Title

Expanding Cellulosomics: Investigation of Novel Cellulolytic Species and Transcriptional Regulation Mechanisms in Thermophilic Anaerobes

ISBN

9798297617094

Author

Takayesu, Allen Turner

Publication Date

2025-10-01

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Expanding Cellulosomics: Investigation of Novel Cellulolytic Species and Transcriptional Regulation Mechanisms in Thermophilic Anaerobes

A dissertation in partial satisfaction of the requirements for the degree Doctor of Philosophy in Biochemistry, Molecular and Structural Biology

by

Allen Turner Takayesu

2025

ABSTRACT OF THE DISSERTATION

Expanding Cellulosomics: Investigation of Novel Cellulolytic Species and Transcriptional Regulation Mechanisms in Thermophilic Anaerobes

by

Allen Turner Takayesu

Doctor of Philosophy in Biochemistry, Molecular and Structural Biology

University of California, Los Angeles, 2025

Professor Robert Thompson Clubb, Chair

Increasing damage from climate change, dwindling petroleum supplies, and the inefficiencies of first-generation biofuels have created a need for improved methods to cost-effectively produce chemicals, biomaterials, and biofuels from carbon-neutral sources. Lignocellulosic plant biomass is an abundant, renewable, and naturally occurring material ideal for bioconversion to these products that does not compete with food production. However, the recalcitrance of lignocellulose to hydrolysis limits its use in a larger-scale industrial setting due to high costs and low yields. One microbial solution to this problem has been the development of engineered microbes capable of efficiently degrading and converting lignocellulose into key bioproducts. At the forefront of this microbial approach are organisms capable of producing large extracellular enzyme assemblies called cellulosomes, which confer high cellulolytic activity to the bacteria that produce them.

In order to improve current efforts towards engineering industrial-use cellulolytic anaerobes, the work described in this dissertation seeks to understand how these organisms regulate the biogenesis and display of cellulosomes by studying the transmembrane receptor RsgI-proteins of *Clostridium thermocellum* (aka *Acetivibrio thermocellus*). Additionally, this work aims to both expand and characterize the set of known bacterial cellulosome-producers in order to reveal the diversity of reported cellulosome assemblies for potential use in ongoing engineering efforts. Chapter 2 of this dissertation describes the ectodomain from one of the nine RsgI-type transmembrane biosensors of *C. thermocellum*. The chapter presents the first electron density reconstruction of an RsgI ectodomain and investigates both the substrate-binding specificity and potential enzymatic activity of two of its three subdomains. In Chapter 3, I solved the solution state structure of the third subdomain within the RsgI ectodomain: an extracellular domain conserved between over 1000+ RsgI homologs. In addition to solving its atomic structure, this study employed molecular dynamics and proteomics experiments to further characterize the mechanism of autoproteolysis within the conserved domain. In Chapter 5, I discuss a broad comparative genomic screen that we conducted across 300k+ curated genomes to catalogue and identify all recorded and putative cellulosome producing bacteria. This genomic screen provided counts and predicted architectures for cellulosomal proteins in each investigated organism, identified 10 novel cellulosome producers, and provided insight into 4 distinct archetypes of cellulosome assemblies. Combined, my thesis research has expanded the genetic toolset available by characterizing the mechanism of RsgI activation in more complex cellulosome producing organisms and growing the set of known cellulosomal bacteria which will provide valuable insight for those attempting to engineer cellulosome producing organisms for bioconversion.

The dissertation of Allen Turner Takayesu is approved.

Carla Marie Koehler

Catherine F. Clarke

Robert P. Gunsalus

Robert Thompson Clubb, Committee Chair

University of California, Los Angeles

2025

DEDICATION

I would like to dedicate this work to my family, in all its forms.

To my parents, who spurred my curiosity and the drive to forever feed it,

To my sister, who has modeled resilience for me,

To my lab mates, peers, and lifetime friends who have redirected me back on course too many times, and

To Dan for his patience, love, and support

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOLWEDGEMENTS

I would like to thank my advisor, Dr. Robert T. Clubb for his mentorship and council throughout my entire doctoral research. My committee members: Dr. Cathy Clarke, Dr. Robert P. Gunsalus, and Dr. Carla Koehler for their input in both research and career advice. In particular, I would like to thank Dr. Robert P. Gunsalus for both his extensive help in providing us with biological samples and his microbiology insight as we expanded into the bioinformatics landscape for several projects. Thank you to many of UCLA's staff researchers: Dr. Mike Sawaya, for his kind words and structural biology expertise, Dr. Robert Peterson for his instruction in the NMR laboratories, and Dr. Mark Arbing for his troubleshooting guidance. Past and present members of the Clubb laboratory: Dr. Kat Ellis-Guardiola, Dr. Scott McConnell, Dr. Orlando Martinez, Dr. Chris Sue, Jess Soule, Andrew Goring, Jordan Ford, Nikki Cheung, Reece Pawlaczyk, Kylie Cheng, Benhur Abraham, and Esmeralda Mendoza Corrales for their constant support, encouragement, and advice throughout my degree. I am indebted to the work, support, and scientific insight provided by my close lab mate, project-partner, and friend, Chrstine Minor. Additional special thanks to Dr. Mahoney for his patience in fielding my constant questions and providing extensive training in NMR, MD, crystallography, and more. My undergraduate student Tobi Jessup for her contributions in the lab and with helping me to publish my first-author publication. I am eternally grateful for having worked in such a supportive lab environment and community.

Chapter two of this dissertation is a version of the publication: Mahoney BJ, Takayesu A, Zhou A, Cascio D, Clubb RT. The structure of the Clostridium thermocellum RsgI9 ectodomain provides insight into the mechanism of biomass sensing. Proteins. 2022 Jul;90(7):1457-1467. https://onlinelibrary.wiley.com/doi/10.1002/prot.26326. This work was supported by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research

program under Award Number DE-FC02-02ER63421 as well as for support from NIH grant R01-AI052217 [R.T.C]).

Chapter three of this dissertation is a version of the publication: Takayesu A, Mahoney BJ, Goring AK, Jessup T, Ogorzalek Loo RR, Loo JA, Clubb RT. Insight into the autoproteolysis mechanism of the RsgI9 anti-σ factor from Clostridium thermocellum. Proteins. 2024 Aug;92(8):946-958. https://onlinelibrary.wiley.com/doi/10.1002/prot.26690. This work was supported by grants from the National Institute of Health (NIH) (grants S10OD025073 and S10OD016336 for partial support of the NMR core facilities, as well as NIH grant R01-AI052217 [R.T.C]).

Chapter four of this dissertation is a version of the publication: Minor CM & Takayesu A, Ha SM, Salwinski L, Sawaya MR, Pellegrini M, Clubb RT. A genomic analysis reveals the diversity of cellulosome displaying bacteria. Front Microbiol. 2024 Oct 30;15:1473396. https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2024.1473396/full. This work was supported by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research program under Award Number DE-FC02-02ER63421. We also acknowledge support from NIH grant R01-AI052217 (R.T.C), a Cellular and Molecular Biology training grant T32GM145388 (CM), and the UCLA Quantitative & Computational Biosciences Collaboratory for assistance.

Chapter five of this dissertation is a version of the manuscript in preparation: Minor CM & Takayesu A, Ha SM, Salwinski L, Pellegrini M, Gunsalus RP, Arbing MA, Sawaya MR, Clubb RT AlphaFold Structural Proteomics Reveals Complex Cellulosome Machinery for Carbohydrate Degradation in Human Gut Bacterium *Ruminococcus Callidus* and Related Species.

VITA

2013 – 2017	Bachelors of Science, Biochemistry
	University of British Columbia
2020 - 2024	Teaching Assistant, Department of Chemistry & Biochemistry
	University of California, Los Angeles
2021	Candidate in Philosophy, Biochemistry and Molecular Biology
	University of California, Los Angeles
2022 – 2024	Nanoscience Education Program Instructor
	California NanoSystems Institute (CNSI), UCLA
2023	Southern California Users of Magnets (SCUM), Best Poster Award
2024	Protein Society 38th Annual Symposium, Best Poster Award
2024	Nanovation Competition Mentor
	California NanoSystems Institute (CNSI), UCLA
2025	Technology Fellow, Life Sciences
	Technology Development Group, UCLA

PUBLICATIONS

Minor CM & **Takayesu A**, Ha SM, Salwinski L, Sawaya MR, Pellegrini M, Clubb RT. A genomic analysis reveals the diversity of cellulosome displaying bacteria. Front Microbiol. 2024 Oct 30;15:1473396. doi: 10.3389/fmicb.2024.1473396.

Takayesu A, Mahoney BJ, Goring AK, Jessup T, Ogorzalek Loo RR, Loo JA, Clubb RT. Insight into the autoproteolysis mechanism of the RsgI9 anti-σ factor from Clostridium thermocellum. Proteins. 2024 Aug;92(8):946-958. doi: 10.1002/prot.26690.

Mahoney BJ, **Takayesu A**, Zhou A, Cascio D, Clubb RT. The structure of the Clostridium thermocellum RsgI9 ectodomain provides insight into the mechanism of biomass sensing. Proteins. 2022 Jul;90(7):1457-1467. doi: 10.1002/prot.26326.

