Cayol, J. L. Tholozan, D. Alazard, and S. Savadi. Characterization of *Sporohalobacter salinus* sp. nov., an anaerobic, halophilic, fermentative bacterium isolated from a hypersaline lake. *Int J Syst Evol Microbiol*, 65(Pt 2):543–8, Feb 2015.

M. J. Bick, V. Lamour, K. R. Rajashankar, Y. Gordiyenko, C. V. Robinson, and S. A. Darst. How to switch off a histidine kinase: crystal structure of *Geobacillus stearothermophilus* KinB with the inhibitor Sda. *J Mol Biol*, 386(1):163–77, Feb 13 2009.

E. Biondi, J. Skerker, M. Arif, M. Prasol, B. Perchuk, and M. Laub. A phosphorelay system controls stalk biogenesis during cell cycle progression in *Caulobacter crescentus*. *Mol Microbiol*, 59:386–401, Jan 2006.

B. Bravo-Ferrada, A. Gómez-Zavaglia, L. Semorile, and E. Tymczyszyn. Effect of the fatty acid composition of acclimated oenological *Lactobacillus plantarum* on the resistance to ethanol. *Lett Appl Microbiol*, 60:155–161, Feb 2015.

H. P. Browne, S. C. Forster, B. O. Anonye, N. Kumar, B. A. Neville, M. D. Stares, D. Goulding, and T. D. Lawley. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature*, 533(7604):543–6, May 26 2016.

H. Bruggemann, S. Baumer, W. Fricke, A. Wiezer, H. Liesegang, I. Decker, C. Herzberg, R. Martinez-Arias, R. Merkl, A. Henne, and G. Gottschalk. The genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. *Proc Natl Acad Sci U S A*, 100:1316–1321, Feb 2003.

R. Brunsing, C. La Clair, S. Tang, C. Chiang, L. Hancock, M. Perego, and J. Hoch. Characterization of sporulation histidine kinases of *Bacillus anthracis*. *J Bacteriol*, 187:6972–6981, Oct 2005.

D. Burbulys, K. Trach, and J. Hoch. Initiation of sporulation in *B. subtilis* is controlled by a multicomponent phosphorelay. *Cell*, 64:545–552, Feb 1991.

D. Burns and N. Minton. Sporulation studies in *Clostridium difficile*. *J Microbiol Methods*, 87: 133–138, Nov 2011.

D. Byrer, F. Rainey, and J. Wiegel. Novel strains of *Moorella thermoacetica* form unusually heat-resistant spores. *Arch Microbiol*, 174:334–339, Nov 2000.

X. Cao, X. Liu, and X. Dong. *Alkaliphilus crotonatoxidans* sp. nov., a strictly anaerobic, crotonate-dismutating bacterium isolated from a methanogenic environment. *Int J Syst Evol Microbiol*, 53 (Pt 4):971–5, Jul 2003.

E. Capra, B. Perchuk, E. Lubin, O. Ashenberg, J. Skerker, and M. Laub. Systematic dissection and trajectory-scanning mutagenesis of the molecular interface that ensures specificity of two-component signaling pathways. *PLoS Genet*, 6:e1001220, Nov 2010.

E. Capra, B. Perchuk, O. Ashenberg, C. Seid, H. Snow, J. Skerker, and M. Laub. Spatial tethering of kinases to their substrates relaxes evolutionary constraints on specificity. *Mol Microbiol*, 86: 1393–1403, Dec 2012a.

E. Capra, B. Perchuk, J. Skerker, and M. Laub. Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. *Cell*, 150:222–232, Jul 2012b.

P. Casino, V. Rubio, and A. Marina. Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. *Cell*, 139:325–336, Oct 2009.

P. Casino, V. Rubio, and A. Marina. The mechanism of signal transduction by two-component systems. *Curr Opin Struct Biol*, 20:763–771, Dec 2010.

N. Catlett, O. Yoder, and B. Turgeon. Whole-genome analysis of two-component signal transduction genes in fungal pathogens. *Eukaryot Cell*, 2:1151–1161, Dec 2003.

Y. Chang, R. Pukall, E. Saunders, A. Lapidus, A. Copeland, M. Nolan, T. Glavina Del Rio, S. Lucas, F. Chen, H. Tice, J. Cheng, C. Han, J. Detter, D. Bruce, L. Goodwin, S. Pitluck, N. Mikhailova, K. Liolios, A. Pati, N. Ivanova, K. Mavromatis, A. Chen, K. Palaniappan, M. Land, L. Hauser, C. Jeffries, T. Brettin, M. Rohde, M. Göker, J. Bristow, J. Eisen, V. Markowitz, P. Hugenholtz, N. Kyrpides, and H. Klenk. Complete genome sequence of *Acidaminococcus fermentans* type strain (vr4). *Stand Genomic Sci*, 3:1–14, Jul 2010.

G. Chee and H. Takami. Alternative splicing by participation of the group II intron ORF in extremely halotolerant and alkaliphilic *Oceanobacillus iheyensis*. *Microbes Environ*, 26:54–60, 2011.

D. Chivian, E. Brodie, E. Alm, D. Culley, P. Dehal, T. DeSantis, T. Gihring, A. Lapidus, L. Lin, S. Lowry, D. Moser, P. Richardson, G. Southam, G. Wanger, L. Pratt, G. Andersen, T. Hazen, F. Brockman, A. Arkin, and T. Onstott. Environmental genomics reveals a single-species ecosystem deep within earth. *Science*, 322:275–278, Oct 2008.

M. Clarke, D. Hughes, C. Zhu, E. Boedeker, and V. Sperandio. The QseC sensor kinase: a bacterial adrenergic receptor. *Proc Natl Acad Sci U S A*, 103:10420–10425, Jul 2006.

M. D. Collins, P. A. Lawson, A. Willems, J. J. Cordoba, J. Fernandez-Garayzabal, P. Garcia, Cai J., H. Hippe, and Farrow J. A. The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. *Int J Syst Bacteriol*, 44(4):812–26, Oct 1994.

G. Crooks, G. Hon, J. Chandonia, and S. Brenner. Weblogo: a sequence logo generator. *Genome Res*, 14(6):1188–90, 2004.

M. Csuros and I. Miklos. Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. *Mol Biol Evol*, 26:2087–2095, Sep 2009.

C. A. Cummins and J. O. McInerney. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. *Syst Biol*, 60(6):833–44, Dec 2011.

E. Dalla Vecchia, M. Visser, A. Stams, and R. Bernier-Latmani. Investigation of sporulation in the *Desulfotomaculum* genus: a genomic comparison with the genera *Bacillus* and *Clostridium*. *Environ Microbiol Rep*, 6:756–766, Dec 2014.

M. de Hoon, P. Eichenberger, and D. Vitkup. Hierarchical evolution of the bacterial sporulation network. *Curr Biol*, 20:R735–R745, Sep 2010.

A. Diaz, L. Core, M. Jiang, M. Morelli, C. Chiang, H. Szurmant, and M. Perego. *Bacillus subtilis* RapA phosphatase domain interaction with its substrate, phosphorylated Spo0F, and its inhibitor, the PhrA peptide. *J Bacteriol*, 194:1378–1388, Mar 2012.

E. Dubnau, J. Weir, G. Nair, L. Carter, C. Moran, and I. Smith. *Bacillus* sporulation gene spo0H codes for sigma 30 (sigma H). *J Bacteriol*, 170(3):1054–1062, Mar 1988.

P. Durre. Ancestral sporulation initiation. *Mol Microbiol*, 80:584–587, May 2011.

P. Durre. Physiology and sporulation in *Clostridium*. *Microbiol Spectr*, 2(4):TBS–0010–2012, Aug 2014.

S. Eddy. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput Biol*, 4:e1000069, May 2008.

M. Eppinger, B. Bunk, M. Johns, J. Edirisinghe, K. Kutumbaka, S. Koenig, H. Creasy, M. Rosovitz, D. Riley, S. Daugherty, M. Martin, L. Elbourne, I. Paulsen, R. Biedendieck, C. Braun, S. Grayburn, S. Dhingra, V. Lukyanchuk, B. Ball, R. Ul-Qamar, J. Seibel, E. Bremer, D. Jahn, J. Ravel, and P. Vary. Genome sequences of the biotechnologically important *Bacillus megaterium* strains QM B1551 and DSM319. *J Bacteriol*, 193:4199–4213, Aug 2011.

F. Ferrari, K. Trach, and J. Hoch. Sequence analysis of the spo0B locus reveals a polycistronic transcription unit. *J Bacteriol*, 161(2):556–562, Feb 1985a.

F. Ferrari, K. Trach, D. LeCoq, J. Spence, E. Ferrari, and J. Hoch. Characterization of the spo0A locus and its deduced product. *Proc Natl Acad Sci U S A*, 82(9):2647–2651, May 1985b.

J. Flint, D. Drzymalski, W. Montgomery, G. Southam, and E. Angert. Nocturnal production of endospores in natural populations of epulopiscium-like surgeonfish symbionts. *J Bacteriol*, 187 (21):7460–7470, Nov 2005.

M. Francis, C. Allen, R. Shrestha, and J. Sorg. Bile acid recognition by the *Clostridium difficile* germinant receptor, CspC, is important for establishing infection. *PLoS Pathog*, 9:e1003356, May 2013.

D. Freier, C. Mothershed, and J. Wiegel. Characterization of *Clostridium thermocellum* JW20. *Appl Environ Microbiol*, 54:204–211, Jan 1988.

M. Fuhrmann, A. Hausherr, L. Ferbitz, T. Schödl, M. Heitzer, and P. Hegemann. Monitoring dynamic expression of nuclear genes in *Chlamydomonas reinhardtii* by using a synthetic luciferase reporter gene. *Plant Mol Biol*, 55(6):869–881, Aug 2004.

M. Galperin. A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts. *BMC Microbiol*, 5:35, Jun 2005.

M. Galperin. A square archaeon, the smallest eukaryote and the largest bacteria. *Environ Microbiol*, 8:1683–1687, Oct 2006.

M. Galperin. Genome diversity of spore-forming Firmicutes. *Microbiol Spectr*, 1, Dec 2013.

M. Galperin, S. Mekhedov, P. Puigbo, S. Smirnov, Y. Wolf, and D. Rigden. Genomic determinants of sporulation in Bacilli and Clostridia: towards the minimal set of sporulation-specific genes. *Environ Microbiol*, Jul 2012.

M. Y. Galperin, V. Brover, I. Tolstoy, and N. Yutin. Phylogenomic analysis of the family Peptostreptococcaceae (*Clostridium* cluster XI) and proposal for reclassification of *Clostridium litorale* (Fendrich et al. 1991) and *Eubacterium acidaminophilum* (Zindel et al. 1989) as *Peptoclostridium litorale* gen. nov. comb. nov. and *Peptoclostridium acidaminophilum* comb. nov. *Int J Syst Evol Microbiol*, 66(12):5506–5513, Dec 2016.

C. Grimshaw, S. Huang, C. Hanstein, M. Strauch, D. Burbulys, L. Wang, J. Hoch, and J. Whiteley. Synergistic kinetic interactions between components of the phosphorelay controlling sporulation in *Bacillus subtilis*. *Biochemistry*, 37:1365–1375, Feb 1998.

S. Gronow, S. Welnitz, A. Lapidus, M. Nolan, N. Ivanova, T. Glavina Del Rio, A. Copeland, F. Chen, H. Tice, S. Pitluck, J. Cheng, E. Saunders, T. Brettin, C. Han, J. Detter, D. Bruce, L. Goodwin, M. Land, L. Hauser, Y. Chang, C. Jeffries, A. Pati, K. Mavromatis, N. Mikhailova, A. Chen, K. Palaniappan, P. Chain, M. Rohde, M. Göker, J. Bristow, J. Eisen, V. Markowitz, P. Hugenholtz, N. Kyrpides, H. Klenk, and S. Lucas. Complete genome sequence of *Veillonella parvula* type strain (Te3). *Stand Genomic Sci*, 2:57–65, Jan 2010.

S. Guindon, J. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*, 59:307–321, May 2010.

C. Han, W. Gu, X. Zhang, A. Lapidus, M. Nolan, A. Copeland, S. Lucas, T. G. Del Rio, H. Tice, JF. Cheng, R. Tapia, L. Goodwin, S. Pitluck, I. Pagani, N. Ivanova, K. Mavromatis, N. Mikhailova, A. Pati, A. Chen, K. Palaniappan, M. Land, L. Hauser, YJ. Chang, C. D. Jeffries, S. Schneider, M. Rohde, M. Goker, R. Pukall, T. Woyke, J. Bristow, JA. Eisen, V. Markowitz, P. Hugenholtz, N. C. Kyrpides, HP. Klenk, and J. C. Detter. Complete genome sequence of *Thermaerobacter marianensis* type strain (7p75a). *Stand Genomic Sci*, 3(3):337–45, Dec 15 2010.

K. Hartwich, A. Poehlein, and R. Daniel. The purine-utilizing bacterium *Clostridium acidurici* 9a: a genome-guided metabolic reconsideration. *PLoS One*, 7:e51662, Dec 2012.

S. Higaki, T. Kitagawa, M. Kagoura, M. Morohashi, and T. Yamagishi. Characterization of *Peptostreptococcus* species in skin infections. *J Int Med Res*, 28:143–147, May/Jun 2000.

D. Higgins and J. Dworkin. Recent progress in *Bacillus subtilis* sporulation. *FEMS Microbiol Rev*, 36:131–148, Jan 2012.

J. Hoch. Genetics of bacterial sporulation. *Adv Genet*, 18:69–98, 1976.

J. Hoch, K. Trach, F. Kawamura, and H. Saito. Identification of the transcriptional suppressor sof-1 as an alteration in the spo0A protein. *J Bacteriol*, 161(2):552–555, Feb 1985.

J. A. Hoch. Genetic analysis of pleiotropic negative sporulation mutants in *Bacillus subtilis*. *J Bacteriol*, 105(3):896–901, Mar 1971.

R. Huber, P. Rossnagel, C. R. Woese, R. Rachel, T. A. Langworthy, and K. O. Stetter. Formation of ammonium from nitrate during chemolithoautotrophic growth of the extremely thermophilic bacterium *Ammonifex degensii* gen. nov. sp. nov. *Syst Appl Microbiol*, 19:40–49, Mar 1996.

H. Imachi, Y. Sekiguchi, Y. Kamagata, S. Hanada, A. Ohashi, and H. Harada. *Pelotomaculum thermopropionicum* gen. nov., sp. nov., an anaerobic, thermophilic, syntrophic propionate-oxidizing bacterium. *Int J Syst Evol Microbiol*, 52:1729–1735, Sep 2002.