Chapter 1

Introduction

1.1 Overview

The effects of climate change and finite petroleum supplies have created a pressing need to find alternative solutions to produce chemicals, biofuel, and materials from renewable and carbon-neutral sources. A promising microbial approach to this problem has been to utilize and engineer cellulolytic bacteria to efficiently degrade and convert abundant biomass to usable bioethanol and other similar bioproducts. Of particular interest are anaerobic thermophiles able to display megadalton, multi-enzyme extracellular complexes known as cellulosomes. Despite their promise as efficient biomass degraders, industrial-scale use of cellulosome producing microbes is hampered by challenging genetic manipulation of the organism and an incomplete understanding of cellulosomal regulation and biogenesis. More complex cellulosome producing microbes, such as *Clostridium thermocellum* (a.k.a. *Acetivibrio thermocellus*), contain a suite of extracytoplasmic function (ECF) sigma factors (SigI) and cognate transmembrane anti-sigma factor receptors (RsgI) responsible for controlling the expression and enzymatic composition of these cellulosomes. In this chapter, I summarize the current state of the microbial-based biofuel production and review the regulatory machinery, components, and general architecture of the cellulosome.

1.2 Lignocellulose Degradation and Processing

The financial and environmental strain from the increasing effects of climate change and over-consumption of limited petroleum resources has created a pressing need for the development of sustainable biofuel, chemical, and biomaterial production [1]. Though initially promising in use, first generation biofuels derived from foods such as corn starch have proven to be environmentally detrimental due to direct competition with food production, excessive cropland area, and increased fossil-fuel derived nitrogen fertilizer usage [2]. Lignocellulosic plant biomass (LCB) is considered one of the best candidate materials to use as a second-generation biofuel source, due to its high abundance (145 billion tons generated per year globally) and ability to be produced sustainably [3, 4]. Although there is a myriad of industrial techniques used to process LCB into usable products, current methods are both cost and product inefficient, with less than 5% of the LCB effectively utilized [5].

Lignocellulose is difficult to fully utilize by traditional processing methods due to both its chemical structure and composition. LCB is comprised of 3 major components, each contributing to LCB's recalcitrance in different ways: cellulose (30-50% dry weight), hemicellulose (20-35%), and lignin (10-25%) (Figure 1.1) [6]. Cellulose, a glucose homopolymer constructed with β-(1,4) glycosidic bonds, forms the core crystalline fibers of LCB, tightly held together by an extensive network of hydrogen bonding [7]. Surrounding these fibers are webs of branched hemicellulose heteropolymers consisting of both 5- and 6-carbon sugars as well as glucuronic acid whose abundance and composition varies between plant species [8]. This heterogeneity contributes most significantly to processing costs due to the required variety of enzymes needed to fully deconstruct the hemicellulose [9]. Lastly, the cellulose and hemicellulose are encased in a sturdy barrier of lignin, a non-crystalline branched phenolic polymer which helps to protect the fibers from oxidative, microbial, and chemical degradation [10].

1.2.1 Current Methods of Lignocellulose Biorefinery

The traditional approach of refining LCB into biofuels and chemicals employs the three-stage process of pretreatment, enzymatic saccharification, and fermentation. Pretreatment of the raw material is necessary in order to physically break up the three main components of the LCB, primarily the lignin and crystalline cellulose [11,12]. This can be accomplished using physical (e.g. mechanical crushing, hydrothermolysis, and irradiation), chemical (e.g. acid, alkali, oxidative treatment), or biological means (e.g. fungi or enzymatic treatment) [13,14]. The pretreated LCB then undergoes enzymatic saccharification wherein a blend of endoglucanases, exoglucanases, and β-glucosidases synergistically break down the hydrolysate into oligosaccharides and then further into monosaccharides [15]. Finally, the produced monosaccharides are converted into the desired fermentation products via an engineered *Saccharomyces cerevisiae* WXY12 strain in a large bioreactor [16].

Current LCB biorefinery methods are not yet suitable for large-scale industrial use owing to several financial and technological limitations. Most physical LCB pretreatment methods require a very high energy input to break up the material which is difficult to implement at industrial scale. Byproducts generated from chemical pretreatment such as furfural must be removed in an additional detoxification step before saccharification, thus increasing processing costs and generated toxic waste [17-19]. As a way to mitigate these effects, biorefineries use a cost-effective combination of both physical and biological methods via a steam-explosion treatment followed by biological pretreatment, leading to higher yields and fewer byproducts [20]. In addition to incurring the largest costs, the cocktail of purified cellulolytic enzymes during saccharification serves as a bottleneck in the biorefinery process from coordinating pH, temperature, and longer incubation time requirements [21,22]. The difficulties of efficient co-fermentation of hexose and pentose sugars and the limited tolerance to metabolic and ethanol byproducts pose additional challenges during the fermentation process [23,24].

1.2.2 Consolidated Bioprocessing

In order to overcome the problems faced by current LCB biorefineries, several research groups have looked towards naturally occurring bacterial and fungal-based biomass deconstruction models as a cost-effective and less labor-intensive alternative [25]. Consolidated bioprocessing (CBP) is an approach to LCB processing and refinement that combines cellulase production, saccharification, and fermentation into a single bioreactor step [26]. An effective CBP system would therefore significantly reduce the production cost by removing the need to produce and purify separate enzymes for a discrete hydrolysis step. As of now, there are three primary CBP models under investigation for industrial use: heterologous expression of cellulolytic genes in model organisms, co-cultured microbial consortia systems, and engineering naturally lignocellulolytic organisms [27].

Model organisms like *Escherichia coli*, *Bacillus subtilis*, and *S. cerevisiae* have a well-established suite of genetic tools available, making them targets for CBP development. Without natural cellulolytic activity, the construction of these CBPs require the heterologous expression of cellulases, hemicellulases, and lytic polysaccharide monooxygenases (LPMOs) [28]. While robust, these model organisms often still struggle to not only express the lignocellulolytic enzymes at a sufficient quantity but also to achieve proper folding and glycosylation of these enzymes for activity [29]. Another CBP strategy is the co-culturing of multiple species in a microbial consortium for an efficient division of labor in lignocellulose degradation and fermentation. Co-culture strategies try to mimic naturally occurring consortia but are usually limited to two to three bacterial and/or fungal species [27,30]. Inoculation times, ratios, pH, and metabolic byproduct build up must be carefully balanced and monitored. Near industrial titers of bioethanol (22 g/L) production have been achieved from the synergistic co-culturing of a cellulase producing bacteria (*Clostridium phytofermentans*) and fermenting yeast [31]. Construction of synthetic lignocellulosic consortia poses significant optimization challenges but

also shows significant promise as a viable CBP approach, especially given advancements in machine learning tools for mapping complex microbial, substrate, and enzyme interactions [32,33].

Several naturally occurring bacteria and fungi contain an array of Carbohydrate-Active Enzymes (CAZymes) with specific activities capable of biomass degradation. However, these organisms often lack the additional pathways required to produce the desired bioproduct or produce too many byproducts which inhibit growth and lower yield. The largest hurdle in engineering these microbes comes in the balance of maintaining high lignocellulosic activity with equal levels of biochemical and biofuel production. Lignocellulosic fungi such as *Trichoderma reesei*, *Aspergillus niger*, and *Myceliophthora thermophila* are promising as many of them produce the same cellulases used in industrial applications [34,35]. In addition to having an extensive set of genetic manipulation tools and established culturing protocols, fungal CBP candidates can achieve 2-5x faster cellulose utilization rates compared to bacteria [36,37]. Though many are thermostable at the optimal catalytic temperature for cellulases, fungal CBP candidates are obligate aerobes meaning cellulase and glycolysis genes are often inhibited under the hypoxic conditions experienced during LCB processing [38].

Alternatively, bacterial CBP candidates, namely thermophilic anaerobes such as *C. thermocellum*, *Clostridium cellulovorans*, and *Ruminiclostridium cellulolyticum*, thrive in industrial conditions because they proliferate rapidly in high temperatures, have a high tolerance for pretreatment byproducts, and are natural bioethanol producers [39]. Of high interest are bacterial CBP organisms that owe their potent cellulolytic ability to the elaborate multi-enzyme surface structures they produce called cellulosomes, discussed further in Section 1.5 (Figure 1.2) [40]. A slow adoption of these bacteria for industrial scale use stems primarily from a lack of genetic editing tools for these anaerobes, an incomplete understanding of cellulosome biogenesis, and low ethanol titers. By gaining an understanding of the molecular machinery

responsible for their efficient lignocellulose degradation and its regulation, CBP development becomes more viable as an alternative biofuel refinement strategy.

1.3 Bacterial Sigma Factors

Bacteria must be able to effectively react and respond to a plethora of extracellular stimuli in order to continue to survive and propagate. As such, they contain an extensive genetically encoded toolkit of adaptive mechanisms for responding to different environmental cues. Control of these adaptive systems happens at all levels of the central dogma – transcriptional, translational, and post-translational – with the initiation of alternative transcription acting as the vital initial step to this process [41,42]. The primary mediators of transcription initiation in bacteria come in the form of sigma (σ) factors. Bacterial σ factors are essential for transcription initiation, acting as the specificity subunit for the RNA polymerase (RNAP) holoenzyme [43]. In the RNAP complex, the σ factor subunit serves three primary functions: it helps to stabilize the other five subunits ($\alpha_2\beta\beta'\omega$) within the holoenzyme, makes direct contact with the DNA via the -35 bp and -10 bp promoter elements upstream of the transcription start site, and contributes to DNA strand separation during promoter escape [43,44]. σ factors preferentially recognize unique DNA promoter sequences and can therefore direct the transcription of distinct sets of genes and operons, or regulons [45]. In controlling the access and availability of these σ factors, bacteria have effective control over a wide variety of transcriptional profiles for use in stress and/or environmental responses, pathogenesis, and cell growth [46,47].