K. Ireton, N. Gunther, and A. Grossman. spo0J is required for normal chromosome segregation as well as the initiation of sporulation in *Bacillus subtilis*. *J Bacteriol*, 176(17):5320–5329, Sep 1994.

N. Ivanova, S. Gronow, A. Lapidus, A. Copeland, T. Glavina Del Rio, M. Nolan, S. Lucas, F. Chen, H. Tice, J. Cheng, E. Saunders, D. Bruce, L. Goodwin, T. Brettin, J. Detter, C. Han, S. Pitluck, N. Mikhailova, A. Pati, K. Mavrommatis, A. Chen, K. Palaniappan, M. Land, L. Hauser, Y. Chang, C. Jeffries, P. Chain, C. Rohde, M. Göker, J. Bristow, J. Eisen, V. Markowitz, P. Hugenholtz, N. Kyrpides, and H. Klenk. Complete genome sequence of *Leptotrichia buccalis* type strain (C-1013-b). *Stand Genomic Sci*, 1:126–132, Sep 2009.

S. Jagadeesan, P. Mann, C. Schink, and P. Higgs. A novel "four-component" two-component signal transduction mechanism regulates developmental progression in *Myxococcus xanthus*. *J Biol Chem*, 284:21435–21445, Aug 2009.

T. Jahns. Unusually stable NAD-specific glutamate dehydrogenase from the alkaliphile *Amphibacillus xylanus*. *Antonie Van Leeuwenhoek*, 70:89–95, Jul 1996.

J. Jalava and E. Eerola. Phylogenetic analysis of *Fusobacterium alocis* and *Fusobacterium sulci* based on 16S rRNA gene sequences: proposal of *Filifactor alocis* (Cato, Moore and Moore) comb. nov. and *Eubacterium sulci* (Cato, Moore and Moore) comb. nov. *Int J Syst Bacteriol*, 49 Pt 4:1375–1379, Oct 1999.

M. Jiang, Y. Tzeng, V. Feher, M. Perego, and J. Hoch. Alanine mutants of the Spo0F response regulator modifying specificity for sensor kinases in sporulation initiation. *Mol Microbiol*, 33: 389–395, Jul 1999.

M. Jiang, W. Shao, M. Perego, and J. Hoch. Multiple histidine kinases regulate entry into stationary phase and sporulation in *Bacillus subtilis*. *Mol Microbiol*, 38:535–542, Nov 2000.

G. Jobb, A. von Haeseler, and K. Strimmer. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol*, 4:18, Jun 28 2004.

P. Junier, M. Frutschi, N. Wigginton, E. Schofield, J. Bargar, and R. Bernier-Latmani. Metal reduction by spores of *Desulfotomaculum reducens*. *Environ Microbiol*, 11:3007–3017, Dec 2009.

V. Juturu and J. Wu. Production of high concentration of l-lactic acid from oil palm empty fruit bunch by thermophilic *Bacillus coagulans* JI12. *Biotechnol Appl Biochem*, Apr 2017.

A. Kaczmarczyk, R. Hochstrasser, J. A. Vorholt, and A. Francez-Charlot. Complex two-component signaling regulates the general stress response in Alphaproteobacteria. *Proc Natl Acad Sci U S A*, 111(48):E5196–204, Dec 2 2014.

C. Kaneuchi, T. Miyazato, T. Shinjo, and T. Mitsuoka. Taxonomic study of helically coiled, spore-forming anaerobes isolated from the intestines of humans and other animals: *Clostridium cocleatum* sp. nov. and *Clostridium spiroforme* sp. nov. *Int J Syst Evol Microbiol*, 29(1):1–12, 1979.

K. Katoh, K. Kuma, H. Toh, and T. Miyata. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, 33(2):511–8, Jan 20 2005.

L. Kimble-Long and M. Madigan. Molecular evidence that the capacity for endosporulation is universal among phototrophic heliobacteria. *FEMS Microbiol Lett*, 199:191–195, May 2001.

K. Kitahara and C. L. Lai. On the spore formation of *Sporolactobacillus inulinus*. *J Gen Appl Microbiol*, 197-203, 13 1967.

H. Klenk, A. Lapidus, O. Chertkov, A. Copeland, T. Del Rio, M. Nolan, S. Lucas, F. Chen, H. Tice, J. Cheng, C. Han, D. Bruce, L. Goodwin, S. Pitluck, A. Pati, N. Ivanova, K. Mavromatis, C. Daum, A. Chen, K. Palaniappan, Y. Chang, M. Land, L. Hauser, C. Jeffries, J. Detter, M. Rohde, B. Abt, R. Pukall, M. Göker, J. Bristow, V. Markowitz, P. Hugenholtz, and J. Eisen. Complete genome sequence of the thermophilic, hydrogen-oxidizing *Bacillus tusciae* type strain (T2) and reclassification in the new genus, *Kyrpidia* gen. nov. as *Kyrpidia tusciae* comb. nov. and emendation of the family Alicyclobacillaceae da Costa and Rainey, 2010. *Stand Genomic Sci*, 5:121–134, Oct 2011.

W. Kloos, D. Ballard, C. George, J. Webster, R. Hubner, W. Ludwig, K. Schleifer, F. Fiedler, and K. Schubert. Delimiting the genus *Staphylococcus* through description of *Macrococcus caseolyticus* gen. nov., comb. nov. and *Macrococcus equipercicus* sp. nov., and *Macrococcus bovicus* sp. nov. and *Macrococcus carouselicus* sp. nov. *Int J Syst Bacteriol*, 48 Pt 3:859–877, Jul 1998.

G. Kozianowski, F. Canganella, F. Rainey, H. Hippe, and G. Antranikian. Purification and characterization of thermostable pectate-lyases from a newly isolated thermophilic bacterium, *Thermoanaerobacter italicus* sp. nov. *Extremophiles*, 1:171–182, Nov 1997.

S. Krishnamurthi, T. Chakrabarti, and E. Stackebrandt. Re-examination of the taxonomic position of *Bacillus silvestris* Rheims et al. 1999 and proposal to transfer it to *Solibacillus* gen. nov. as *Solibacillus silvestris* comb. nov. *Int J Syst Evol Microbiol*, 59:1054–1058, May 2009.

S. Kuehne, J. Heap, C. Cooksley, S. Cartman, and N. Minton. ClosTron-mediated engineering of *Clostridium*. *Methods Mol Biol*, 765:389–407, 2011.

T. Kunisawa. Evolutionary relationships of completely sequenced Clostridia species and close relatives. *Int J Syst Evol Microbiol*, 65:4276–4283, Nov 2015.

C. Lai, B. Males, P. Dougherty, P. Berthold, and M. Listgarten. *Centipeda periodontii* gen. nov., sp. nov. from human periodontal lesions. *Int J Syst Evol Microbiol*, 33(3):628–635, Jul 1983.

J. Larkin and J. Stokes. Taxonomy of psychrophilic strains of *Bacillus*. *J Bacteriol*, 94:889–895, Oct 1967.

M. Laub and M. Goulian. Specificity in two-component signal transduction pathways. *Annu Rev Genet*, 41:121–145, 2007a.

M. Laub, E. Biondi, and J. Skerker. Phosphotransfer profiling: systematic mapping of two-component signal transduction pathways and phosphorelays. *Methods Enzymol*, 423:531–548, 2007.

M. T. Laub and M. Goulian. Specificity in two-component signal transduction pathways. *Annu Rev Genet*, 121-45, 41 2007b.

J. P. Lavigne, N. Bouziges, A. Sotto, J. L. Leroux, and S. Michaux-Charachon. Spondylodiscitis due to *Clostridium ramosum* infection in an immunocompetent elderly patient. *J Clin Microbiol*, 41(5):2223–6, May 2003.

P. Lawson, D. Citron, K. Tyrrell, and S. Finegold. Reclassification of *Clostridium difficile* as *Clostridioides difficile* (Hall and O'Toole 1935) Prévot 1938. *Anaerobe*, 40:95–99, Aug 2016.

J. LeDeaux, N. Yu, and A. Grossman. Different roles for KinA, KinB, and KinC in the initiation of sporulation in *Bacillus subtilis*. *J Bacteriol*, 177:861–863, Feb 1995.

Y. Lee, M. Jain, C. Lee, S. Lowe, and J. Zeikus. Taxonomic distinction of saccharolytic thermophilic anaerobes: description of *Thermoanaerobacterium xyanolyticum* gen. nov., sp. nov., and *Thermoanaerobacterium saccharolyticum* gen. nov., sp. nov.; reclassification of *Thermoanaerobium brockii*, *Clostridium thermosulfurogenes*, and *Clostridium thermohydrosulfuricum* e100-69 as *Thermoanaerobacter brockii* comb. nov., *Thermoanaerobacterium thermosulfurigenes* comb. nov., respectively; and transfer of *Clostridium thermohydrosulfuricum* 39e. 43(1):41–51, 1993.

R. Lewis, D. Scott, J. Brannigan, J. Ladds, M. Cervin, G. Spiegelman, J. Hoggett, I. Barák, and A. Wilkinson. Dimer formation and transcription activation in the sporulation response regulator Spo0A. *J Mol Biol*, 316:235–245, Feb 2002.

R. J. Lewis, J. A. Brannigan, K. Muchova, I. Barak, and A. J. Wilkinson. Phosphorylated aspartate in the structure of a response regulator protein. *J Mol Biol*, 294(1):9–15, Nov 19 1999.

R. J. Lewis, S. Krzywda, J. A. Brannigan, J. P. Turkenburg, K. Muchova, E. J. Dodson, I. Barak, and A. J. Wilkinson. The trans-activation domain of the sporulation response regulator Spo0A revealed by x-ray crystallography. *Mol Microbiol*, 38(2):198–212, Oct 2000.

M. Lindström and H. Korkeala. Laboratory diagnostics of botulism. *Clin Microbiol Rev*, 19: 298–314, Apr 2006.

C. Liu, S. Finegold, Y. Song, and P. Lawson. Reclassification of *Clostridium coccoides*, *Ruminococcus hansenii*, *Ruminococcus hydrogenotrophicus*, *Ruminococcus luti*, *Ruminococcus productus* and *Ruminococcus schinkii* as *Blautia coccoides* gen. nov., comb. nov., *Blautia hansenii* comb. nov., *Blautia hydrogenotrophica* comb. nov., *Blautia luti* comb. nov., *Blautia producta* comb. nov., *Blautia schinkii* comb. nov. and description of *Blautia wexlerae* sp. nov., isolated from human faeces. *Int J Syst Evol Microbiol*, 58:1896–1902, Aug 2008.

M. Long and K. Thornton. Gene duplication and evolution. *Science*, 293(5535), Aug 2001.

M. Long, E. Betran, K. Thornton, and W. Wang. The origin of new genes: glimpses from the young and old. *Nat Rev Genet*, 4(11):865–75, Nov 2003.

J. Lu, Y. Nogi, and H. Takami. *Oceanobacillus iheyensis* gen. nov., sp. nov., a deep-sea extremely halotolerant and alkaliphilic species isolated from a depth of 1050 m on the Iheya ridge. *FEMS Microbiol Lett*, 205:291–297, Dec 2001.

W. Ludwig, K. H. Schleifer, and W. B. Whitman. *Revised Road Map to the Phylum Firmicutes*, pages 1–13. John Wiley & Sons, Ltd, 2015. ISBN 9781118960608.

J. Maignel-Ludop, M. Huchet, and J. Krupp. Botulinum neurotoxins serotypes A and B induce paralysis of mouse striated and smooth muscles with different potencies. *Pharmacol Res Perspect*, 5:e00289, Feb 2017.

T. Malvar, C. Gawron-Burke, and J. A. Baum. Overexpression of *Bacillus thuringiensis* HknA, a histidine protein kinase homology, bypasses early Spo mutations that result in CryIIIA overproduction. *J Bacteriol*, 176(15):4742–9, Aug 1994.

I. Mandic-Mulec, P. Stefanic, and J. van Elsas. Ecology of Bacillaceae. *Microbiol Spectr*, 3: TBS–0017–2013, Apr 2015.

M. Marounek, K. Fliegrova, and S. Bartos. Metabolism and some characteristics of ruminal strains of *Megasphaera elsdenii*. *Appl Environ Microbiol*, 55:1570–1573, Jun 1989.

A. Mattoo, M. Saif Zaman, G. Dubey, A. Arora, A. Narayan, et al. Spo0B of *Bacillus anthracis* - a protein with pleiotropic functions. *FEBS J*, 275:739–752, Feb 2008.

H. Maughan, C. W. Birky Jr., and W. L. Nicholson. Transcriptome divergence and the loss of plasticity in *Bacillus subtilis* after 6,000 generations of evolution under relaxed selection for sporulation. *J Bacteriol*, 191(1):428–33, Jan 2009.

K. Mavromatis, N. Ivanova, I. Anderson, A. Lykidis, S.D. Hooper, H. Sun, V. Kunin, A. Lapidus, P. Hugenholtz, B. Patel, and N. C. Kyrpides. Genome analysis of the anaerobic thermohalophilic bacterium *Halothermothrix orenii*. *PLoS One*, e4192, 4(1) 2009.

E. B. Mearls and L. R. Lynd. The identification of four histidine kinases that influence sporulation in *Clostridium thermocellum*. *Anaerobe*, 28:109–19, Aug 2014.

N. Mesbah, D. Hedrick, A. Peacock, M. Rohde, and J. Wiegel. *Natranaerobius thermophilus* gen. nov., sp. nov., a halophilic, alkalithermophilic bacterium from soda lakes of the Wadi An Natrun, Egypt, and proposal of Natranaerobiaceae fam. nov. and Natranaerobiales ord. nov. *Int J Syst Evol Microbiol*, 57(11):2507–2512, Nov 2007.

D. Miller, G. Suen, D. Bruce, A. Copeland, J. Cheng, C. Detter, L. Goodwin, C. Han, L. Hauser, M. Land, A. Lapidus, S. Lucas, L. Meincke, S. Pitluck, R. Tapia, H. Teshima, T. Woyke, B. Fox, E. Angert, and C. Currie. Complete genome sequence of the cellulose-degrading bacterium *Cellulosilyticum lentocellum*. *J Bacteriol*, 193:2357–2358, May 2011.

T. Mizuno. His-Asp phosphotransfer signal transduction. *J Biochem*, 123:555–563, Apr 1998.

V. Molle, M. Fujita, S. Jensen, P. Eichenberger, J. González-Pastor, J. Liu, and R. Losick. The Spo0A regulon of *Bacillus subtilis*. *Mol Microbiol*, 50:1683–1701, Dec 2003.

C. Moon, D. Pacheco, W. Kelly, S. Leahy, D. Li, J. Kopecny, and G. Attwood. Reclassification of *Clostridium proteoclasticum* as *Butyrivibrio proteoclasticus* comb. nov., a butyrate-producing ruminal bacterium. *Int J Syst Evol Microbiol*, 58:2041–2045, Sep 2008.

W. Moore, J. Johnson, and L. Holdeman. Emendation of bacteroidaceae and *Butyrivibrio* and descriptions of *Desulfomonas* gen. nov. and ten new species in the genera *Desulfomonas*, *Butyrivibrio*, *Eubacterium*, *Clostridium*, and *Ruminococcus*. *Int. J. Syst. Bacteriol.*, 26:238–252, 1976.

M. Mourez, D. Lacy, K. Cunningham, R. Legmann, B. Sellman, J. Mogridge, and R. Collier. 2001: a year of major advances in anthrax toxin research. *Trends Microbiol*, 10:287–293, Jun 2002.

K. Muchová, R. Lewis, D. Perecko, J. Brannigan, J. Ladds, A. Leech, A. Wilkinson, and I. Barák. Dimer-induced signal propagation in Spo0A. *Mol Microbiol*, 53:829–842, Aug 2004.

D. Murdoch and H. Shah. Reclassification of *Peptostreptococcus magnus* (prevot 1933) holdeman and moore 1972 as *Finegoldia magna* comb. nov. and *Peptostreptococcus micros* (prevot 1933) smith 1957 as *Micromonas micros* comb. nov. *Anaerobe*, 5(5):555–559, Oct 1999.

L. Nakamura, M. Roberts, and F. Cohan. Relationship of *Bacillus subtilis* clades associated with strains 168 and W23: a proposal for *Bacillus subtilis* subsp. *subtilis* subsp. nov. and *Bacillus subtilis* subsp. *spizizenii* subsp. nov. *Int J Syst Bacteriol*, 49 Pt 3:1211–1215, Jul 1999.

J. Narula, A. Kuchina, D. Lee, M. Fujita, G. Süel, and O. Igoshin. Chromosomal arrangement of phosphorelay genes couples sporulation and DNA replication. *Cell*, 162:328–337, Jul 2015.

S. Naser, F. Thompson, B. Hoste, D. Gevers, K. Vandemeulebroecke, I. Cleenwerck, C. Thompson, M. Vancanneyt, and J. Swings. Phylogeny and identification of *Enterococci* by atpA gene sequence analysis. *J Clin Microbiol*, 43:2224–2230, May 2005.

T. Nazina, T. Tourova, A. Poltaraus, E. Novikova, A. Grigoryan, A. Ivanova, A. Lysenko, V. Petrunyaka, G. Osipov, S. Belyaev, and M. Ivanov. Taxonomic study of aerobic thermophilic bacilli: descriptions of *Geobacillus subterraneus* gen. nov., sp. nov. and *Geobacillus uzenensis* sp. nov. from petroleum reservoirs and transfer of *Bacillus stearothermophilus*, *Bacillus thermocatenulatus*, *Bacillus thermoleovorans*, *Bacillus kaustophilus*, *Bacillus thermodenitrificans* to *Geobacillus* as the new combinations *G. stearothermophilus*, *G. th. Int J Syst Evol Microbiol*, 51:433–446, Mar 2001.

P. R. Norris, D. A. Clark, J. P. Owen, and S. Waterhouse. Characteristics of *Sulfobacillus acidophilus* sp. nov. and other moderately thermophilic mineral-sulphide-oxidizing bacteria. *Microbiology*, 142(Pt 4):775–83, Apr 1996.

I. Ntaikou, H. Gavala, M. Kornaros, and G. Lyberatos. Hydrogen production from sugars and sweet sorghum biomass using *Ruminococcus albus*. *Int J Hydrogen Energ*, 33:1153–1163, 2008.

S. O-Thong, P. Prasertsan, D. Karakashev, and I. Angelidaki. Thermophilic fermentative hydrogen production by the newly isolated *Thermoanaerobacterium thermosaccharolyticum* psu-2. *Int J Hydrogen Energy*, 33:1204–1214, 2008.

N. Obana, R. Nakao, K. Nagayama, K. Nakamura, H. Senpuku, and N. Nomura. Immunoactive Clostridial membrane vesicle production is regulated by a sporulation factor. *Infect Immun*, 85, May 2017.

M. Ohno, H. Shiratori, M. J. Park, Y. Saitoh, Y. Kumon, N. Yamashita, A. Hirata, H. Nishida, K. Ueda, and T. Beppu. *Symbiobacterium thermophilum* gen. nov., sp. nov., a symbiotic thermophile that depends on co-culture with a *Bacillus* strain for growth. *Int J Syst Evol Microbiol*, 50(Pt 5):1829–32, Sep 2000.

N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, Rajput B., B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, 44(D1):D733–45, Jan 4 2016.

M. Ontiveros Corpus, L. Hernández Andrade, J. López Mendez, and V. Tenorio Gutierrez. Prevention of Blackleg by an immunogen of *Clostridium chauvoei*. *Ann N Y Acad Sci*, 1149:303–305, Dec 2008.

R. Onyenwoke, V. Kevbrin, A. Lysenko, and J. Wiegel. *Thermoanaerobacter pseudethanolicus* sp. nov., a thermophilic heterotrophic anaerobe from Yellowstone National Park. *Int J Syst Evol Microbiol*, 57:2191–2193, Oct 2007.

A. Oren, P. Gurevich, and Y. Henis. Reduction of nitrosubstituted aromatic compounds by the halophilic anaerobic eubacteria *Haloanaerobium praevalens* and *Sporohalobacter marismortui*. *Appl Environ Microbiol*, 57(11):3367–70, Nov 1991.

A. Palop, I. Alvarez, J. Raso, and S. Condón. Heat resistance of *Alicyclobacillus acidocaldarius* in water, various buffers, and orange juice. *J Food Prot*, 63:1377–1380, Oct 2000.

C. Paredes, K. Alsaker, and E. Papoutsakis. A comparative genomic view of Clostridial sporulation and physiology. *Nat Rev Microbiol*, 3:969–978, Dec 2005.

D. Paredes-Sabja, B. Setlow, P. Setlow, and M. Sarker. Characterization of *Clostridium perfringens* spores that lack spoVA proteins and dipicolinic acid. *J Bacteriol*, 190:4648–4659, Jul 2008.

D. Paredes-Sabja, P. Setlow, and M. Sarker. Germination of spores of Bacillales and Clostridiales species: mechanisms and proteins involved. *Trends Microbiol*, 19:85–94, Feb 2011.

D. Paredes-Sabja, A. Shen, and J. Sorg. *Clostridium difficile* spore biology: sporulation, germination, and spore structural proteins. *Trends Microbiol*, 22:406–416, Jul 2014.

S. Park, S. Park, and S. Choi. Characterization of sporulation histidine kinases of *Paenibacillus polymyxa*. *Res Microbiol*, 163:272–278, May 2012.

C. Parker, R. Russell, J. Njoroge, A. Jimenez, R. Taussig, and V. Sperandio. Genetic and mechanistic analyses of the periplasmic domain of the enterohemorrhagic *Escherichia coli* QseC histidine sensor kinase. *J Bacteriol*, 199, Apr 2017.

A. Parte. LPSN–list of prokaryotic names with standing in nomenclature. *Nucleic Acids Res*, 42: D613–D616, Jan 2014.

G. Patel, A. Khan, B. Agnew, and J. Colvin. Isolation and characterization of an anaerobic, cellulolytic microorganism, *Acetivibrio cellulolticus* gen. nov., sp. nov. *Int J Syst Evol Microbiol*, 30 (1):179–185, Jan 1980.

S. Perchat, A. Talagas, S. Poncet, N. Lazar, I. Li de la Sierra-Gallay, M. Gohar, D. Lereclus, and S. Nessler. How quorum sensing connects sporulation to necrotrophism in *Bacillus thuringiensis*. *PLoS Pathog*, 12:e1005779, Aug 2016.

M. Perego, S. Cole, D. Burbulys, K. Trach, and J. Hoch. Characterization of the gene for a protein kinase which phosphorylates the sporulation-regulatory proteins Spo0A and Spo0F of *Bacillus subtilis*. *J Bacteriol*, 171(11):6187–6196, Nov 1989.

E. Petitdemange, F. Caillet, J. Giallo, and C. Gaudin. *Clostridium cellulolyticum* sp. nov., a cellulolytic, mesophilic species from decayed grass. *Int. J. Syst. Bacteriol.*, 34:155159, 1984.

E. Pikuta, A. Lysenko, N. Chuvilskaya, U. Mendrock, H. Hippe, N. Suzina, D. Nikitin, G. Osipov, and K. Laurinavichius. *Anoxybacillus pushchinensis* gen. nov., sp. nov., a novel anaerobic, alkaliphilic, moderately thermophilic bacterium from manure, and description of *Anoxybacillus flavitherms* comb. nov. *Int J Syst Evol Microbiol*, 50 Pt 6:2109–2117, Nov 2000.

S. Pitluck, M. Yasawong, C. Munk, M. Nolan, A. Lapidus, S. Lucas, T. Glavina Del Rio, H. Tice, J. F. Cheng, D. Bruce, C. Detter, R. Tapia, C. Han, L. Goodwin, K. Liolios, N. Ivanova, K. Mavromatis, N. Mikhailova, A. Pati, A. Chen, K. Palaniappan, M. Land, L. Hauser, Y. J. Chang, C. D. Jeffries, M. Rohde, S. Spring, J. Sikorski, M. Goker, T. Woyke, J. Bristow, J. A. Eisen, V. Markowitz, P. Hugenholtz, N. C. Kyrpides, and Klenk H. P. Complete genome sequence of *Thermosediminibacter oceani* type strain (JW/IW-1228P). *Stand Genomic Sci*, 3(2): 108–16, Sep 28 2010.

A. Podgornaia and M. Laub. Pervasive degeneracy and epistasis in a protein-protein interface. *Science*, 347:673–677, Feb 2015.

M. N. Price, P. S. Dehal, and A. P. Arkin. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, 5(3):e9490, Mar 10 2010.

A. Procaccini, B. Lunt, H. Szurmant, T. Hwa, and M. Weigt. Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: orphans and crosstalks. *PLoS One*, 6 (5):e19729, May 9 2011.

M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn. The Pfam protein families database. *Nucleic Acids Res*, 40(Database issue):D290–301, Jan 2012.

G. Reddy, J. Prakash, M. Vairamani, S. Prabhakar, G. Matsumoto, and S. Shivaji. *Planococcus antarcticus* and *Planococcus psychrophilus* spp. nov. isolated from cyanobacterial mat samples collected from ponds in Antarctica. *Extremophiles*, 6:253–261, Jun 2002.

W. Robertson, P. Franzmann, and B. Mee. Spore-forming, *Desulfosporosinus*-like sulphate-reducing bacteria from a shallow aquifer contaminated with gasoline. *J Appl Microbiol*, 88: 248–259, Feb 2000.

M. Roessler and V. Müller. Chloride, a new environmental signal molecule involved in gene regulation in a moderately halophilic bacterium, *Halobacillus halophilus*. *J Bacteriol*, 184: 6207–6215, Nov 2002.

M. Rogosa. Transfer of *Veillonella* Prévot and *Acidaminococcus* Rogosa from Neisseriaceae to Veillonellaceae fam. nov., and the inclusion of *Megasphaera* Rogosa in Veillonellaceae. *Int J Syst Bacteriol*, 21(3):231–33, July 1 1971.

D. Roush, D. Elias, and M. Mormile. Metabolic capabilities of the members of the order Halanaerobiales and their potential biotechnological applications. *Curr Biotech*, 3(1):3–9, 2014.

S. Rowland, W. Burkholder, K. Cunningham, M. Maciejewski, A. Grossman, and G. King. Structure and mechanism of action of Sda, an inhibitor of the histidine kinases that regulate initiation of sporulation in *Bacillus subtilis*. *Mol Cell*, 13:689–701, Mar 2004.

D. Rudner, J. LeDeaux, K. Ireton, and A. Grossman. The spo0K locus of *Bacillus subtilis* is homologous to the oligopeptide permease locus and is required for sporulation and competence. *J Bacteriol*, 173(4):1388–1398, Feb 1991.

M. Salinas, M. Fardeau, P. Thomas, J. Cayol, B. Patel, and B. Ollivier. *Mahella australiensis* gen. nov., sp. nov., a moderately thermophilic anaerobic bacterium isolated from an Australian oil well. *Int J Syst Evol Microbiol*, 54:2169–2173, Nov 2004.

M. Sánchez-Sutil, F. Marcos-Torres, J. Pérez, M. Ruiz-González, E. García-Bravo, M. Martínez-Cayuela, N. Gómez-Santos, A. Moraleda-Muñoz, and J. Muñoz-Dorado. Dissection of the sensor domain of the copper-responsive histidine kinase CorS from *Myxococcus xanthus*. *Environ Microbiol Rep*, 8:363–370, Jun 2016.

E. Sayers, T. Barrett, D. Benson, S. Bryant, K. Canese, V. Chetvernin, D. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. Lipman, T. Madden, D. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. Pruitt, G. Schuler, E. Sequeira, S. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. Tatusova, L. Wagner, E. Yaschenko, and J. Ye. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 37:D5–15, Jan 2009.

Y. Sekiguchi, Y. Kamagata, K. Nakamura, A. Ohashi, and H. Harada. *Syntrophothermus lipocalidus* gen. nov., sp. nov., a novel thermophilic, syntrophic, fatty-acid-oxidizing anaerobe which utilizes isobutyrate. *Int J Syst Evol Microbiol*, 50 Pt 2:771–779, Mar 2000.

I. Sela, H. Ashkenazy, K. Katoh, and T. Pupko. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res*, 43 (W1):W7–14, Jul 1 2015.

J. Seo, H. Kim, G. Jung, M. Nam, J. Chung, J. Kim, J. Yoo, C. Kim, and O. Kwon. Psychrophilicity of *Bacillus psychrosaccharolyticus*: a proteomic study. *Proteomics*, 4:3654–3659, Nov 2004.

S. Seredick, B. Seredick, D. Baker, and G. Spiegelman. An A257V mutation in the *Bacillus subtilis* response regulator Spo0A prevents regulated expression of promoters with low-consensus binding sites. *J Bacteriol*, 191:5489–5498, Sep 2009.