1.3.1 σ^{70}-Family σ Factors

There are two structurally distinct flavors of σ factors in bacteria: σ^{70} (or SigA in gram-positive bacteria), comprised of primary and some alternative σ factors, and σ^{54}, which is

responsible for the regulation of several other adaptive cellular responses and virulence factors. The σ^{70} family encompasses nearly 20 times more members than those within the σ^{54} family and can be subdivided into four distinct groups based on sequence conservation, domain organization, and to a lesser extent, the types of genes they regulate [48]. These four distinct groups within the σ^{70} family are defined by the presence or absence of four conserved helical subdomains (σ_1 – σ_4), which themselves can be further subdivided into ~2-4 subregions for each subdomain (Figure 1.3) [49,50]. These subdomains mediate important interactions with the promoter DNA of the genes they help to transcribe, with the C-terminal σ_4 subdomain interacting with the -35 promoter region and the σ_2 and σ_3 subdomains forming contacts with the -10 region [51].

Group 1, or “Primary”, σ^{70} proteins contain all 4 subdomains and are associated with the regulation and transcription of housekeeping genes [52]. Group 2 σ factors are non-essential σ factors usually involved in regulating genes implicated in stress responses during the stationary phase and contain all 4 major subdomains like group 1 but lack a $\sigma_{1.1}$ [53,54]. Flagellum biosynthesis, heat shock, cell wall stress, and sporulation genes constitute the four broad regulomes under the control of Group 3 σ factors which contain only σ_2, σ_3, and σ_4 subdomains [55,56]. Lastly, the remainder of this section will focus on Group 4, or Extracytoplasmic Function (ECF) σ factors, which are associated with genes activated in response to stimuli sourced from outside the cell or from within the cell membrane.

1.3.2 Extracytoplasmic Function (ECF) σ Factors

ECFs account for the largest and most sequence diverse of the four σ^{70} families, consisting of over 43 major phylogenetically distinct sub-groups and involved in the regulation of a multitude of cellular stress responses as well as virulence pathways [57,58]. These regulons include genes involved in responses to oxidative stress, envelope stress, iron transport, nutrient limitation, and denatured extracellular proteins [59,60]. ECFs are the smallest of the σ^{70} factors

with an average length of 200 residues, only containing the essential σ_2 and σ_4 conserved subdomains for promoter recognition as well as a non-conserved σ_2 / σ_4 linker connecting them [61]. *In vitro* structures of the individual σ_2 and σ_4 subdomains with corresponding -10 and -35 promoter elements, respectively, suggest an alternative mode of DNA binding during transcription [62,63]. Indeed, both crystal and cryo-EM structures of the full transcription initiation complex containing an ECF suggest that the non-conserved σ_2 / σ_4 linker plays a key role in transcription initiation akin to the structurally similar $\sigma_{3.2}$ subdomain [64-66].

In addition to their distinct structure and DNA binding mode compared to the other σ^{70} families, ECFs are defined by three unique functional hallmarks as well. Despite the drastically smaller size and set of recognition elements compared to Group 1-3 σ factors, each Group 4 σ factor recognizes a unique promoter and is much less promiscuous to promoter sequence identity as is seen with primary σ factor binding [41,67]. This stringency in promoter recognition also suggests why there is little cross talk seen between regulons controlled by ECF σ factors [68]. Regulomes involving ECFs take full advantage of this specificity via positive autoregulation, as each ECF gene has its corresponding ECF-specific promoter upstream of it [60]. Lastly, ECF σ factors are often defined by the presence of a cognate anti-σ factor as a negative regulator and binding partner. These anti-σ factors are modular, using a conserved domain to hold their cognate σ factor in an inactive state until an extracellular sensory domain detects a signal, either directly or through a signal-relay system, thus releasing the σ factor for transcription initiation. The ECF/anti-σ factor system will be discussed in more detail in the following section with other similar regulatory mechanisms seen in cellulolytic bacteria.

1.4 Regulatory Systems in Cellulolytic Bacteria

Microbial LCB degradation presents a cleaner alternative for converting cellulosic biopolymers into usable simple sugars or fermentation products and chemicals. However, the slower processing times and lower yields currently capable with microbial-based methods present significant hurdles to their industrial use [69]. In order to fully harness the potential of these CBP organisms, research groups have focused primarily on two different approaches: Using model organisms such as *S. cerevisiae* for the heterologous expression of cellulolytic enzymes and genetic engineering of already existing, highly lignocellulolytic species such as anaerobic thermophiles and fungi [70]. In both approaches, an understanding of the core regulatory machinery and mechanisms in these microbes is crucial to both improving cellulolytic activity and biofuel production. This section will summarize the four common regulatory systems employed by gram-positive and gram-negative bacteria for the expression of CAZyme genes: Hybrid Two-Component Systems (HTCSs), Carbon Catabolite Repression (CCR), Selective RNA Processing and Stabilization (SRPS), and ECF/anti-σ factor systems (Figure 1.4).

1.4.1 Hybrid Two-Component Systems

HTCSs are ubiquitous and highly conserved signaling mechanisms that utilize a sensor histidine kinase (HK) that, when activated, triggers a phosphorylation cascade eventually leading to alternative transcription [71]. Extracellular signals or substrates, such as a cellulose polymer, are sensed by the sensor domain of the HK leading to signal transfer through the membrane to initiate autophosphorylation of the HK at a conserved histidine [72]. After phosphorylation of the HK, the phosphoryl group is transferred to a response regulator (RR) domain, triggering a conformational change of the RR into an active transcriptional regulator to modulate transcription [73,74]. Even though phosphorylation between paralogous, non-cognate HTCS domains could enable sizeable cross talk networks to form, stringency between HK/RR pairs is maintained due to specific interface residues and tight control of HK/RR stoichiometry [76-79]. In cellulolytic microbes, HTCS systems have been noted to regulate the expression of

CAZyme and related transporters in both gram-negative (e.g. *Cellvibrio japonicus*, *Bacteroides thetaiotaomicron*) and gram-positive bacteria (e.g. *Ruminiclostridium cellulolyticum*, *Geobacillus stearothermophilus*) [80-83].

1.4.2 Carbon Catabolite Repression

CCR encompasses a broad set of different regulatory mechanisms which all allow microbes to selectively use preferred carbon substrates – often glucose – in a mixed media by modulating the expression of secondary catabolic systems [84]. CCR manifests differently in gram-positive and gram-negative microbes. Basal expression of genes for non-preferred carbon source metabolism in gram-negative bacteria is normally very low [85]. Low levels of glucose cause an accumulation of phosphorylated EII_A, a subunit of a glucose transport complex, which in turn phosphorylates and activates adenylate cyclase. The buildup of cAMP from adenylate cyclase activity then leads to an activation of cAMP receptor proteins that promote the upregulation of alternative carbon metabolism genes [86]. Unlike in gram-negative bacteria, gram-positive microbes actively suppress expression of non-glucose metabolism genes by employing a negative transcriptional regulator in the form of the carbon catabolite protein A (CcpA) and phosphorylated histidine-containing protein (HPr ~ P) [87]. In the absence of glucose, and therefore low fructose 1,6-bisphosphate (FBP) levels, HPr cannot be phosphorylated and dimerize with CcpA to bind to operator catabolite responsive elements (*cre*) [88]. As a result of binding to the *cre* box downstream of the transcription start site, transcription elongation of non-glucose metabolism genes is blocked [89,90].

CCR poses a significant obstacle to genetically engineering CBP microbes to simultaneously metabolize the pentose and hexose sugars present in LCB, leading to longer fermentation times and added costs. As such, there is a significant trade-off within the CBP field as engineered model organisms which can co-ferment C5/C6-carbon sources without a CCR effect, such as *S. cerevisiae* and *E. coli*, are not naturally cellulolytic like an ideal CBP organism

[91,92]. In spite of this, engineered yeast continues to be a strong model for CBP development as several attempts have shown moderate success at increasing the cellulolytic activity by increasing cellulose-adherence or via heterologous expression of extracellular cellulolytic machinery [93,94]. Other groups have successfully demonstrated that co-utilization of pentose and hexose sugars is possible in engineered organisms with a CCR system like *C. thermocellum* by introducing exogenous xylose-catabolizing enzymes [95].

1.4.3 Posttranscriptional Regulatory Mechanisms

Cellulolytic microbes are also able to regulate relevant lignocellulose degradation genes at the posttranscriptional level in several ways. One such regulatory mechanism used in both gram-positive and gram-negative microbes is Selective RNA Processing and Stabilization (SRPS). As with all bacterial genomes, lignocellulolytic bacteria often organize functionally related genes into operons which are often transcribed as a single, polycistronic mRNA transcript. Secondary structural elements within the resultant transcripts in the form of stem-loops can modulate the stability of the mRNA by making it more recalcitrant or susceptible to exo- and endonuclease degradation [96]. The longer the stem-loop formed between genes within the transcript, the more stable the gene within the transcript. Clostridium and related cellulolytic species such as *C. thermocellum*, *Acetivibrio cellulolyticus*, and *R. cellulolyticum* employ this strategy in certain operons containing a mixture of CAZymes and scaffolding proteins used to arrange them on the cell surface [97-99]. In this way, microbes are not only able to control the abundance of select transcripts but also their stoichiometries. Strategic introduction of stem loops into operons and between genes also offers another tool for the genetic engineering of CBP Clostridia species.