P. Setlow. Spore resistance properties. *Microbiol Spectr*, 2, Oct 2014.

E. Shelobolina, K. Nevin, J. Blakeney-Hayward, C. Johnsen, T. Plaia, P. Krader, T. Woodard, D. Holmes, C. Vanpraagh, and D. Lovley. *Geobacter pickeringii* sp. nov., *Geobacter argillaceus* sp. nov. and *Pelosinus fermentans* gen. nov., sp. nov., isolated from subsurface kaolin lenses. *Int J Syst Evol Microbiol*, 57:126–135, Jan 2007.

O. Shida, H. Takagi, K. Kadowaki, and K. Komagata. Proposal for two new genera, *Brevibacillus* gen. nov. and *Aneurinibacillus* gen. nov. *Int J Syst Bacteriol*, 46:939–946, Oct 1996.

T. Shimada, H. Takada, K. Yamamoto, and A. Ishihama. Expanded roles of two-component response regulator OmpR in *Escherichia coli*: genomic SELEX search for novel regulation targets. *Genes Cells*, 20:915–931, Nov 2015.

J. Sieber, D. Sims, C. Han, E. Kim, A. Lykidis, A. Lapidus, E. McDonnald, L. Rohlin, D. Culley, R. Gunsalus, and M. McInerney. The genome of *Syntrophomonas wolfei*: new insights into syntrophic metabolism and biohydrogen production. *Environ Microbiol*, 12:2289–2301, Aug 2010.

J. Sikorski, A. Lapidus, O. Chertkov, S. Lucas, A. Copeland, T. Glavina Del Rio, M. Nolan, H. Tice, J. F. Cheng, C. Han, E. Brambilla, S. Pitluck, K. Liolios, N. Ivanova, K. Mavromatis, N. Mikhailova, A. Pati, D. Bruce, C. Detter, R. Tapia, L. Goodwin, A. Chen, K. Palaniappan, M. Land, L. Hauser, Y. J. Chang, C. D. Jeffries, M. Rohde, M. Goker, S. Spring, T. Woyke, J. Bristow, J. A. Eisen, V. Markowitz, P. Hugenholtz, N. C. Kyrpides, and Klenk H. P. Complete genome sequence of *Acetohalobium arabaticum* type strain (z-7288). *Stand Genomic Sci*, 3(1): 57–65, Aug 20 2010.

J. Skerker, B. Perchuk, A. Siryaporn, E. Lubin, O. Ashenberg, M. Goulian, and M. Laub. Rewiring the specificity of two-component signal transduction systems. *Cell*, 133:1043–1054, Jun 2008.

J. M. Skerker, M. S. Prasol, B. S. Perchuk, E. G. Biondi, and M. T. Laub. Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. *PLoS Biol*, 3(10):e334, Oct 2005.

T. Sokolova, J. González, N. Kostrikina, N. Chernyh, T. Slepova, E. Bonch-Osmolovskaya, and F. Robb. *Thermosinus carboxydivorans* gen. nov., sp. nov., a new anaerobic, thermophilic, carbon-monoxide-oxidizing, hydrogenogenic bacterium from a hot pool of yellowstone national park. *Int J Syst Evol Microbiol*, 54:2353–2359, Nov 2004.

F. Soriano, R. Fernández-Roblas, R. Calvo, and G. García-Calvo. *In vitro* susceptibilities of aerobic and facultative non-spore-forming gram-positive bacilli to HMR 3647 (RU 66647) and 14 other antimicrobials. *Antimicrob Agents Chemother*, 42:1028–1033, May 1998.

S. Spring, M. Visser, M. Lu, A. Copeland, A. Lapidus, S. Lucas, J. Cheng, C. Han, R. Tapia, L. Goodwin, S. Pitluck, N. Ivanova, M. Land, L. Hauser, F. Larimer, M. Rohde, M. Göker, J. Detter, N. Kyrpides, T. Woyke, P. Schaap, C. Plugge, G. Muyzer, J. Kuever, I. Pereira, S. Parshina, R. Bernier-Latmani, A. Stams, and H. Klenk. Complete genome sequence of the sulfate-reducing firmicute *Desulfotomaculum ruminis* type strain (dl(t)). *Stand Genomic Sci*, 7: 304–319, Dec 2012.

E. Stackebrandt, H.Pohla, R.Kroppenstedt, and H.Hippe andC. R. Woese. 16S rRNA analysis of *Sporomusa*, *Selenomonas*, and *Megasphaera*: on the phylogenetic origin of gram-positive eubacteria. *Arch Microbiol*, 143:270276, 1985.

T. Stadtman and L. McClung. *Clostridium sticklandii* nov. spec. *J Bacteriol*, 73:218–219, Feb 1957.

A. Stamatakis. Phylogenetic models of rate heterogeneity: a high performance computing perspective. *Proceedings 20th IEEE International Parallel & Distributed Processing Symposium*, page 8, 2006.

A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30:1312–1313, May 2014.

T. Stanton and D. C. Savage. *Roseburia cecicola* gen. nov., a motile obligately anaerobic bacterium from a mouse cecum. *Int J Syst Bacteriol*, 33:618–627, 1983.

E. Steiner, A. Dago, D. Young, J. Heap, N. Minton, J. Hoch, and M. Young. Multiple orphan histidine kinases interact directly with Spo0A to control the initiation of endospore formation in *Clostridium acetobutylicum*. *Mol Microbiol*, 80:641–654, May 2011.

K. Stephenson and J. Hoch. Evolution of signalling in the sporulation phosphorelay. *Mol Microbiol*, 46:297–304, Oct 2002.

K. Stephenson and R. Lewis. Molecular insights into the initiation of sporulation in Gram-positive bacteria: new technologies for an old phenomenon. *FEMS Microbiol Rev*, 29:281–301, Apr 2005.

H. Szurmant and J. Hoch. Interaction fidelity in two-component signaling. *Curr Opin Microbiol*, 13:190–197, Apr 2010.

P. K. Talukdar, V. Olgun-Araneda, Alnoman M., D. Paredes-Sabja, and M. R. Sarker. Updates on the sporulation process in *Clostridium* species. *Res Microbiol*, 166(4):225–35, May 2015.

Y. Tan, M. Ayob, M. Osman, and K. Matthews. Antibacterial activity of different degree of hydrolysis of palm kernel expeller peptides against spore-forming and non-spore-forming bacteria. *Lett Appl Microbiol*, 53:509–517, Nov 2011.

D. Taras, R. Simmering, M. Collins, P. Lawson, and M. Blaut. Reclassification of *Eubacterium formicigenerans* Holdeman and Moore 1974 as *Dorea formicigenerans* gen. nov., comb. nov., and description of *Dorea longicatena* sp. nov., isolated from human faeces. *Int J Syst Evol Microbiol*, 52:423–428, Mar 2002.

D. Tena, N. Martínez, J. Casanova, J. García, E. Román, M. Medina, and J. Sáez-Nieto. Possible *Exiguobacterium sibiricum* skin infection in human. *Emerg Infect Dis*, 20:2178–2179, Dec 2014.

C. Thompson, R. Vier, A. Mikaelyan, T. Wienemann, and A. Brune. '*Candidatus* arthromitus' revised: segmented filamentous bacteria in arthropod guts are members of *Lachnospiraceae*. *Environ Microbiol*, 14:1454–1465, Jun 2012.

E. Tocheva, E. Matson, D. Morris, F. Moussavi, J. Leadbetter, and G. Jensen. Peptidoglycan remodeling and conversion of an inner membrane into an outer membrane during sporulation. *Cell*, 146:799–812, Sep 2011.

E. Tocheva, A. Dekas, S. McGlynn, D. Morris, V. Orphan, and G. Jensen. Polyphosphate storage during sporulation in the gram-negative bacterium *Acetonema longum*. *J Bacteriol*, 195:3940–3946, Sep 2013.

E. Tocheva, D. Ortega, and G. Jensen. Sporulation, bacterial cell envelopes and the origin of life. *Nat Rev Microbiol*, 14:535–542, Aug 2016.

K. Trach and J. Hoch. Multisensory activation of the phosphorelay initiating sporulation in *Bacillus subtilis*: identification and sequence of the protein kinase of the alternate pathway. *Mol Microbiol*, 8:69–79, Apr 1993.

K. Trach, J. Chapman, P. Piggot, and J. Hoch. Deduced product of the stage 0 sporulation gene spo0F shares homology with the Spo0A, OmpR, and SfrA proteins. *Proc Natl Acad Sci U S A*, 82(21):7260–7264, Nov 1985.

A. Tsong, M. Miller, R. Raisner, and A. Johnson. Evolution of a combinatorial transcriptional circuit: a case study in yeasts. *Cell*, 115(4):389–399, Nov 2003.

K. Ueda, A. Yamashita, J. Ishikawa, M. Shimada, T. O. Watsuji, K. Morimura, H. Ikeda, M. Hattori, and T. Beppu. Genome sequence of *Symbiobacterium thermophilum*, an uncultivable bacterium that depends on microbial commensalism. *Nucleic Acids Res*, 32(16):4937–44, Sep 21 2004.

T. Ueki and S. Inouye. Transcriptional activation of a heat-shock gene, lond, of *Myxococcus xanthus* by a two component histidine-aspartate phosphorelay system. *J Biol Chem*, 277(8): 6170–7, Feb 22 2002.

L. E. Ulrich and I. B. Zhulin. The MiST2 database: a comprehensive genomics resource on microbial signal transduction. *Nucleic Acids Res*, 38(Database issue):D401–7, Jan 2010.

S. Underwood, S. Guan, V. Vijayasubhash, S. Baines, L. Graham, R. Lewis, M. Wilcox, and K. Stephenson. Characterization of the sporulation initiation pathway of *Clostridium difficile* and its role in toxin production. *J Bacteriol*, 191:7296–7305, Dec 2009.

V. Vacic, L. M. Iakoucheva, and P. Radivojac. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*, 22(12):1536–7, Jun 15 2006.

I. Vishniakov, S. Levitski, and S. Borkhsenius. Effect of heat shock on cells of phytopathogenic mycoplasma *Acholeplasma laidlawii* pg-8a. *Tsitologiia*, 57:5–13, 2015.

P. Vos, G. Garrity, D. Jones, N.R. Krieg, W. Ludwig, and F.A. Rainey. *Bergey's Manual of Systematic Bacteriology*. Springer-Verlag New York, 2009. ISBN 9780387950419.

T. Warnick, B. Methé, and S. Leschine. *Clostridium phytofermentans* sp. nov., a cellulolytic mesophile from forest soil. *Int J Syst Evol Microbiol*, 52:1155–1160, Jul 2002.

Y. Watanabe, F. Nagai, and M. Morotomi. Characterization of *Phascolarctobacterium succinatutens* sp. nov., an asaccharolytic, succinate-utilizing bacterium isolated from human feces. *Appl Environ Microbiol*, 78:511–518, Jan 2012.

M. S. Wei Wang. Phylogenetic relationships between *Bacillus* species and related genera inferred from 16s rDNA sequences. *Braz J Microbiol*, 2009.

M. Westerholm, S. Roos, and A. Schnürer. *Tepidanaerobacter acetatoxydans* sp. nov., an anaerobic, syntrophic acetate-oxidizing bacterium isolated from two ammonium-enriched mesophilic methanogenic processes. *Syst Appl Microbiol*, 34(4):260–266, Jun 2011.

F. Widdel and N. Pfennig. A new anaerobic, sporing, acetate-oxidizing, sulfate-reducing bacterium, *Desulfotomaculum* (emend.) *acetoxidans*. *Arch Microbiol*, 112:119–122, Feb 1977.

K. Willquist, A. Zeidan, and E. van Niel. Physiological characteristics of the extreme thermophile *Caldicellulosiruptor saccharolyticus*: an efficient hydrogen cell factory. *Microb Cell Fact*, 9:89, Nov 2010.

Y. Wolf and E. Koonin. Genome reduction as the dominant mode of evolution. *Bioessays*, 35: 829–837, Sep 2013.

K. Wörner, H. Szurmant, C. Chiang, and J. Hoch. Phosphorylation and functional analysis of the sporulation initiation factor Spo0A from *Clostridium botulinum*. *Mol Microbiol*, 59:1000–1012, Feb 2006.

L. Wu, P. Lewis, R. Allmansberger, P. Hauser, and J. Errington. A conjugation-like mechanism for prespore chromosome partitioning during sporulation in *Bacillus subtilis*. *Genes Dev*, 9: 1316–1326, Jun 1995.

M. Wu, Q. Ren, A. Durkin, S. Daugherty, L. Brinkac, R. Dodson, R. Madupu, S. Sullivan, J. Kolonay, D. Haft, W. Nelson, L. Tallon, K. Jones, L. Ulrich, J. Gonzalez, I. Zhulin, F. Robb, and J. Eisen. Life in hot carbon monoxide: the complete genome sequence of *Carboxydothermus hydrogenoformans* z-2901. *PLoS Genet*, 1:e65, Nov 2005.

K. Wuichet and I. B. Zhulin. Origins and diversification of a complex signal transduction system in prokaryotes. *Sci Signal*, 3(128):ra50, Jun 29 2010.

D. Xing, N. Ren, Q. Li, M. Lin, A. Wang, and L. Zhao. *Ethanoligenens harbinense* gen. nov., sp. nov., isolated from molasses wastewater. *Int J Syst Evol Microbiol*, 56:755–760, Apr 2006.

S. Yamada, H. Sugimoto, M. Kobayashi, A. Ohno, H. Nakamura, and Y. Shiro. Structure of pas-linked histidine kinase and the response regulator complex. *Structure*, 17(10):1333–1344, Oct 2009.

Q. Ye, Y. Roh, S. Carroll, B. Blair, J. Zhou, C. Zhang, and M. Fields. Alkaline anaerobic respiration: isolation and characterization of a novel alkaliphilic and metal-reducing bacterium. *Appl Environ Microbiol*, 70:5595–5602, Sep 2004.

N. Yutin and M. Y. Galperin. A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia. *Environ Microbiol*, 15:2631–41, Oct 2013.

J. Zapf, U. Sen, Madhusudan, J. A. Hoch, and K. I. Varughese. A transient interaction between two phosphorelay proteins trapped in a crystal lattice reveals the mechanism of molecular recognition and phosphotransfer in signal transduction. *Structure*, 8(8):851–62, Aug 15 2000.

D. G. Zavarzina, T. N. Zhilina, B. B. Kuznetsov, T. V. Kolganova, G. A. Osipov, M. S. Kotelev, and G. A. Zavarzin. *Natranaerobaculum magadiense* gen. nov., sp. nov., an anaerobic, alkalithermophilic bacterium from soda lake sediment. *Int J Syst Evol Microbiol*, 63(Pt 12):4456–61, Dec 2013.

T. N. Zhilina, G. A. Zavarzin, E. N. Detkova, and F. A. Rainey. *Natroniella acetigena* gen. nov., sp. nov., an extremely haloalkaliphilic, homoacetic bacterium: A new member of Haloanaerobiales. *Curr Microbiol*, 32(6):320–6, Jun 1996.

T. N. Zhilina, D. G. Zavarzina, A. N. Panteleeva, G. A. Osipov, N. A. Kostrikina, T. P. Tourova, and G. A. Zavarzin. *Fuchsiella alkaliacetigena* gen. nov., sp. nov., an alkaliphilic, lithoautotrophic homoacetogen from a soda lake. *Int J Syst Evol Microbiol*, 62(Pt 7):1666–73, Jul 2012.

T. N. Zhilina, D. G. Zavarzina, E. N. Detkova, E. O. Patutina, and B. B. Kuznetsov. *Fuchsiella ferrireducens* sp. nov., a novel haloalkaliphilic, lithoautotrophic homoacetogen capable of iron reduction, and emendation of the description of the genus *Fuchsiella*. *Int J Syst Evol Microbiol*, 65(8):2432–40, Aug 2015.

P. Zuber and R. Losick. Role of AbrB in Spo0A- and Spo0B-dependent utilization of a sporulation promoter in *Bacillus subtilis*. *J Bacteriol*, 169:2223–2230, May 1987.

Appendix

Supplementary tables

Table A1: Set of 84 Firmicutes genomes used in the bioinformatic analysis of the Spo0 pathway including species tree construction and survey of Spo0 components as described in Chapter 3.

Species Name	Sporulates	Note
Acetivibrio cellulolyticus CD2	No	Patel et al. [1980]
Acetonema longum DSM 6540	Yes	Tocheva et al. [2011]
Acholeplasma laidlawii PG-8A	No	Vishniakov et al. [2015]
Acidaminococcus fermentans DSM 20731	No	Chang et al. [2010]
Alicyclobacillus acidocaldarius subsp. *acidocaldarius* DSM 446	Yes	Palop et al. [2000]
Alkaliphilus metalliredigens QYMF	Yes	Ye et al. [2004]
Ammonifex degensii KC4	No	Huber et al. [1996]
Anoxybacillus flavithermus WK1	Yes	Pikuta et al. [2000]
Bacillus anthracis str. Ames	Yes	
Bacillus clausii KSM-K16	Yes	
Bacillus halodurans C-125	Yes	
Bacillus megaterium DSM 319	Yes	Eppinger et al. [2011]
Bacillus psychrosaccharolyticus ATCC 23296	Yes	Seo et al. [2004]
Bacillus subtilis subsp. *subtilis* str. 168	Yes	Nakamura et al. [1999]
Blautia hansenii DSM 20583	No	Liu et al. [2008]
Brevibacillus brevis NBRC 100599	Yes	Shida et al. [1996]
Butyrivibrio proteoclasticus B316	No	Moon et al. [2008]

Table A1: (continued)

Species Name	Sporulates	Note
Caldicellulosiruptor saccharolyticus DSM 8903	No	Willquist et al. [2010]
Candidatus Arthromitus sp. *SFB-mouse-Yit*	Yes	Thompson et al. [2012]
Candidatus Desulforudis audaxviator MP104C	Yes	Chivian et al. [2008]
Carboxydothermus hydrogenoformans Z-2901	Yes	Wu et al. [2005]
Centipeda periodontii DSM 2778	No	Lai et al. [1983]
Clostridium acetobutylicum ATCC 824	Yes	Steiner et al. [2011]
Clostridium botulinum A str. Hall	Yes	Lindström and Korkeala [2006]
Clostridium lentocellum DSM 5427	Yes	Miller et al. [2011]
Clostridium perfringens ATCC 13124	Yes	Paredes-Sabja et al. [2008]
Clostridium tetani E88	Yes	Bruggemann et al. [2003]
Desulfosporosinus orientis DSM 765	Yes	Robertson et al. [2000]
Desulfotomaculum acetoxidans DSM 771	Yes	Widdel and Pfennig [1977]
Desulfotomaculum reducens MI-1	Yes	Junier et al. [2009]
Desulfotomaculum ruminis DSM 2154	Yes	Spring et al. [2012]
Dorea formicigenerans ATCC 27755	No	Taras et al. [2002]
Enterococcus faecalis V583	No	Naser et al. [2005]
Erysipelatoclostridium ramosum DSM 1402	Yes	Yutin and Galperin [2013]
Erysipelatoclostridium spiroforme DSM 1552	Yes	Yutin and Galperin [2013]
Erysipelothrix rhusiopathiae str. Fujisawa	No	Soriano et al. [1998]
Ethanoligenens harbinense YUAN-3	No	Xing et al. [2006]
Eubacterium eligens ATCC 27750	No	
Eubacterium rectale ATCC 33656	No	
Exiguobacterium sibiricum 255-15	No	Tena et al. [2014]
Filifactor alocis ATCC 35896	No	Jalava and Eerola [1999]
Finegoldia magna ATCC 29328	No	Murdoch and Shah [1999]
Geobacillus kaustophilus HTA426	Yes	Nazina et al. [2001]
Gottschalkia acidurici 9a	Yes	Hartwich et al. [2012]

Table A1: (continued)

Species Name	Sporulates	Note
Heliobacterium modesticaldum Ice1	Yes	Kimble-Long and Madigan [2001]
Kyrpidia tusciae DSM 2912	Yes	Klenk et al. [2011]
Lachnoclostridium phytofermentans ISDg	Yes	Warnick et al. [2002]
Lachnoclostridium symbiosum WAL-14163	Yes	Allen et al. [2003]
Lactobacillus casei ATCC 334	No	
Leptotrichia buccalis C-1013-b	No	Ivanova et al. [2009]
Listeria monocytogenes EGD-e	No	Tan et al. [2011]
Lysinibacillus fusiformis ZC1	Yes	Ahmed et al. [2007]
Lysinibacillus sphaericus C3-41	Yes	Ahmed et al. [2007]
Macrococcus caseolyticus JCSC5402	No	Kloos et al. [1998]
Mageeibacillus indolicus UPII9-5	No	Austin et al. [2015]
Mahella australiensis 50-1 BON	Yes	Salinas et al. [2004]
Megasphaera elsdenii DSM 20460	No	Marounek et al. [1989]
Moorella thermoacetica ATCC 39073	Yes	Byrer et al. [2000]
Oceanobacillus iheyensis HTE831	Yes	Lu et al. [2001]
Paenibacillus polymyxa E681	Yes	Park et al. [2012]
Pelosinus fermentans DSM 17108	Yes	Shelobolina et al. [2007]
Pelotomaculum thermopropionicum SI	Yes	Imachi et al. [2002]
Peptoclostridium difficile 630	Yes	Paredes-Sabja et al. [2014]
Peptoclostridium sticklandii DSM 519	No	Stadtman and McClung [1957]
Peptostreptococcus anaerobius VPI 4330	No	Higaki et al. [2000]
Phascolarctobacterium succinatutens YIT 12067	No	Watanabe et al. [2012]
Planococcus antarcticus DSM 14505	No	Reddy et al. [2002]
Roseburia hominis A2-183	No	Stanton and Savage [1983]
Ruminiclostridium cellulolyticum H10	Yes	Petitdemange et al. [1984]
Ruminiclostridium leptum DSM 753	Yes	Moore et al. [1976]
Ruminiclostridium thermocellum ATCC 27405	Yes	Freier et al. [1988]
Ruminococcus albus 7	No	Ntaikou et al. [2008]

Table A1: (continued)

Species Name	Sporulates	Note
Selenomonas ruminantium subsp. *lactilytica* TAM6421	No	Stackebrandt et al. [1985]
Solibacillus silvestris StLB046	Yes	Krishnamurthi et al. [2009]
Staphylococcus aureus subsp. *aureus* NCTC 8325	No	
Streptococcus pneumoniae TIGR4	No	
Syntrophomonas wolfei subsp. *wolfei* str. Goettingen	No	Sieber et al. [2010]
Syntrophothermus lipocalidus DSM 12680	No	Sekiguchi et al. [2000]
Thermincola potens JR	No	Sokolova et al. [2004]
Thermoanaerobacter italicus Ab9	Yes	Kozianowski et al. [1997]
Thermoanaerobacter pseudethanolicus ATCC 33223	Yes	Onyenwoke et al. [2007]
Thermoanaerobacterium thermosaccharolyticum DSM 571	Yes	O-Thong et al. [2008]
Thermoanaerobacterium xylanolyticum LX-11	Yes	Lee et al. [1993]
Thermosinus carboxydivorans Nor1	No	Sokolova et al. [2004]
Veillonella parvula DSM 2008	No	Gronow et al. [2010]

Table A2: Multiple studies have been undertaken to elucidate the sporulation kinases in a total of eight different genomes. Kinases that have been demonstrated by *in vitro* phosphotransfer, have an effect on sporulation when deleted, or can rescue sporulation in a sporulation deficient background are included. All phosphorelay sporulation kinases were identified in Class Bacilli. All direct phosphorylation architecture kinases have been identified in Class Clostridia. The kinases that encode PAS domains are indicated.

Species	Kinase Locus	Spec. Res.	PAS	Citation
Bacillus anthracis	BA_1356	TSGFQL		Brunsing et al. [2005]
Bacillus anthracis	BA_2291	TSGFKL		Brunsing et al. [2005]
Bacillus anthracis	BA_4223	TVGFQL		Brunsing et al. [2005]
Bacillus anthracis	BA_5029	ASGFQL		Brunsing et al. [2005]
Bacillus subtilis	BSU13530	TAGFQL	✓	Jiang et al. [2000]
Bacillus subtilis	BSU13660	TGGFQL		Jiang et al. [2000]
Bacillus subtilis	BSU13990	TAGFQL	✓	Perego et al. [1989]
Bacillus subtilis	BSU14490	TSGFQI	✓	LeDeaux et al. [1995]
Bacillus subtilis	BSU31450	TVGFQL		Trach and Hoch [1993]
Paenibacillus polymyxa	PPE_01377	TAGFQL		Park et al. [2012]
Paenibacillus polymyxa	PPE_01038	QVGFQL	✓	Park et al. [2012]
Geobacillus stearothermophilus	GT94_02755	TAGFQL		Bick et al. [2009]
Desulfotomaculum acetoxidans	Dtox_1918	TSGFQL	✓	This work
Clostridia acetobutylicum	CA_C0323	NVSAQV		Steiner et al. [2011]
Clostridia acetobutylicum	CA_C0903	NISAQL	✓	Steiner et al. [2011]
Clostridia acetobutylicum	CA_C3319	SVGLQL		Steiner et al. [2011]
Clostridioides difficile	CD1579	NLSSQV	✓	Underwood et al. [2009]
Clostridioides difficile	CD2492	NVSSQL	✓	Underwood et al. [2009]
Ruminiclostridium thermocellum	Clo1313_0286	ASGLQL		Mearls and Lynd [2014]
Ruminiclostridium thermocellum	Clo1313_1942	NISTQI	✓	Mearls and Lynd [2014]
Ruminiclostridium thermocellum	Clo1313_2735	SVGAQL		Mearls and Lynd [2014]

Table A3: Experimentally verified Spo0Fs in eight genomes.

Species	Spo0F Locus	Spec. Res.	Citation
Bacillus subtilis	BSU37130	QGILEVD	Burbulys et al. [1991]

Table A4: Experimentally verified Spo0Bs in eight genomes.

Species	Spo0B Locus	Spec. Res.	Citation
Bacillus subtilis	BSU27930	QLGNSL	Burbulys et al. [1991]
Bacillus anthracis	GBAA_4673	QLGNSL	Mattoo et al. [2008]

Table A5: Experimentally verified Spo0As in eight genomes.

Species	Spo0A Locus	Spec. Res.	Citation
Bacillus subtilis	BSU24220	NELLEYD	Burbulys et al. [1991]; Hoch et al. [1985]
Clostridium acetobutylicum	CA_C2071	NEFIDYD	Steiner et al. [2011]
Clostridium botulinum	CBO1872	NEFIDYD	Wörner et al. [2006]
Ruminiclostridium thermocellum	Clo1313_1409	NEFIEYD	Mearls and Lynd [2014]
Paenibacillus polymyxa	PPE_02831	NEFLEYD	Park et al. [2012]
Clostridioides difficile	CD630_12140	NDFVEYD	Underwood et al. [2009]

Table A6: Spo0F genome neighborhood information. For each putative Spo0F, includes the loci of marker genes identified by their characteristic domain content. If a locus is blank, that marker gene was not encoded in the genome neighborhood of the putative Spo0F.