The use of RNA regulatory machinery in cellulolytic microbes can also overlap with CCR processes. In *Ruminiclostridium papyrosolvens*, a cellulolytic thermophile and anaerobe, the 5'-UTR of a polycistronic mRNA containing a cluster of xylanases has been shown to block

transcription [100]. In the presence of some alternative carbon sources, the 5'-UTR suppression is released, suggesting the presence of a riboswitch-like mechanism that allows for expression of cellulolytic genes only when the appropriate carbohydrate substrate is present. Other cellulolytic microbes also make use of *cis*-encoded small regulatory RNAs (sRNA) to control transcription levels of genes dedicated for polysaccharide transport and utilization, such as in several *Bacteroides* species [101]. Bacterial antisense *cis*-encoded sRNA is transcribed from the complementary strand of their target gene, which leads to a lack of translation or complete degradation of the original transcript due to the strong RNA duplex formed [102].

1.4.4 ECF/anti-σ factor Systems

Similar to a two-component system, ECF/anti-σ factor systems use a transmembrane protein capable of sensing an external signal which is transmitted through the membrane to trigger the activity of a cytoplasmic cognate protein. Within this system, the anti-σ factor is typically comprised of a C-terminal, extracellular sensing domain and a conserved cytoplasmic N-terminal domain which binds to its cognate ECF-σ factor [60]. The C-terminal sensor domain is highly varied between anti-σ factors, each determining the stimuli that activates it, with signals ranging from substrates like glycans, metals, and chemical stressors or mechanosensing signals such as cell-wall stress [103-106]. Cognate binding between ECF/anti-σ factor pairs is specific, but potential cross talk between non-cognate pairs has been reported in rare cases such as RpoE in *Azospirillum brasilense* and FpvR in *Pseudomonas aeruginosa* [107,108].

ECF/Anti-σ factor pairs present a wide sequence and structural diversity, with 157 ECF groups currently classified amongst both gram-positive and gram-negative bacteria, several of which do not involve transmembrane sensor proteins [109]. The σ^E and σ^R of *Rhodobacter sphaeroides* and *Streptomyces coelicolor*, respectively, are held by soluble anti-σ factors which undergo a conformational change in response to different forms of redox stress experienced by the cell [110-112]. Some soluble ECFs rely on a C-terminal regulatory domain attached to the

ECF (e.g. CorE from *Myxococcus xanthus*, ECF41, and ECF42) to control activity [113-115]. Genomic screens of similar soluble ECF families such as ECF52 and ECF53 have also predicted the presence of hemicellulolytic glycosyl hydrolase domains [59]. Other forms of ECF regulation can be found across other non-cellulolytic species, including ECFs activated via membrane-bound Ser/Thr kinase phosphorylation [116] and systems that combine two-component systems and ECFs using an ECF σ factor mimic as an anti-anti-σ factor [117,118].

Most ECF/anti-σ factor pairs involved in lignocellulose gene regulation are transmembrane proteins that are activated via Regulated Intramembrane Proteolysis (RIP). RIP is a three-step process wherein the membrane-anchored anti-σ factor is successively degraded by a set of three separate proteases/protease complexes. Signal activation via the C-terminal domain of the anti-σ domain allows access for a membrane-anchored protease (e.g. DegS, PrsW) to proteolytically cleave a conserved site (Site-1) within the extracytoplasmic domain [119-121]. With the release of the cleaved extracytoplasmic domain, an additional intramembrane protease (e.g. RseP, RasP) is free to make a proteolytic cut at second site within the transmembrane helix of the anti-σ factor (Site-2) causing the subsequent release of the N-terminal anti-σ factor and bound cognate ECF [122,123]. Lastly, the remainder of the anti-σ factor is degraded by the work of cytoplasmic ClpXP proteases, allowing the ECF to initiate transcription of cellulolytic genes [124].

Select cellulosome producing bacteria discussed in this work utilize transmembrane anchored ECF/anti-σ factors which have been implicated in cellulosome biogenesis and regulation [125,126]. Indeed, a majority of the microbes that produce more complex arrangements of cellulosomes contain a suite of anti-σ factors which were first identified due to their shared sequence homology to “Regulator of σ^{I}” (RsgI) in *B. subtilis* [128]. Organisms such as *C. thermocellum* and *Pseudobacteroides cellulosolvens* can encode as many as 9 (Figure 1.5) and 14 RsgI proteins, respectively [125,127]. Extracellular sensor domains for these RsgI

often take the form of carbohydrate binding modules (CBMs), CAZymes (e.g. GH10), or other domains predicted to interact with other polysaccharide substrates (e.g. PA14, FN3). Different RsgI proteins within the same species can also encode extracellular domains of the same predicted carbohydrate binding domain. While several studies have confirmed the substrate sensing capability of a few RsgI within these cellulosome producing species, it is unclear why this redundancy occurs [129-131]. Prior to the work described in this thesis, the precise mechanism of signal activation at a site-1/2 for these RsgI-type proteins was also poorly understood.

1.5 The Cellulosome

As described in previous sections, cellulosome producing bacteria are high-interest targets for CBP development given their naturally high lignocellulolytic activity. Cellulosomes, first identified in 1983 as large extracellular protrusions, are extracellular megadalton structures comprised of multiple cellulolytic enzymes assembled onto a series of protein platforms (Figure 1.2) [132,133]. At the core of all cellulosomes are two fundamental protein domain building blocks which mediate the interactions holding the assembly together: cohesins and dockerins [40,134,135]. Multiple cohesin (Coh) domains on a single polypeptide, called “scaffoldin”, serve as protein loading sites for their non-covalent binding partners, dockerins (Doc), which are themselves covalently linked to cellulolytic enzymes or other scaffoldin molecules. Cohesin-dockerin (Coh:Doc) binding pairs represent very high-affinity protein-protein interactions (K_d~ < 10^{-9} – 10^{-11} M) [136]. Dockerins and cohesins of the same type are able to change binding partners in a modular fashion, allowing for new and/or different Doc-enzymes to populate the scaffoldin to change cellulosome activity [137].

The structure of the cellulosome confers higher degrees of lignocellulosic activity compared to bacteria who freely secrete CAZymes for three key reasons. (1) Enzyme co-localization within the cellulosome promotes enzyme-enzyme synergies and enzyme-proximity effects [138,139]. The incorporation of enzymes of multiple substrate-specificities and complementary activities onto the same protein-loading platform allows for fast and efficient digestion of complex polysaccharides. Enzyme-enzyme synergy also manifests by limiting product inhibition effects [140]. (2) Carbohydrate-binding domains present on both the dockerin-fused glycoside hydrolases (DocGH) and the scaffoldins themselves facilitate cellulose-enzyme-microbe contact, enabling the host microbes to have easy access to the resultant hydrolyzed products [141]. (3) Due to its modular nature, cellulosomes can and do change their enzyme composition in response to the presence of different substrates, ensuring that the correct CAZymes are displayed to degrade different carbohydrate polymers [142]. Additionally, scaffoldins containing dockerins not only attach to the cell surface but can also enable different scaffoldin-scaffoldin arrangements, thus creating poly-cellulosomal structures, with some cellulosomes capable of incorporating up to 160 enzymes [134,143].

In the remainder of this section, I will describe the individual components that comprise cellulosomes (cohesins & dockerins), how their known structures contribute to activity, and the diversity of cellulosome architectures seen in nature.

1.5.1 Cohesins & Dockerins

Cohesins-dockerin pairs (Coh:Doc) can be divided into different families or types based on sequence homology, which also roughly translates to the distinct roles they play in cellulosome assembly [135]. Generally, type I is associated with enzyme loading between scaffoldin cohesins and DocGHs whereas type II includes Coh:Doc pairs that join together two scaffoldins noncovalently. There are two additional Coh:Doc types that uniquely appear in select genera or species; including type III which exclusively appears in *Ruminococcus* cellulosome

producers and type R which has only been recorded in *Pseudobacteroides cellulosolvens* [144-146]. Cohesin-dockerin binding is both type-specific and species-specific, barring a few exceptions where cross-species type II Coh:Doc binding was observed [137,147]. Coh:Doc conventions are also broken in *P. cellulosolvens* where the functional roles of type I and type II cohesins and dockerins are reversed [146].

Although they can be divided into distinct sequence homology-based families, overall cohesin and dockerin topologies are conserved between types. A canonical cohesin domain contains ~150 residues arranged in a two-sheeted, nine β-stranded jelly roll-type fold, with strands 4,7,2,1,9 in one sheet and strands 5,6,3,8 on the other [148]. The flat second sheet (β5,6,3,8) acts as the hydrophobic interaction surface for Coh:Doc binding. Dockerin domains are much smaller, consisting of only ~70 residues arranged into three α-helices (α1-3) with two additional EF-hand Ca^{2+}-binding loop-helix structures at the N-terminal ends of α1 and α3 [149,150]. Conserved aspartate and/or asparagine residues are present in both EF-hand motifs for coordinating the Ca^{2+} which is essential for proper dockerin folding and therefore Coh:Doc binding. The intervening α2 helix stabilizes both the EF-hands and the two ~22 residue α1 and α3 helices, which are nearly symmetrical in sequence [151].