Genome Name	Locus of marker gene:			
	Spo0F	FbaA	Transaldolase	CTP synthase
Acetonema longum DSM 6540	ALO_07188		ALO_07178	ALO_07203
Alicyclobacillus acidocaldarius subsp. *acidocaldarius* DSM 446	Aaci_2784		Aaci_2783	Aaci_2785
Alkaliphilus metalliredigens QYMF	Amet_0316		Amet_0319	
Ammonifex degensii KC4	Adeg_0054	Adeg_0056	Adeg_0057	Adeg_0051
Anoxybacillus flavithermus WK1	Aflv_2730	Aflv_2729	Aflv_2728	Aflv_2732
Bacillus anthracis str. Ames	BA_5581	BA_5580		BA_5583
Bacillus clausii KSM-K16	ABC3884	ABC3883	ABC3882	ABC3886
Bacillus halodurans C-125	BH3787	BH3786	BH3785	
Bacillus megaterium DSM 319	BMD_5162	BMD_5161	BMD_5160	BMD_5164
Bacillus psychrosaccharolyticus ATCC 23296	WYK_RS12080	WYK_RS12075	WYK_RS12060	WYK_RS12090
Bacillus subtilis subsp. *subtilis* str. 168	BSU37130	BSU37120	BSU37110	BSU37150

Table A6: (continued)

Genome Name	Locus of marker gene:			
	Spo0F	FbaA	Transaldolase	CTP synthase
Brevibacillus brevis NBRC 100599	BBR47_54900	BBR47_54890	BBR47_54880	BBR47_54920
Candidatus Desulforudis audaxviator MP104C	Daud_2177	Daud_2175	Daud_2174	Daud_2180
Carboxydothermus hydrogenoformans Z-2901	CHY_0126	CHY_0128	CHY_0129	CHY_0125
Clostridium acidurici 9a	Curi_c01760	Curi_c01780	Curi_c01790	
Desulfosporosinus orientis DSM 765	Desor_5545	Desor_5543	Desor_5542	Desor_5548
Desulfotomaculum acetoxidans DSM 771	Dtox_0055	Dtox_0057	Dtox_0058	Dtox_0054
Desulfotomaculum reducens MI-1	Dred_3178	Dred_3176	Dred_3175	Dred_3181
Desulfotomaculum ruminis DSM 2154	Desru_3715	Desru_3713	Desru_3712	Desru_3718
Exiguobacterium sibiricum 255-15	Exig_2732	Exig_2731	Exig_2730	Exig_2734
Geobacillus kaustophilus HTA426	GK3387	GK3386	GK3385	GK3389
Heliobacterium modesticaldum Ice1	HM1_1074	HM1_1076	HM1_1077	
Kyrpidia tusciae DSM 2912	Btus_3270	Btus_3269	Btus_3268	Btus_3271

Table A6: (continued)

Genome Name	Locus of marker gene:			
	Spo0F	FbaA	Transaldolase	CTP synthase
Lysinibacillus fusiformis ZC1	BFZC1_18170	BFZC1_18175	BFZC1_18180	BFZC1_18160
Lysinibacillus sphaericus C3-41	Bsph_0981	Bsph_0982	Bsph_0983	Bsph_0979
Macrococcus caseolyticus JCSC5402	MCCL_1780	MCCL_1779		MCCL_1782
Moorella thermoacetica ATCC 39073	Moth_2406	Moth_2404	Moth_2403	Moth_2409
Natranaerobius thermophilus JW/NM-WN-LF	Nther_2879	Nther_2878	Nther_2877	
Oceanobacillus iheyensis HTE831	OB3006	OB3005	OB3004	OB3007
Paenibacillus polymyxa E681	PPE_00136	PPE_00137		PPE_00135
Pelosinus fermentans DSM 17108	FR7_2724			FR7_2728
Pelotomaculum thermopropionicum SI	PTH_2840	PTH_2838	PTH_2837	PTH_2843
Planococcus antarcticus DSM 14505	A1A1_06247	A1A1_06242	A1A1_06237	A1A1_06257
Roseburia hominis A2-183	RHOM_02530	RHOM_02510		

Table A6: (continued)

Genome Name	Locus of marker gene:			
	Spo0F	FbaA	Transaldolase	CTP synthase
Solibacillus silvestris StLB046	SSIL_3398	SSIL_3397	SSIL_3396	SSIL_3400
Syntrophomonas wolfei subsp. *wolfei* str. Goettingen	Swol_2412	Swol_2411	Swol_2410	
Syntrophothermus lipocalidus DSM 12680	Slip_2311	Slip_2310	Slip_2309	
Thermincola potens JR	TherJR_2919	TherJR_2917	TherJR_2915	TherJR_2922
Thermoanaerobacter italicus Ab9	Thit_0118	Thit_0120		
Thermoanaerobacter pseudethanolicus ATCC 33223	Teth39_2116	Teth39_2115		
Thermoanaerobacterium thermosaccharolyticum DSM 571	Tthe_2542	Tthe_2541	Tthe_2540	
Thermoanaerobacterium xylanolyticum LX-11	Thexy_2205	Thexy_2204	Thexy_2203	
Thermosinus carboxydivorans Nor1	TcarDRAFT_0981		TcarDRAFT_0979	

Table A7: Spo0B genome neighborhood information. For each putative Spo0B, includes the loci of marker genes identified by their characteristic domain content. If a locus is blank, that marker gene was not encoded in the genome neighborhood of the putative Spo0B.

Genome Name	Locus of marker gene:			
	Spo0B	ObgE (GTP1_OBG)	L27	L21
Acetonema longum DSM 6540	ALO_08033	ALO_08038	ALO_08028	ALO_08018
Alicyclobacillus acidocaldarius subsp. *acidocaldarius* DSM 446	Aaci_1814	Aaci_1813	Aaci_1815	Aaci_1817
Alkaliphilus metalliredigens QYMF	Amet_2306	Amet_2307	Amet_2305	Amet_2303
Ammonifex degensii KC4	Adeg_0184	Adeg_0185		
Anoxybacillus flavithermus WK1	Aflv_0697	Aflv_0698	Aflv_0696	Aflv_0694
Bacillus anthracis str. Ames	BA_4673	BA_4672	BA_4674	BA_4676
Bacillus clausii KSM-K16	ABC1541	ABC1542		
Bacillus halodurans C-125	BH1212	BH1213		
Bacillus megaterium DSM 319	BMD_4644	BMD_4643	BMD_4645	BMD_4647
Bacillus psychrosaccharolyticus ATCC 23296	WYK_RS19930	WYK_RS19925	WYK_RS19935	WYK_RS19950
Bacillus subtilis subsp. *subtilis* str. 168	BSU27930	BSU27920	BSU27940	BSU27960

Table A7: (continued)

Genome Name	Locus of marker gene:			
	Spo0B	ObgE-like	L27	L21
Brevibacillus brevis NBRC 100599	BBR47_18480	BBR47_18490	BBR47_18470	BBR47_18450
Candidatus Desulforudis audaxviator MP104C	Daud_1874	Daud_1873		
Carboxydothermus hydrogenoformans Z-2901	CHY_0369	CHY_0370	CHY_0368	CHY_0367
Centipeda periodontii DSM 2778	HMPREF9081_1411	HMPREF9081_1415	HMPREF9081_1416	HMPREF9081_1418
Clostridium acidurici 9a	Curi_c19820	Curi_c19810	Curi_c19830	Curi_c19850
Clostridium tetani E88	CTC02057	CTC02056	CTC02058	CTC02060
Desulfosporosinus orientis DSM 765	Desor_5065	Desor_5064	Desor_5066	Desor_5068
Desulfotomaculum acetoxidans DSM 771	Dtox_3313	Dtox_3312	Dtox_3314	Dtox_3315
Desulfotomaculum reducens MI-1	Dred_2533	Dred_2532	Dred_2534	Dred_2535
Desulfotomaculum ruminis DSM 2154	Desru_1034	Desru_1035	Desru_1033	Desru_1032
Exiguobacterium sibiricum 255-15	Exig_2106	Exig_2105	Exig_2107	Exig_2109
Geobacillus kaustophilus HTA426	GK2607	GK2606	GK2608	GK2610

Table A7: (continued)

Genome Name	Locus of marker gene:			
	Spo0B	ObgE-like	L27	L21
Heliobacterium modesticaldum Ice1	HM1_2706	HM1_2705	HM1_2709	HM1_2711
Kyrpidia tusciae DSM 2912	Btus_0684	Btus_0685		
Lysinibacillus fusiformis ZC1	BFZC1_06453	BFZC1_06458	BFZC1_06448	BFZC1_06438
Lysinibacillus sphaericus C3-41	Bsph_3947	Bsph_3946	Bsph_3948	Bsph_3950
Macrococcus caseolyticus JCSC5402	MCCL_1283	MCCL_1282		MCCL_1285
Moorella thermoacetica ATCC 39073	Moth_0560	Moth_0561	Moth_0559	Moth_0557
Oceanobacillus iheyensis HTE831	OB2043	OB2042	OB2044	OB2046
Paenibacillus polymyxa E681	PPE_03655	PPE_03654	PPE_03656	PPE_03658
Pelosinus fermentans DSM 17108	FR7_4318	FR7_4319	FR7_4317	FR7_4315
Pelotomaculum thermopropionicum SI	PTH_0832	PTH_0833	PTH_0831	PTH_0830
Planococcus antarcticus DSM 14505	A1A1_03757	A1A1_03752	A1A1_03762	A1A1_03772

Table A7: (continued)

Genome Name	Locus of marker gene:			
	Spo0B	ObgE-like	L27	L21
Solibacillus silvestris StLB046	SSIL_1468	SSIL_1469	SSIL_1467	SSIL_1465
Syntrophomonas wolfei subsp. *wolfei* str. Goettingen	Swol_1610	Swol_1609	Swol_1611	Swol_1613
Syntrophothermus lipocalidus DSM 12680	Slip_0708	Slip_0709	Slip_0707	Slip_0705
Thermoanaerobacter italicus Ab9	Thit_0854	Thit_0855	Thit_0853	Thit_0851
Thermoanaerobacter pseudethanolicus ATCC 33223	Teth39_1432	Teth39_1431	Teth39_1433	Teth39_1435
Thermoanaerobacterium thermosaccharolyticum DSM 571	Tthe_1141	Tthe_1142	Tthe_1140	Tthe_1138
Thermoanaerobacterium xylanolyticum LX-11	Thexy_1559	Thexy_1558	Thexy_1560	Thexy_1562
Thermosinus carboxydivorans Nor1	TcarDRAFT_1643	TcarDRAFT_1642	TcarDRAFT_1644	TcarDRAFT_1646

Table A8: Specificity residues for putative Spo0F, Spo0B, and Spo0A proteins identified in the 84 genomes of the representative set (see Table S1). If specificity residues are not reported, that component was not identified in that genome.

Species	Spore-former?	Predicted Architecture	Spo0 Component Specificity Residues			
			Spo0F	Spo0B	Spo0A	Spo0A
Acetivibrio cellulolyticus CD2	No				NEFLEYD	NEFIEYD
Acholeplasma laidlawii PG-8A	No					
Blautia hansenii DSM 20583	No				NKMLNLD	
Dorea formicigenerans ATCC 27755	No				NRILEMS	
Erysipelothrix rhusiopathiae str. Fujisawa	No					
Ethanoligenens harbinense YUAN-3	No				SEMVATD	
Eubacterium eligens ATCC 27750	No				NNIANVD	
Eubacterium rectale ATCC 33656	No				DFAENYE	
Filifactor alocis ATCC 35896	No					
Finegoldia magna ATCC 29328	No					
Mageeibacillus indolicus UPII9-5	No					
Peptoclostridium sticklandii DSM 519	No					
Peptostreptococcus anaerobius VPI 4330	No					
Roseburia hominis A2-183	No		YQIEDAD		NRMLDID	
Ruminococcus albus 7	No					
Exiguobacterium sibiricum 255-15	No		EGILQLD	QLSYAM	--MLQHS	
Ammonifex degensii KC4	No		ELVVEAD	QVGFDL	NELAEFS	
Enterococcus faecalis V583	No					
Lactobacillus casei ATCC 334	No					
Listeria monocytogenes EGD-e	No					
Macrococcus caseolyticus JCSC5402	No		ENILEIT	QLTYQF	SELSSHN	

Table A8: (continued)

Species	Spore-former?	Predicted Architecture	Spo0 Component Specificity Residues			
			Spo0F	Spo0B	Spo0A	Spo0A
Planococcus antarcticus DSM 14505	No		QGILEVD	QLLNDL	NELLDYE	
Streptococcus pneumoniae TIGR4	No					
Syntrophomonas wolfei subsp. wolfei str. Goettingen	No		QGVLELD	QVGYEL	NEFIDYN	
Thermincola potens JR	No		QGVLEAD		NEFLEFD	
Thermosinus carboxydivorans Nor1	No		QGILEVD	QVAMQL	NEFIEFN	
Acidaminococcus fermentans DSM 20731	No					
Butyrivibrio proteoclasticus B316	No					
Caldicellulosiruptor saccharolyticus DSM 8903	No				NQFLEVD	
Centipeda periodontii DSM 2778	No			LAIFLI		
Leptotrichia buccalis C-1013-b	No					
Megasphaera elsdenii DSM 20460	No					
Phascolarctobacterium succinatutens YIT 12067	No					
Selenomonas ruminantium subsp. lactilytica TAM6421	No					
Staphylococcus aureus subsp. aureus NCTC 8325	No					
Syntrophothermus lipocalidus DSM 12680	No		QGILELD	QVGLEL	NEFIDFN	
Veillonella parvula DSM 2008	No					
Candidatus Arthromitus sp. SFB-mouse-Yit	Yes	Direct			NEFIDYD	

Table A8: (continued)

Species	Spore-former?	Predicted Architecture	Spo0 Component Specificity Residues			
			Spo0F	Spo0B	Spo0A	Spo0A
Clostridium acetobutylicum ATCC 824	Yes	Direct			NEFIDYD	
Clostridium botulinum A str. Hall	Yes	Direct			NEFIDYD	
Clostridium perfringens ATCC 13124	Yes	Direct			NEFIDYD	
Clostridium tetani E88	Yes	Direct		QVGYQI	NEFIDYD	
Peptoclostridium difficile 630	Yes	Direct			NDFVEYD	
Erysipelatoclostridium ramosum DSM 1402	Yes	Direct			NDLNDSD	NLLENKD
Erysipelatoclostridium spiroforme DSM 1552	Yes	Direct			NDLNESD	DLLSNKN
Mahella australiensis 50-1 BON	Yes	Direct			NDFIEYD	
Clostridium lentocellum DSM 5427	Yes	Direct			NDFIEFD	
Lachnoclostridium phytofermentans ISDg	Yes	Direct			NRMLEVD	
Lachnoclostridium symbiosum WAL-14163	Yes	Direct			NQTMEIT	
Ruminiclostridium cellulolyticum H10	Yes	Direct			NEFIEYD	NLATDND
Ruminiclostridium leptum DSM 753	Yes	Direct			AEFEKIN	
Ruminiclostridium thermocellum ATCC 27405	Yes	Direct			NEFIEYD	NEFLEYD
Thermoanaerobacter italicus Ab9	Yes	Phosphorelay	NGILEID	QVGYQL	NEFIEYD	
Thermoanaerobacter pseudethanolicus ATCC 33223	Yes	Phosphorelay	NGILEID	QLGYQL	NEFIEYD	
Thermoanaerobacterium thermosaccharolyticum DSM 571	Yes	Phosphorelay	NGVLEID	QVGYQL	NEFIDYD	

Table A8: (continued)