As its symmetry about two of its α-helices suggests, some dockerins are capable of adopting a dual-binding mode, wherein the two possible binding conformations for a Coh:Doc pair are related by a 180° rotation [152,153]. Though initially believed to occur only in type I Coh:Doc pairs for sterically optimizing Doc-enzyme incorporation into the cellulosome, the dual-binding mode has also been observed in Type II -III Coh:Docs [154]. Conversely, lack of 2-fold symmetry between the two helices, occurring either naturally or from genetic engineering, abolishes dual-binding capability while still maintaining a high-affinity Coh:Doc interaction. Select dockerins have an additional intramolecular clasp formed via stacking interactions between a key proline and an aromatic residue located on the N- and C-terminus of helix α1

and α3, respectively [155]. Not only does this contribute to the stability of the dockerin but also may play a role as an allosteric regulator in determining the favored binding mode of the dockerin [156]. Some cellulosome producers have pushed the utility of Coh:Doc binding modes even further. For example, *A. cellulolyticus* contains asymmetrical dockerins that have a dual-cohesin specificity, suggesting a potential partner switching role in cellulosome assembly [157]. In *R. flavefaciens*, a truncated group of dockerins (i.e. only containing 2 α-helices and one EF-hand motif) exhibited a tripartite binding mode *in vitro*, capable of binding two separate cohesin domains simultaneously [158].

While scaffoldins serve as the assembly platform, dockerin-fused enzymes constitute the workhorses of the cellulosome. Expectedly, cellulosomal enzymes are primarily made of glycoside hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate esterases (CE), and other CAZymes [159]. Doc-enzymes like DocGHs are modular, typically comprised of one dockerin domain, a short linker, and often an additional CBM to improve enzyme substrate proximity [160,161]. Doc-enzyme linkers are low complexity and tightly packed, which contributes to the rigid, extended structures that the proteins adopt [162,163]. Dockerin-fused genes are not restricted to just CAZymes. Other dockerin-fused genes have predicted functionalities such as serpins, proteases, and several domains of unknown function [164-166].

1.5.2 Cellulosome Types and Architectural Diversity

Categorization of different cellulosomal systems is chiefly determined by the types of scaffoldins that make up them. Broadly, scaffoldins are cohesin-containing proteins that facilitate the assembly of additional enzymes and/or other scaffoldin proteins. Genomically, scaffoldin genes are clustered together in a *cip-cel* operon often along with additional DocGH enzymes all under the control of a SigI/RsgI promoter [99]. Like Doc-enzymes, scaffoldins are modular. Threonine-rich linkers between the multiple cohesin domains confer a high degree of flexibility, allowing the full scaffoldin to switch between extended and compact states for optimal enzyme

positioning [167]. To account for this flexibility and to combat the strong mechanical forces present near the cell-surface, cohesins found at the connection points between the cell and the substrate are more robust than those found outside of these anchoring points [168]. Besides cohesins, scaffoldins can also contain several other types of modules such as a commonly present CBM for cellulose adhesion or structural X2 domains which are thought to improve adjacent dockerin stability and even participate in type II Coh:Doc recognition [169,170].

Scaffoldin proteins can be categorized into four general types based on their function [159]. Primary scaffoldins are the main docking platform for DocGHs and other cellulosomal enzymes. In turn, primary scaffoldins are attached to the cell-surface either directly in the case of simpler cellulosome systems seen in *R. papyrosolvens* and *R. cellulolyticum* or via a Coh:Doc interaction to an anchoring scaffoldin for more elaborate cellulosomes such as *C. thermocellum* and *P. cellulosolvens* [146,171,172]. Anchoring scaffoldins are cohesin-containing proteins that contain an additional domain responsible for cell-surface attachment, either covalently (e.g. sortase-mediated attachment) or non-covalently (e.g. Surface Layer Homology (SLH) domains) [145,173,174]. The remaining two types, cell-free and adaptor scaffoldins, are exclusively seen in complex cellulosome producers just as anchoring scaffoldins are. These scaffoldins improve the reach and size of cellulosomes by acting as non-dockerin fused enzyme-loading platforms (cell-free scaffoldin) or as means to attach additional scaffoldins or enzyme types to the cellulosome (adaptor scaffoldins). Further details on the categorization of cellulosome types, namely simple and complex, as well as the scaffoldin types that comprise them will be discussed in chapter 4.

1.6 Scope of Dissertation

Research described in this thesis focuses on elucidation of RsgI signal activation in cellulosome producing anaerobic thermophiles as well as the characterization of cellulosomal genes and predicted assemblies across all genera. In particular, this work helped to define the

precise mechanism through which RsgI proteins become primed for signal activation via autoproteolysis. Additionally, this research provides a comprehensive set and categorization of cellulosome types, components, and observed patterns which were previously disorganized within the cellulosomics field. With this work, I hope to expand the genetic toolset available by identifying essential cellulosome regulatory machinery which will provide valuable insight for those attempting to engineer cellulosome producing organisms for bioconversion of LCB into bio-based materials, chemicals, and biofuels.

CHAPTER 2 describes the structure of the biomass-sensing ectodomain from RsgI9, an RsgI anti-σ from *C. thermocellum* that lacks both a predicted cognate σ^{I}-factor and substrate [175]. In this work, we presented the first electron density reconstruction of an RsgI ectodomain as well as obtained an atomic-resolution structure for the NTF2-like/S1Peptidase bidomain present within RsgI9. The shape and rigidity of all subdomains within the ectodomain were analyzed using a combination of SAXS and crystallography data. Additional biochemical characterization of the ectodomain's binding capability and possible proteolytic activity were examined as well.

In CHAPTER 3, I determined the solution state structure of a key conserved regulatory domain in RsgI9 which is conserved among 1000+ homologous RsgI-type transmembrane anti-σ factors [176]. The structure, supported by proteomics and molecular dynamics experiments, provided valuable insight into mechanism responsible for triggering autoproteolysis within the conserved domain. Further studies of this system, particularly into the processes involved in the physical detachment of the autoproteolyzed ectodomain upon signal activation, may lead to an improved regulatory toolbox for CBP engineering in cellulosome producers.

CHAPTER 4 describes a comparative genomic analysis of cellulosome producing organisms across all phyla [177]. This work provides a comprehensive analysis of cellulosomal counts and assemblies across all cellulosome producing organisms, which was previously

poorly annotated despite several detailed reviews and publications. The genomic-centric approach of 300k+ curated genomes led to the categorization of all known and potential cellulosome producers, including the identification of 10 novel putative cellulosome producing organisms. Enzyme specificities, enzyme counts, and predicted cellulosome architectures between genera are also compared. This analysis provides both a catalogued inventory of known cellulosomal species but also a strong framework through which additional species for future CBP development may be identified.

The work in CHAPTER 5 builds upon that of the previous chapter by implementing a structural homology based screen in *Ruminococcus* type species in order to identify cryptic cohesins previously not detected based on sequence homology search methods. From this screen, we have identified 170 previously unidentified putative cohesin domains across 26 *Ruminococcus* species. We report *R. callidus* as a novel cellulosome producing microbe containing at least two new potential cohesin types, for which I have determined a representative crystal structure of. This work lays the foundation for using similar structural-homology based methods in identifying sequence-divergent families of proteins, which can be further used to identify additional cellulosomal microbes.

Lastly, I attempted to find the cognate σ factor for RsgI9 and to demonstrate potential crosstalk interactions between non-cognate SigI/RsgI pairs which I outline in the APPENDIX. In order to search for a “SigI9”, I utilized a pull-down affinity assay using an immobilized RsgI N-terminal anti-σ factor domain as bait for a *C. thermocellum* lysate prey. Co-eluted proteins were visualized by SDS-PGE and subsequently excised for identification via LC-MS/MS. I also employed 2D-NMR chemical shift perturbation (CSP) studies on purified ^{15}N-labeled N-terminal RsgI and unlabeled C-terminal SigI fragments to determine if RsgI/SigI crosstalk was feasible in *C. thermocellum*.

1.7 Figures

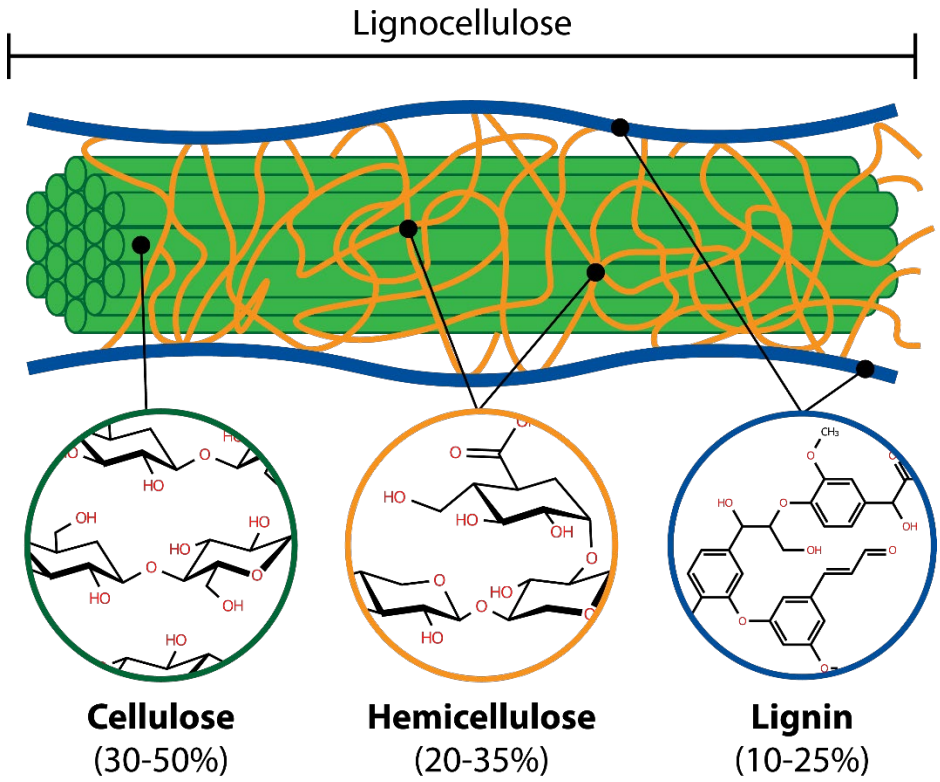


Figure 1.1 Structure of lignocellulosic biomass (LCB) in plants

Cellulose fibers made from rigid long chains of β-(1,4)-linked d-glucose monomers stabilized by hydrogen bonding makes the core of LCB. Hemicellulose is a branched heteropolysaccharide that is comprised of both hexoses (e.g. glucose, galactose, mannose) and/or pentoses (e.g. arabinose, xylose) with additional uronic and acetic acid groups which decorate the main chain. The hemicellulose forms a webbed matrix which encases the cellulose fibers which is further surrounded by a layer of lignin, a rigid and recalcitrant heteropolymer made of phenolic monomers (e.g. p-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol) [178].