Species	Spore-former?	Predicted Architecture	Spo0 Component Specificity Residues			
			Spo0F	Spo0B	Spo0A	Spo0A
Thermoanaerobacterium xylanolyticum LX-11	Yes	Phosphorelay	NGVLELD	QVGYQL	NEFIDYD	
Alkaliphilus metalliredigens QYMF	Yes	Phosphorelay	QRILEVD	QTGYQL	NDFIEYD	
Gottschalkia acidurici 9a	Yes	Phosphorelay	QGILEID	QIGYQL	NDFIEYD	
Acetonema longum DSM 6540	Yes	Phosphorelay	QGILEVD	QVALQL	NEFIEFN	
Candidatus Desulforudis audaxviator MP104C	Yes	Phosphorelay	QAILEAD	QVGLQL	NEFTDFD	
Carboxydothermus hydrogenoformans Z-2901	Yes	Phosphorelay	QGILDAS	QIGFQL	NEFLEFD	
Desulfosporosinus orientis DSM 765	Yes	Phosphorelay	QGVLETD	QVGNQL	NEFVEYD	NNLMNYD
Desulfotomaculum acetoxidans DSM 771	Yes	Phosphorelay	QGILEVD	QVGLQL	NEFLDFD	
Desulfotomaculum reducens MI-1	Yes	Phosphorelay	QGVLEAD	QVGLQL	NDFLEFD	
Desulfotomaculum ruminis DSM 2154	Yes	Phosphorelay	QGVLETD	QVGLQL	NEFLEFD	
Heliobacterium modesticaldum Ice1	Yes	Phosphorelay	QGVLEAD	QTGYQV	NEFIDYD	
Moorella thermoacetica ATCC 39073	Yes	Phosphorelay	QGVLEAD	QVGYQL	NEFVEYD	
Pelosinus fermentans DSM 17108	Yes	Phosphorelay	QGILEVD	QVAMQM	NEFIEYN	
Pelotomaculum thermopropionicum SI	Yes	Phosphorelay	QGVLEAD	QVGLQL	NEFLEFD	
Alicyclobacillus acidocaldarius subsp. acidocaldarius DSM 446	Yes	Phosphorelay	QGILEVD	QIALQM	HEFLEFD	
Anoxybacillus flavithermus WK1	Yes	Phosphorelay	QGILEVD	QLGNSL	NELLEYD	
Bacillus anthracis str. Ames	Yes	Phosphorelay	QGILEVD	QMGNSL	NELMSYD	
Bacillus clausii KSM-K16	Yes	Phosphorelay	QGILEVD	QLGNAI	NELLEYD	

Table A8: (continued)

Species	Spore-former?	Predicted Architecture	Spo0 Component Specificity Residues			
			Spo0F	Spo0B	Spo0A	Spo0A
Bacillus halodurans C-125	Yes	Phosphorelay	QGILEID	QLGYAL	NELLDYD	
Bacillus megaterium DSM 319	Yes	Phosphorelay	QGILEVD	QLGNAL	NELLDYD	
Bacillus psychrosaccharolyticus ATCC 23296	Yes	Phosphorelay	QGILEVD	QLGNEL	NELLEYD	
Bacillus subtilis subsp. subtilis str. 168	Yes	Phosphorelay	QGILEVD	QLGNSL	NELLEYD	
Brevibacillus brevis NBRC 100599	Yes	Phosphorelay	QGILEVD	QVGYHL	NEFLEYD	
Geobacillus kaustophilus HTA426	Yes	Phosphorelay	QGILEVD	QLGHAL	NELLEYD	
Kyrpidia tusciae DSM 2912	Yes	Phosphorelay	QGILELD	QLGYQL	NEFLDYD	
Lysinibacillus fusiformis ZC1	Yes	Phosphorelay	QGILEVD	QLMNDL	NELTQYE	
Lysinibacillus sphaericus C3-41	Yes	Phosphorelay	QGILEVD	QLMNDL	NELTQYE	
Oceanobacillus iheyensis HTE831	Yes	Phosphorelay	QGILEVN	QLGYSM	NELMDYD	
Paenibacillus polymyxa E681	Yes	Phosphorelay	QGILEVD	QVGYRM	NEFLEYD	
Solibacillus silvestris StLB046	Yes	Phosphorelay	---LEVN	HLMNDL	NELMVYE	

Table A9: Supporting information on potential sporulation kinases. All experimentally verified sporulation kinases have been orphans and commonly encode PAS domains, suggesting that unknown sporulation kinases may be identified using these as criteria. Thus, here I tabulate all orphan kinases in spore-formers including specificity residues, PAS domain encoding, and catlytic domain subtype. Catalytic domain subtypes were annotated by Agfam v1 [Alexander and Zhulin, 2007].

Specificity residues are more similar within architecture than between architectures. Further, orphan kinases in genomes where a phosphorelay has been predicted are more likely to encode the HK_CA:3 subtype catalytic domain, while orphan kinases in direct phosphorylation architectures are more likely to encode HK_CA:2 subtype catalytic domains.

Species	Predicted Architecture	Spec. Res.	PAS	HK_CA	Spec. Res.	PAS	HK_CA
Pelosinus fermentans DSM 17108	Phosphorelay	TTGYQM	✓	3	TTGYQV	✓	3
		TTGYQV	✓	3	TTGYQV	✓	3
		TTGYQL		3	TTGFQM		3
		TTGFQM	✓	3	TSGFQF	✓	3
		TTGYQY	✓	3	TTGYRY		3
		AAGYQF	✓	3	TVGYQF	✓	3
		TTGFET	✓	2	TTGFQS		3
		TTGLQF					
Acetonema longum DSM 6540	Phosphorelay	TTGYQV	✓	3	TTGYQL	✓	3
		TTGFQI	✓	3	TVGYQF		3
		TAGFET		2	TVGYQL	✓	3
		AYNNEG		3	AVANDI		2
		AITNEV		2	TVGHEG		2
Carboxydothermus hydrogenoformans Z-2901	Phosphorelay	TTGFQL	✓	3	FTSYEL		3

Table A9: (continued)

Species	Predicted Architecture	Spec. Res.	PAS	HK_CA	Spec. Res.	PAS	HK_CA
Candidatus Desulforudis audaxviator MP104C	Phosphorelay	TAGFQM	✓	3	TVGYQL		3
		VSTFEL	✓	3			
Desulfotomaculum ruminis DSM 2154	Phosphorelay	TTGFQM	✓	3	TTGFQM	✓	3
		TTGFQM	✓	3	TSGLQI		3
		TTGLQL		3	TSGFQL	✓	3
Desulfotomaculum reducens MI-1	Phosphorelay	TTGFQM	✓	3	TTGLQL		3
		TTGLQI		3	TSGFQL	✓	3
		TSGFQL	✓	3	TSGYKA		2
Pelotomaculum thermopropionicum SI	Phosphorelay	TTGFQL	✓	3	TAGFQL	✓	3
		TAGFQI	✓	3	TVGFQL		3
Desulfotomaculum acetoxidans DSM 771	Phosphorelay	TTGFQL	✓	3	TTGFQL	✓	3
		TTGFQM	✓	3	TTGFQM	✓	3
		TTGFQM	✓	3	TTGFQF	✓	3
		TAAFEL					
Desulfosporosinus orientis DSM 765	Phosphorelay	TTGYQL	✓	3	TTGYQL	✓	3
		TTGYQL	✓	3	TTGYQL	✓	3
		TTGYQL		3	TTGFQL	✓	3
		TTGFQL		3	TTGFQL	✓	3
		TTGFQL	✓	3	TTGFQL	✓	3
		TTGFQL	✓	3	TTGFQL	✓	3
		TTGFQM		3	TTGFQF	✓	3
		GIATQV					

Table A9: (continued)

Species	Predicted Architecture	Spec. Res.	PAS	HK_CA	Spec. Res.	PAS	HK_CA
Moorella thermoacetica ATCC 39073	Phosphorelay	TSGFET	✓	2	TSGFQL		3
Heliobacterium modesticaldum Ice1	Phosphorelay	TTGFQI NAGIDL	✓	3	TTGFQL ASGYDM	✓	3
Thermoanaerobacter italicus Ab9	Phosphorelay	GFSNKN	✓	3	TSGFQL	✓	3
Thermoanaerobacter pseudethanolicus ATCC 33223	Phosphorelay	TSGFQL	✓	3			
Thermoanaerobacterium xylanolyticum LX-11	Phosphorelay	TAGFQI		2	TSGFQL	✓	3
Thermoanaerobacterium thermosaccharolyticum DSM 571	Phosphorelay	TSGFQL	✓	3			
Clostridium acidurici 9a	Phosphorelay	TSGYQL ASLFTI NISTQL NVSAQL	 ✓ 	2 2 2 2	TTGFQI NVSIQL NMSTQV NVSAQL	 	3 2 2 2
Alkaliphilus metalliredigens QYMF	Phosphorelay	TVGYQL NVGTQL NVGSQL	 ✓ ✓	3 2 2	NVGTQL NIGSQL NVGAQL	 ✓	2 2 2
Bacillus subtilis subsp. *subtilis* str. 168	Phosphorelay	TAGFQL TSGFQI TGGFQL	✓ ✓ 	3 3 3	TAGFQL TVGFQL	✓ 	3 3
Solibacillus silvestris StLB046	Phosphorelay	TSGFQM TSGLEV	✓ ✓	3 3	TVGYQM	✓	3

Table A9: (continued)

Species	Predicted Architecture	Spec. Res.	PAS	HK_CA	Spec. Res.	PAS	HK_CA
Lysinibacillus fusiformis ZC1	Phosphorelay	TSTYDL		2	TSGFEL		3
Lysinibacillus sphaericus C3-41	Phosphorelay	TSTYDL		2	TSGFDL		3
		TAAYES		2			
Bacillus megaterium DSM 319	Phosphorelay	TTGFQL	✓	3	TAGFQM		3
		TAGFQL	✓	3	TAGFQL	✓	3
		TAGFQL	✓	3	TAGFQL	✓	3
		TSGFQV	✓	3	TVGFQL		3
		QSGFQF		3	AVGFSL		3
		TSGFHL	✓	3	TSGFQL	✓	3
		TSGFQL	✓	3	TSGFQL	✓	3
Bacillus anthracis str. Ames	Phosphorelay	TVGFQL		3	TVGFQL		3
		SVSFEG		2	ALAYMG		2
		TGGLAL		3	ASGFQL		3
		TSGFKL		3	TSGFTL	✓	3
		TSGFQL		3			
Geobacillus kaustophilus HTA426	Phosphorelay	TTGYQL	✓	3	TSGFQM	✓	3
		TAGFQL	✓	3	TAGFQL		3
		TAGFQL		3	TVGFQL		3
Anoxybacillus flavithermus WK1	Phosphorelay	TSGFQM	✓	3	TTGFQF		3
		TTGYQF	✓	3	TAGFQL	✓	3
		TVGFQL		3	TVGFQL		3
		AGGFQL		3	AVGFAF		3

Table A9: (continued)

Species	Predicted Architecture	Spec. Res.	PAS	HK_CA	Spec. Res.	PAS	HK_CA
Bacillus halodurans C-125	Phosphorelay	TSGFQY		3	SVGFQL		
		AVGFSL		3	TPGFNL	✓	3
Bacillus clausii KSM-K16	Phosphorelay	TTGFQL		3	TVGFQL		3
		TVSYKL		2	SVGFQL		3
		TSGFQL	✓	3			
Oceanobacillus iheyensis HTE831	Phosphorelay	TSGFQL	✓	3			
Paenibacillus polymyxa E681	Phosphorelay	TTGFQL	✓	3	TSGFEY		2
		TAGFQL	✓	3	TVGFQM		3
		TEGADL			NSNLQI		
		NGGMGL	✓		QVGFQL		3
		AVGFQL	✓	3	TYGYNV		2
Brevibacillus brevis NBRC 100599	Phosphorelay	TSGFQM	✓	3	TAGFQY	✓	3
		TSGFQF	✓	3	TAGFQL	✓	3
		TAGFQL	✓	3	TVGFQL		3
		TVGFQL		3	TAGFKL		3
		CAGFKL		3	TSGFRL		3
		TSGYDL		2	TSGFQL	✓	3
Alicyclobacillus acidocaldarius subsp. *acidocaldarius* DSM 446	Phosphorelay	TTGFQL		3	TTGFQL	✓	3
		TTGFQL	✓	3	AAQMKA		3
		TAGFKL	✓	3	TAGFDL	✓	3
Kyrpidia tusciae DSM 2912	Phosphorelay	TAGFQL	✓	3	TSGFQI		3

Table A9: (continued)

Species	Predicted Architecture	Spec. Res.	PAS	HK_CA	Spec. Res.	PAS	HK_CA
Clostridium spiroforme DSM 1552	Direct	TSTYDL		2	TSAFII		
		TSGYKL		2			
Clostridium ramosum DSM 1402	Direct	No orphan kinases encoded					
Mahella australiensis 50-1 BON	Direct	TSGFET	✓	2			
Clostridium lentocellum DSM 5427	Direct	AIGREL		2	NIATQL	✓	2
		NVMTQL	✓	2	NVSNQL		2
		TAGYQL		2			
Clostridium phytofermentans ISDg	Direct	TSGYQL		2	TSGYEL		2
Clostridium symbiosum WAL-14163	Direct	TTGYSL		2	TSVLDS		2
		TAGYDS		2	TSGYEA		2
		TAGYDI					
Clostridium thermocellum ATCC 27405	Direct	VVSAQA		2	SVGAQL		2
		ASGLQL		3			
Clostridium cellulolyticum H10	Direct	NLSSKL	✓	2	TAGYQL		2
Clostridium leptum DSM 753	Direct	TSGFEL		2			
Candidatus Arthromitus sp. *SFB-mouse-Yit*	Direct	TLSSQA	✓	2			
Clostridium perfringens ATCC 13124	Direct	TSGFET	✓	2	TNSHEA		2
		NVATQL		2	NISTQL	✓	2
		NVSSQV		2	NVSAQL		2
		NLSAQL	✓	2	NISCQL	✓	2
		TSGYYL		2			

Table A9: (continued)

Species	Predicted Architecture	Spec. Res.	PAS	HK_CA	Spec. Res.	PAS	HK_CA
Clostridium acetobutylicum ATCC 824	Direct	TTGYQM		2	TAGYKF		2
		NISAQV	✓	2	NVSAQV		2
		NISAQL	✓	2	SVGLQL		3
		TSGYEG		2			
Clostridium tetani E88	Direct	TSGYQL		2	SSIYDI		2
		TVSSYM		2	NISSQL		2
		TSGWIT		2			
Clostridium botulinum A str. Hall	Direct	TSGYQM		2	TISSQL	✓	2
		NISAQV		2	NVTAQL	✓	2
		NVTAQI	✓	2	TAGYEL		2
Clostridium difficile 630	Direct	TIGNDL		2	NNSSQL	✓	2
		NLSSQV	✓	2	NVSSQL	✓	2
		SPGYEL		2			

Constructs

Expression and protein purification were performed as described in Laub et al. [2007] and Skerker et al. [2008]. The destination vectors used add an N-terminal tagging protein (either Thioredoxin or Maltose Binding Protein) and a six histidine tag for increased folding fidelity and Ni-NTA column purification, respectively. Another destination vector only adds the six histidine tags. These were generously provided by Michael Laub.

The reference sequences for proteins used in this thesis are listed in table A11. Dtox_1918 was truncated N-terminally truncated (residues 301-535) for solubility. Dtox_2041 (residues 1-129) was truncated in the initial construct as the C-terminal output domain does not effect interaction. Ca_C0903 (residues 244-683) was truncated in accordance with the constructs used by Steiner et al. [2011].

The *D. acetoxidans* and *C. acetobutylicum* proteins were encoded by synthetic DNA sequence with codon usage and GC content optimized for expression in an *E. coli* system. Codon usage recoding was performed using the Graphical Codon Usage Analyzer [Fuhrmann et al., 2004]. GC% was designed in consultation with GeneWiz Inc. Synthetic sequences obtained from GeneWiz Inc. and Thermofisher respectively.

B. subtilis proteins were expressed from plasmids generously provided in expression constructs by Michael Laub.

Table A10: Strains and plasmids used in my experiments for the characterization of the *D. acetoxidans* phosphorelay (Section 3.3.2), interchange of *B. subtilis* and *D. acetoxidans* phosphorelays, and cross-species interaction experiments between *B. subtilis*, *D. acetoxidans*, and *C. acetobutylicum*.

Organism	Strain Name	Description	Source or Reference
E. coli	BL21	Protein expression and purification	
E. coli	TOP10	Plasmid propagation	

Plasmid Category	Plasmid Name	Description	Source or Reference
General purpose vectors	pUC19	High-copy replicon for Genewiz Synthetic Sequences	Genewiz
	pUC19-Dtox_1918	Plasmid containing synthetic *D. acetoxidans* HK Dtox_1918 gene; re-coded for expression in E. coli	
	pUC19-Dtox_0055	Plasmid containing synthetic *D. acetoxidans* spo0F gene; re-coded for expression in E. coli	
	pUC19-Dtox_3313	Plasmid containing synthetic *D. acetoxidans* spo0B gene; re-coded for expression in E. coli	
	pUC19-Dtox_2675	Plasmid containing synthetic *D. acetoxidans* spo0A gene; re-coded for expression in E. coli (synthesized only residues 1-129)	
Entry clones	pENTR-Dtox_1918	*D. acetoxidans* HK Dtox_1918 Entry vector in pENTR/D-TOPO (kanR); Dtox_1918 N-terminally truncated at residue 301 by PCR	This study, Genewiz
	pENTR-Dtox_0055	*D. acetoxidans* Spo0F Entry vector in pENTR/D-TOPO (kanR)	This study, Genewiz

Table A10: (continued)

Plasmid Category	Plasmid Name	Description	Source or Reference
Entry clones	pENTR-Dtox_3313	*D. acetoxidans* Spo0B Entry vector in pENTR/D-TOPO (kanR)	This study, Genewiz
	pENTR-Dtox_2675	*D. acetoxidans* Spo0A Entry vector in pENTR/D-TOPO (kanR)	This study, Genewiz
	pENTR-Ca_C0903	*C. acetobutylicum* HK Ca_C0903 Entry vector in pENTR	This study, Thermofisher
	pENTR-Ca_C3319	*C. acetobutylicum* HK Ca_C3319 Entry vector in pENTR	This study, Thermofisher
	pENTR-Ca_C2071	*C. acetobutylicum* Spo0A Entry vector in pENTR	This study, Thermofisher
Destination vectors	ML310	pET-TRX-His6-TEV (ampR, chlorR); Adds N-terminal 6-Histidine Thioredoxin tag	Skerker et al. [2008]
	ML333	pET-His6-MBP-TEV (ampR, chlorR); Adds N-terminal 6-Histidine and Maltose Binding Protein tag	Skerker et al. [2008]
	ML375	pET-His6-TEV (ampR, chlorR); Adds N-terminal 6-Histidine tag	Skerker et al. [2008]
Expression Vectors	ML333-Dtox_1918	Expression vector for *D. acetoxidans* HK Dtox_1918 with His-MBP tag	This study
	ML310-Dtox_0055	Expression vector for *D. acetoxidans* spo0F with His-TRX tag	This study
	ML333-Dtox_3313	Expression vector for *D. acetoxidans* spo0B with His-MBP tag	This study

Table A10: (continued)

Plasmid Category	Plasmid Name	Description	Source or Reference
Expression Vectors	ML375-Dtox_2675	Expression vector for *D. acetoxidans* spo0Awith His tag	This study
	ML333-Ca_C0903	Expression vector for *C. acetobutylicum* HK Ca_C0903 with His-MBP tag	This study
	ML333-Ca_C3319	Expression vector for *C. acetobutylicum* HK Ca_C3319 with His-MBP tag	This study
	ML310-Ca_C2071	Expression vector for *C. acetobutylicum* Spo0A with His-TRX tag	This study
	ML333-Bs_KinA	Expression vector for *B. subtilis* HK KinA with His-MBP tag	Laub Lab
	ML310-Bs_Spo0F	Expression vector for *B. subtilis* Spo0F with His-TRX tag	Laub Lab
	ML310-Bs_Spo0B	Expression vector for *B. subtilis* Spo0B with His-TRX tag	Laub Lab
	ML310-Bs_Spo0A	Expression vector for *B. subtilis* Spo0A with His-TRX tag	Laub Lab

Table A11: Accession numbers and truncation location for protein sequences used

Species Name	Spo0 Component	Locus Name	Accession Number	Truncation
Desulfotomaculum acetoxidans DSM 771	kinase	Dtox_1918	WP_015757463.1	301-535(end)
	Spo0F	Dtox_0055	WP_012813470.1	
	Spo0B	Dtox_3313	WP_015758734.1	
	Spo0A	Dtox_2041	WP_042317062.1	(start)1-129
Clostridium acetobutylicum ATCC 824	kinase	Ca_C0903	NP_347539.1	244-683(end)
	kinase	Ca_C3319	NP_349911.1	
	Spo0A	Ca_C2071	NP_348690.1	
Bacillus subtilis subsp. *subtilis* str. 168	Bs_KinA	BSU13990	NP_389282.1	
	Spo0F	BSU37130	NP_391594.1	
	Spo0B	BSU27930	NP_390671.1	
	Spo0A	BSU24220	NP_390302.1	

Sequences

D. acetoxidans

Dtox_1918-synth

ATGGAGGAGCTGTACAAGGCCATCACCGAAATCGACATCCTGAGCCCTCCGAAAGACTAT
GTGCGCAATCTGTTACGCGTGATCTGTGACGAGCTGGGTTACAGCTACGGCAGCGTGATC
GATGTGGATAGCCAGGGCAAGGGCTTCATTTTTGCAAGCTACAACTTACCTAACAACTAC
CCGCAGACAATCAATCAGGTGAACGTGCCGGTGCTGAGCAGTCCGAGCGGTGAGGCCATC
AAGAAATGCAAGATCGTTATCGTGCAGAACCCGTTAAGCGAGGAACGTCTGGTGCCGTGG
CATGAGATCATCCGCACCAACAATATCCAGACCATCACCTGGGTTCCGCTGTTAAGCAAG
GGCCGCGCATTAGGCACCTACATCCTGTACGACACCAAGGTGCGTGACATTAGCGAGGAG
AAGCAGCAGCTGCTGAAGCAAATTGGCGTTATGATCAGTATTGCAATTAGCAGTAACCAG
TACCTGGACCAGTTAAACCAGAAGACCGGCGAGCTGATCGAAGAGGTTCACAAACGCAAG
CGTGCCCAGTTAGCACTGAGCGAAAGTGAGCGCTTCAGCAGCGGTGTTTTCGAGAGCATC
CGCGACAACCTGTGCGTTATTGATAACCAGTACAACATCCTGAAAGTGAACCCGGTTCTG
AACCGCCTGTATGATACCAAGCAGCTGATCGGCGAGAAGTGTTACTACGCCTTTCACGAC
CTGGAGCAGATCTGCAAAAACTGTCCGGGCAAACGCGTGCTGGAGACAGGTAAGAGCGCA
AGTGGCATTATCACCCTGAAGAACAGCAAAGGTGAGATCAAGGGCTGGTTCGAGGTTCAT
AGCCACCCGTTCATCGACATGAACACCGGCGAGATCAAAGGTGTGATCAAGTACGGCCGC
GACATTACCGACAAGGTTAAGATGGAACAGGAGATCGCCCGCCTGGATCGCCTGAATCTG
GTGGGCGAGATGGCCGCCAGTATTAGCCACGAGGTTCGCAACCCTATGACCACAGTGCGC
GGCTTCCTGCAGCTGTTACAGAAAAAGGAAGACTGCAAAAAGTACAACGATTACTTCAAC
CTGATGATCGAAGAGCTGGACCGCGCAAACAGCATCATCACCGAGTTCCTGAGCCTGGCC
AAAGACCGCCCGCTGGACCTGATCATGAATAATATCAACAGTGTGGTGGAGATCCTGCAA
CCGCTGATCATCGCCCACGCCGTGCTGTACGATACCACCGTGAAGATTGAGCTGAGCGAG
ATCCCGAATCTGTACCTGGACGAAAAGGAGATCCGCCAGCTGATCCTGAACCTGGTTCGT
AATGGTCTGGAGGCCATGAGCCCTGGTAAAACCCTGACCATCAGCACATACATCGAGAAC
GACACAGTGATCTTAGCCGTTGGCGACCAGGGTGAGGGTATCAAAAGCGAGATCATGGCC
ATGATCGGCACCCCGTTCTTCACAACCAAGGAACAGGGTACAGGCCTGGGCCTGGCCGTT
TGCTACCGTATTGCCGCACGTCACAACGCCCGCATCAACATCAAGACAGACAGCACCGGC
ACAACCTTCGAGGTGCGCTTTAAAGCAAGCAGCCTGGACAACAAGTAA

Dtox_1918 N-terminally Truncated
A32

CGCGACATTACCGACAAGGTTAAGATGGAACAGGAGATCGCCCGCCTGGATCGCCTGAAT
CTGGTGGGCGAGATGGCCGCCAGTATTAGCCACGAGGTTCGCAACCCTATGACCACAGTG
CGCGGCTTCCTGCAGCTGTTACAGAAAAAGGAAGACTGCAAAAAGTACAACGATTACTTC
AACCTGATGATCGAAGAGCTGGACCGCGCAAACAGCATCATCACCGAGTTCCTGAGCCTG
GCCAAAGACCGCCCGCTGGACCTGATCATGAATAATATCAACAGTGTGGTGGAGATCCTG
CAACCGCTGATCATCGCCCACGCCGTGCTGTACGATACCACCGTGAAGATTGAGCTGAGC
GAGATCCCGAATCTGTACCTGGACGAAAAGGAGATCCGCCAGCTGATCCTGAACCTGGTT
CGTAATGGTCTGGAGGCCATGAGCCCTGGTAAAACCCTGACCATCAGCACATACATCGAG
AACGACACAGTGATCTTAGCCGTTGGCGACCAGGGTGAGGGTATCAAAAGCGAGATCATG
GCCATGATCGGCACCCCGTTCTTCACAACCAAGGAACAGGGTACAGGCCTGGGCCTGGCC
GTTTGCTACCGTATTGCCGCACGTCACAACGCCCGCATCAACATCAAGACAGACAGCACC
GGCACAACCTTCGAGGTGCGCTTTAAAGCAAGCAGCCTGGACAACAAGTAA

Dtox_Spo0A-synth

ATGATCATGATGCGCAAGGCCATCAAGCTGCTGATTGCAGACGACAACCGCGAGTTCTGC
GAGCTGATGAAGGACTTCTTCAGTAAGCAGGACGACCTGGAGATCTGCGGCATCGCCTAC
AACGGCCTGGAGGCCATCGACATCATCAAGGAGCACAAGCCTGACATCGCCATCCTGGAC
ATCATCATGCCGCACCTGGACGGCATCGGTGTGCTGGAGAAGCTGGCCAGCGGCATCGTG
AGCAAACGCCCGAAGGTGATCATGCTGACCGCCTTCGGCCAGGAGAGTGTGACACAGCGT
GCCGTGGAACTGGGCGCCGACTACTACGTTCTGAAGCCGTTCGACTTCAGCGTGCTGGCA
ACCCGTGTGCGTCAGCTGAGCGACGGCATCACCGTTAACCAGTAA

Dtox_Spo0B-synth

ATGCAAATCAAGGACCTGCTGAAAGTGATCCAGGCCCAGCGTCACGACTTCCTGAACCAC
CTGCAGGTGATCAGCGGCCTGGTGCAGCTGGGCAAGATCGATCGCGTGAACGAGTACATC
GGCGAGGTGTGCATGAATATGGGCCTGATGAGCCTGGTGAACCGCATCAAAATCCCGGAG
ATCAGCGCCGCCCTGATGATCGGCCACAACCTGGCCTTTGAACAGCAGGTGCAGGTGGAG
TACCGCCTGGAGGACGAGCTGGGTGACACCGTGCTGCCTGCCGAGATGCTGGCCGACATG
CTGATCCATATGCAGAAGATCATCTTCGACCTGGCAGCAGCCGCAGCCGAAGTGAGCACC
GTGACCCTGATCATGGAGCGCCTGAAGAACAACTACATTTGCCGCCTGCAACTGAAGATC
CCGGATGAGATCCCTGCCGCCACCGAGCAGATCTTCGAGGAGCTGCAGCAGATCGCCAAG
GTTAACATGGGTCTGGACAACGGCCGCCTGGAGATCACCAACCTGAACGATACCCTGGAA
CTGGCCCTGATCTTTCGCAAAGTGGACATCAATAAGATGAATGGCTAA

Dtox_Spo0F-synth

ATGGGCGACTGCGATCTGCTGATCGTGGACGATCAGCCGGGCATCCGCCGTCTGCTGTAC
GAGGTGCTGAGCGAGGACGGCTATCGTGTGGAGGCAGTGGCCGGTGGTCGTGAAGCCATC
GACAAGGCCGTGCTGTTACGCCCGCGCCTGATCCTGCTGGACATGAAGATGCCGGGCATG
AACGGCCTGGAGACCCTGATCGAGCTGAACAAGGTGTACAAGGACGCCACCGCCGTGATC
CTGATGACCGCCTACGGCGAGCTGGAGATCGTGATGCAGGCCCAGAAGCTGGGCGTGAAC
CACTACATCAACAAGCCGTTCGACCTGGAGGACATCCGCGCCCTGATCAAGAGCCTGATC
CCGGACGGCGCAGAGATTGATCGCAGCCGCGACATCGTGTAA