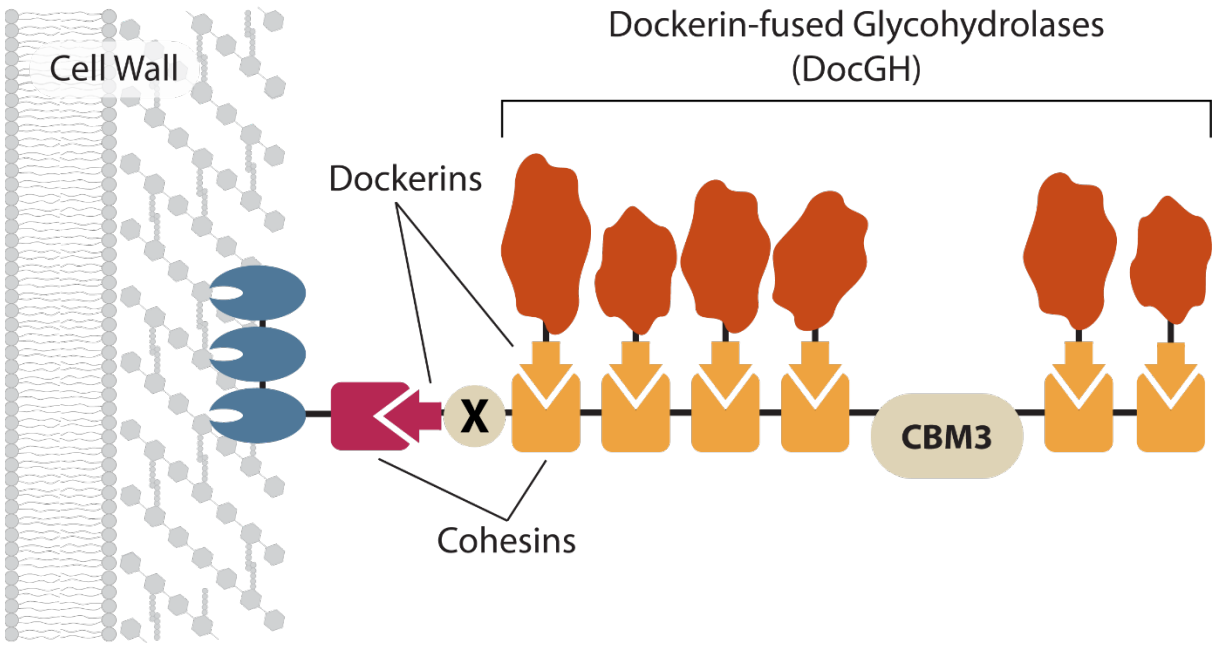


Figure 1.2 General structure of naturally occurring cellulosomes

Cohesin and dockerin domains form the core non-covalent interactors responsible for construction of cellulosomal assemblies. Multi-cohesin proteins called “scaffoldins” act as protein-loading platforms for dockerin-fused proteins which take the form of dockerin-fused enzymes (e.g. Glycohydrolases, DocGHs) or other scaffoldins with a C-terminal dockerin. Multiple DocGHs of varying activities can be loaded onto a single primary scaffoldin, thus increasing cellulolytic efficiency by holding the enzymes, substrate, and cell in close proximity. The primary scaffoldin in turn is non-covalently attached to the cell surface, often through a cohesin-dockerin interaction to a different surface-attached scaffoldin. Scaffoldins can also include several additional accessory domains, such as a carbohydrate-binding module (CBM) or an X-domain (X) which can aid in substrate adhesion and structural stability, respectively.

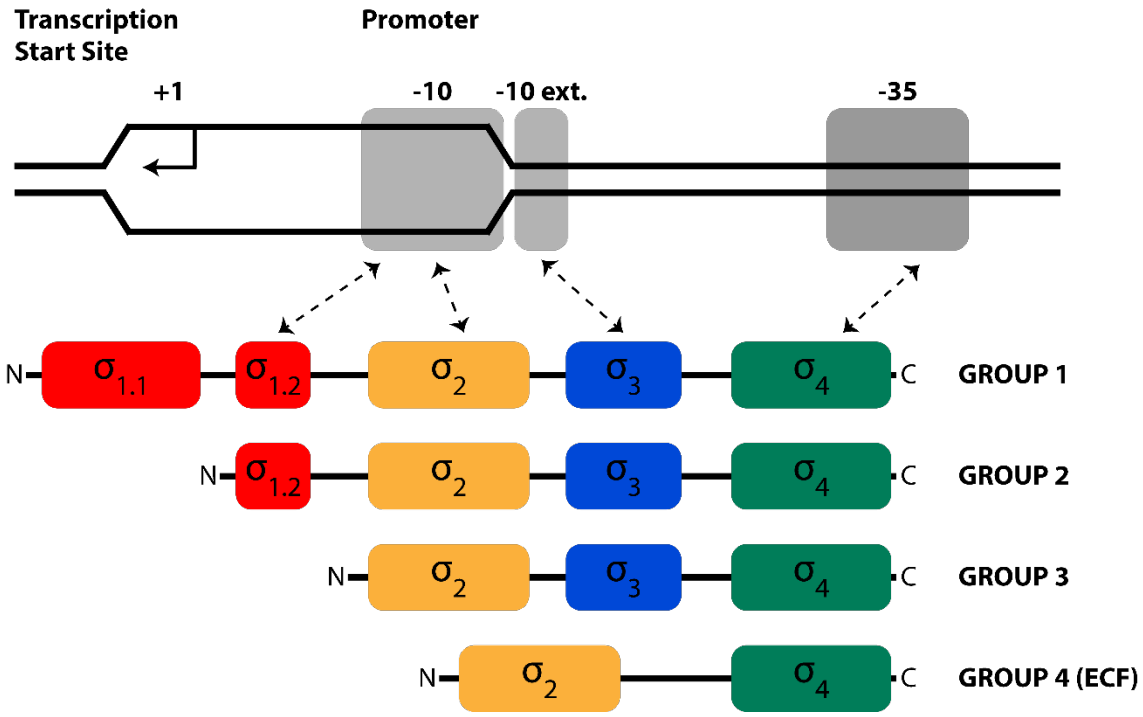


Figure 1.3 Domain organization for the four σ^{70}-family σ factors and subdomain-promoter recognition regions.

A combination of the four subdomains present in the primary σ^{70} are conserved between the four σ^{70}-family σ factors in bacteria. Group 1 σ factors have all 4 major subdomains present whereas ECFs only contain the σ_2 (yellow) and σ_4 (green) which are necessary for making key contacts with the -10 and -35 elements (light grey), respectively, of the promoter region for the genes they help transcribe. The presence of an additional σ_3 subdomain (blue) in Group 1-3 σ factors allows for interaction with the two nucleotide extended -10 elements. The $\sigma_{1.2}$ subdomain (red) unique to Group 1 and 2 σ factors and makes contacts a "discriminator" element immediately downstream of the -10 sequence [51].

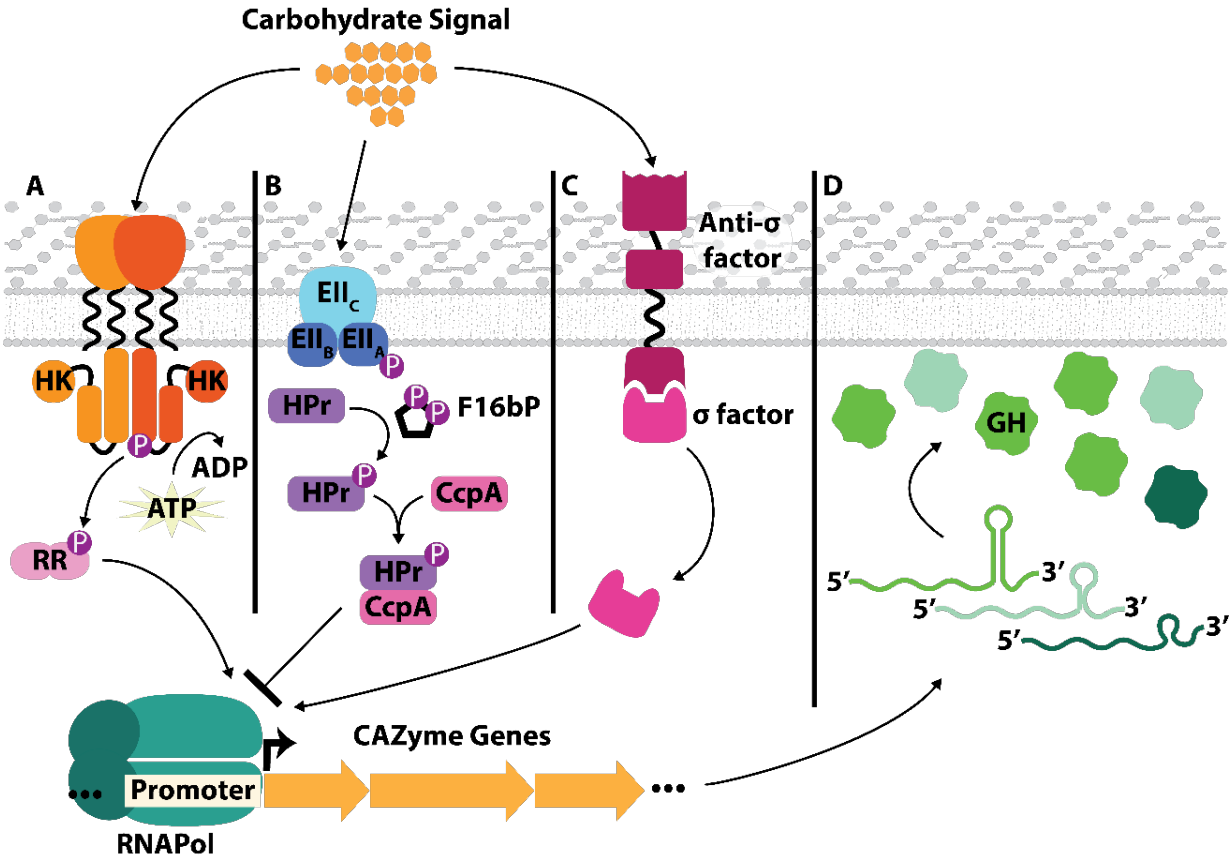


Figure 1.4 Transcription regulatory systems in cellulolytic bacteria

Cartoon schematic of the four major types of transcriptional control of genes related to cellulose deconstruction and metabolism in bacteria. (A) Hybrid Two-Component systems. After signal activation, a histidine kinase (HK) phosphorylates a corresponding response regulator (RR) to recruit the RNA-polymerase (RNAPol) to CAZyme-containing operons. (B) Carbon Catabolite Repression (CCR). In gram-positive bacteria, while glycolysis is active, there is an accumulation of fructose-1,6-bisphosphate which leads to the phosphorylation of Histidine Protein (HPr). After phosphorylation, HPr dimerizes with Carbon-Catabolite Control Protein A (CcpA) to block transcription of genes responsible for metabolizing non-preferred carbon sources (C) ECF/anti-σ factor systems. Signal activation leads to the degradation of the anti-σ and release of the cognate σ factor to bind with RNAPol (D) Selective RNA Processing and Stabilization (SRPS). The presence of larger end-stem loops formed at the 3'- (or 5'-) end of CAZyme-encoding (e.g.

Glycoside hydrolases, GH) transcripts protects the transcript against endonuclease activity, thereby controlling transcript abundance.

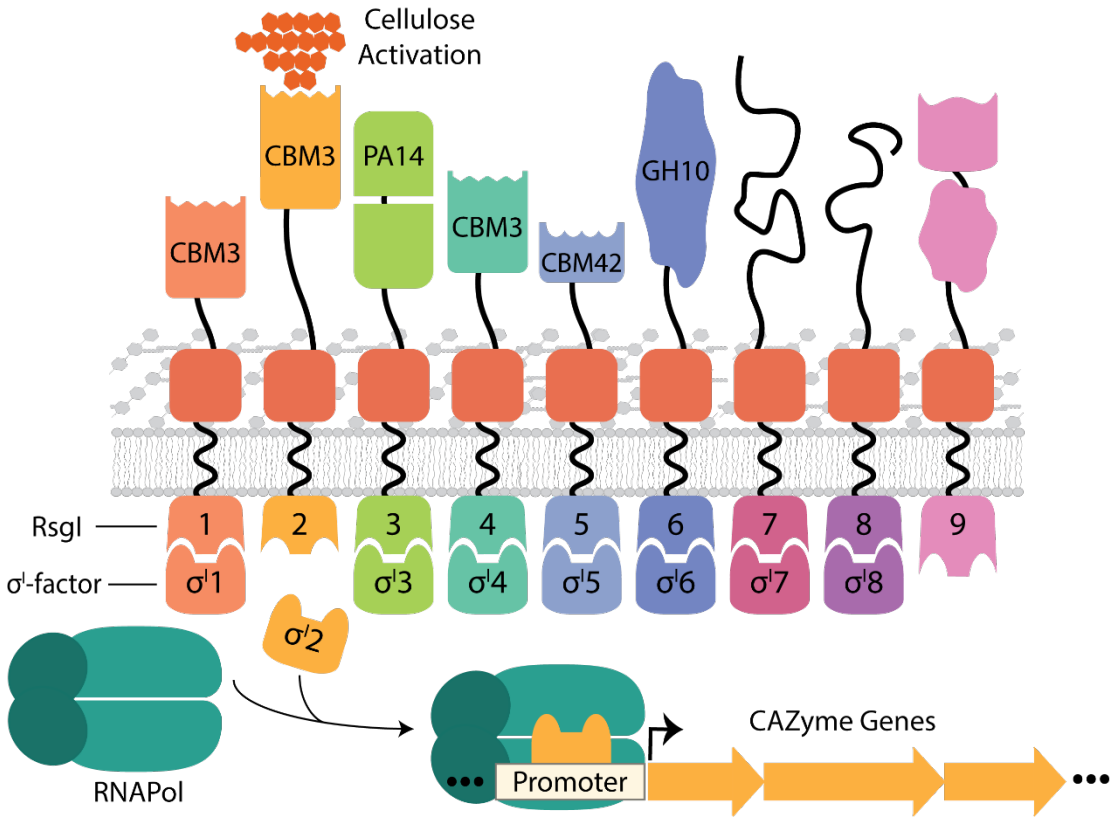


Figure 1.5 Set of nine biomass-sensing transmembrane RsgI proteins in *C. thermocellum*

RsgI proteins all contain both a conserved cytoplasmic N-terminal region responsible for binding its cognate σ^I-factor and extracellular domain that lies above the cell membrane. RsgI contain additional extracellular domains which are responsible for substrate sensing and are often associated with carbohydrate-binding such as CBM3 (cellulose), PA14 (pectin), and CBM42 (arabinoxylan). After signal activation and transduction through the RsgI, the σ factor is released so that it may guide the RNAPol holoenzyme to transcribe cellulolytic genes. Neither the cognate σ^I nor substrate-binding specificity of RsgI9 is known.

1.8 References

1. Robertson, G. P.; Dale, V. H.; Doering, O. C.; Hamburg, S. P.; Melillo, J. M.; Wander, M. M.; Parton, W. J.; Adler, P. R.; Barney, J. N.; Cruse, R. M.; Duke, C. S.; Fearnside, P. M.; Follett, R. F.; Gibbs, H. K.; Goldemberg, J.; Mladenoff, D. J.; Ojima, D.; Palmer, M. W.; Sharpley, A.; Wallace, L.; Weathers, K. C.; Wiens, J. A.; Wilhelm, W. W. Agriculture. Sustainable Biofuels Redux. Science 2008, 322 (5898), 49–50. https://doi.org/10.1126/science.1161525.
2. Lark, T. J.; Hendricks, N. P.; Smith, A.; Pates, N.; Spawn-Lee, S. A.; Bougie, M.; Booth, E. G.; Kucharik, C. J.; Gibbs, H. K. Environmental Outcomes of the US Renewable Fuel Standard. Proc. Natl. Acad. Sci. U. S. A. 2022, 119 (9), e2101084119. https://doi.org/10.1073/pnas.2101084119.
3. Mukherjee, P. K.; Das, B.; Bhardwaj, P. K.; Tampha, S.; Singh, H. K.; Chanu, L. D.; Sharma, N.; Devi, S. I. Socio-Economic Sustainability with Circular Economy - An Alternative Approach. Sci. Total Environ. 2023, 904, 166630. https://doi.org/10.1016/j.scitotenv.2023.166630.
4. Bar-On, Y. M.; Phillips, R.; Milo, R. The Biomass Distribution on Earth. Proc. Natl. Acad. Sci. U. S. A. 2018, 115 (25), 6506–6511. https://doi.org/10.1073/pnas.1711842115.
5. Shen, X.; Sun, R. Recent Advances in Lignocellulose Prior-Fractionation for Biomaterials, Biochemicals, and Bioenergy. Carbohydr. Polym. 2021, 261, 117884. https://doi.org/10.1016/j.carbpol.2021.117884.
6. Liguori, R.; Faraco, V. Biological Processes for Advancing Lignocellulosic Waste Biorefinery by Advocating Circular Economy. Bioresour. Technol. 2016, 215, 13–20. https://doi.org/10.1016/j.biortech.2016.04.054.

7. He, H.; Zhang, R.; Zhang, P.; Wang, P.; Chen, N.; Qian, B.; Zhang, L.; Yu, J.; Dai, B. Functional Carbon from Nature: Biomass-Derived Carbon Materials and the Recent Progress of Their Applications. Adv. Sci. Weinh. Baden-Wurtt. Ger. 2023, 10 (16), e2205557. https://doi.org/10.1002/advs.202205557.

8. Peng, X.; Nie, S.; Li, X.; Huang, X.; Li, Q. Characteristics of the Water- and Alkali-Soluble Hemicelluloses Fractionated by Sequential Acidification and Graded-Ethanol from Sweet Maize Stems. Mol. Basel Switz. 2019, 24 (1), 212. https://doi.org/10.3390/molecules24010212.

9. Zheng, B.; Yu, S.; Chen, Z.; Huo, Y.-X. A Consolidated Review of Commercial-Scale High-Value Products from Lignocellulosic Biomass. Front. Microbiol. 2022, 13, 933882. https://doi.org/10.3389/fmicb.2022.933882.

10. Wu, Z.; Peng, K.; Zhang, Y.; Wang, M.; Yong, C.; Chen, L.; Qu, P.; Huang, H.; Sun, E.; Pan, M. Lignocellulose Dissociation with Biological Pretreatment towards the Biochemical Platform: A Review. Mater. Today Bio 2022, 16, 100445. https://doi.org/10.1016/j.mtbio.2022.100445.

11. Chen, W.-H.; Pen, B.-L.; Yu, C.-T.; Hwang, W.-S. Pretreatment Efficiency and Structural Characterization of Rice Straw by an Integrated Process of Dilute-Acid and Steam Explosion for Bioethanol Production. Bioresour. Technol. 2011, 102 (3), 2916–2924. https://doi.org/10.1016/j.biortech.2010.11.052.

12. Haldar, D.; Purkait, M. K. A Review on the Environment-Friendly Emerging Techniques for Pretreatment of Lignocellulosic Biomass: Mechanistic Insight and Advancements. Chemosphere 2021, 264 (Pt 2), 128523. https://doi.org/10.1016/j.chemosphere.2020.128523.

13. Dahunsi, S. O. Mechanical Pretreatment of Lignocelluloses for Enhanced Biogas Production: Methane Yield Prediction from Biomass Structural Components. Bioresour. Technol. 2019, 280, 18–26. https://doi.org/10.1016/j.biortech.2019.02.006.

14. Chen, S.; Davaritouchaee, M. Nature-Inspired Pretreatment of Lignocellulose - Perspective and Development. Bioresour. Technol. 2023, 369, 128456. https://doi.org/10.1016/j.biortech.2022.128456.

15. Sutaoney, P.; Rai, S. N.; Sinha, S.; Choudhary, R.; Gupta, A. K.; Singh, S. K.; Banerjee, P. Current Perspective in Research and Industrial Applications of Microbial Cellulases. Int. J. Biol. Macromol. 2024, 264 (Pt 1), 130639. https://doi.org/10.1016/j.ijbiomac.2024.130639.

16. Zhao, W.; Zhao, F.; Zhang, S.; Gong, Q.; Chen, G. Ethanol Production by Simultaneous Saccharification and Cofermentation of Pretreated Corn Stalk. J. Basic Microbiol. 2019, 59 (7), 744–753. https://doi.org/10.1002/jobm.201900117.

17. Kumar, R.; Wyman, C. E. Effect of Additives on the Digestibility of Corn Stover Solids Following Pretreatment by Leading Technologies. Biotechnol. Bioeng. 2009, 102 (6), 1544–1557. https://doi.org/10.1002/bit.22203.

18. Millati, R.; Wikandari, R.; Ariyanto, T.; Putri, R. U.; Taherzadeh, M. J. Pretreatment Technologies for Anaerobic Digestion of Lignocelluloses and Toxic Feedstocks. Bioresour. Technol. 2020, 304, 122998. https://doi.org/10.1016/j.biortech.2020.122998.

19. Tan, J.; Li, Y.; Tan, X.; Wu, H.; Li, H.; Yang, S. Advances in Pretreatment of Straw Biomass for Sugar Production. Front. Chem. 2021, 9, 696030. https://doi.org/10.3389/fchem.2021.696030.

20. Li, C.; Huang, X.; Xi, J. Steam Explosion Pretreatment to Enhance Extraction of Active Ingredients: Current Progress and Future Prospects. Crit. Rev. Food Sci. Nutr. 2024, 64 (20), 7172–7180. https://doi.org/10.1080/10408398.2023.2181760.

21. Liu, G.; Zhang, J.; Bao, J. Cost Evaluation of Cellulase Enzyme for Industrial-Scale Cellulosic Ethanol Production Based on Rigorous Aspen Plus Modeling. Bioprocess Biosyst. Eng. 2016, 39 (1), 133–140. https://doi.org/10.1007/s00449-015-1497-1.

22. Agrawal, D.; Tsang, A.; Chadha, B. S. Economizing the Lignocellulosic Hydrolysis Process Using Heterologously Expressed Auxiliary Enzymes Feruloyl Esterase D (CE1) and β-Xylosidase (GH43) Derived from Thermophilic Fungi Scytalidium Thermophilum. Bioresour. Technol. 2021, 339, 125603. https://doi.org/10.1016/j.biortech.2021.125603.

23. Qiu, Y.; Wu, M.; Bao, H.; Liu, W.; Shen, Y. Engineering of Saccharomyces Cerevisiae for Co-Fermentation of Glucose and Xylose: Current State and Perspectives. Eng. Microbiol. 2023, 3 (3), 100084. https://doi.org/10.1016/j.engmic.2023.100084.

24. Minnaar, L. S.; Kruger, F.; Fortuin, J.; Hoffmeester, L. J.; den Haan, R. Engineering Saccharomyces Cerevisiae for Application in Integrated Bioprocessing Biorefineries. Curr. Opin. Biotechnol. 2024, 85, 103030. https://doi.org/10.1016/j.copbio.2023.103030.

25. Bhardwaj, N.; Kumar, B.; Agrawal, K.; Verma, P. Current Perspective on Production and Applications of Microbial Cellulases: A Review. Bioresour. Bioprocess. 2021, 8 (1), 95. https://doi.org/10.1186/s40643-021-00447-6.

26. Re, A.; Mazzoli, R. Current Progress on Engineering Microbial Strains and Consortia for Production of Cellulosic Butanol through Consolidated Bioprocessing. Microb. Biotechnol. 2023, 16 (2), 238–261. https://doi.org/10.1111/1751-7915.14148.

27. Li, Z.; Waghmare, P. R.; Dijkhuizen, L.; Meng, X.; Liu, W. Research Advances on the Consolidated Bioprocessing of Lignocellulosic Biomass. Eng. Microbiol. 2024, 4 (2), 100139. https://doi.org/10.1016/j.engmic.2024.100139.

28. Bissaro, B.; Várnai, A.; Røhr, Å. K.; Eijsink, V. G. H. Oxidoreductases and Reactive Oxygen Species in Conversion of Lignocellulosic Biomass. Microbiol. Mol. Biol. Rev. MMBR 2018, 82 (4), e00029-18. https://doi.org/10.1128/MMBR.00029-18.

29. Singhania, R. R.; Patel, A. K.; Singh, A.; Haldar, D.; Soam, S.; Chen, C.-W.; Tsai, M.-L.; Dong, C.-D. Consolidated Bioprocessing of Lignocellulosic Biomass:

Technological Advances and Challenges. Bioresour. Technol. 2022, 354, 127153. https://doi.org/10.1016/j.biortech.2022.127153.

30. Cragg, S. M.; Beckham, G. T.; Bruce, N. C.; Bugg, T. D. H.; Distel, D. L.; Dupree, P.; Etxabe, A. G.; Goodell, B. S.; Jellison, J.; McGeehan, J. E.; McQueen-Mason, S. J.; Schnorr, K.; Walton, P. H.; Watts, J. E. M.; Zimmer, M. Lignocellulose Degradation Mechanisms across the Tree of Life. Curr. Opin. Chem. Biol. 2015, 29, 108–119. https://doi.org/10.1016/j.cbpa.2015.10.018.

31. Zuroff, T. R.; Xiques, S. B.; Curtis, W. R. Consortia-Mediated Bioprocessing of Cellulose to Ethanol with a Symbiotic Clostridium Phytofermentans/Yeast Co-Culture. Biotechnol. Biofuels 2013, 6 (1), 59. https://doi.org/10.1186/1754-6834-6-59.

32. Gao, L.; Yu, Z.; Wang, S.; Hou, Y.; Zhang, S.; Zhou, C.; Wu, X. A New Paradigm in Lignocellulolytic Enzyme Cocktail Optimization: Free from Expert-Level Prior Knowledge and Experimental Datasets. Bioresour. Technol. 2023, 388, 129758. https://doi.org/10.1016/j.biortech.2023.129758.

33. Huang, X.-Y.; Ao, T.-J.; Zhang, X.; Li, K.; Zhao, X.-Q.; Champreda, V.; Runguphan, W.; Sakdaronnarong, C.; Liu, C.-G.; Bai, F.-W. Developing High-Dimensional Machine Learning Models to Improve Generalization Ability and Overcome Data Insufficiency for Mixed Sugar Fermentation Simulation. Bioresour. Technol. 2023, 385, 129375. https://doi.org/10.1016/j.biortech.2023.129375.

34. Gusakov, A. V. Alternatives to Trichoderma Reesei in Biofuel Production. Trends Biotechnol. 2011, 29 (9), 419–425. https://doi.org/10.1016/j.tibtech.2011.04.004.

35. Gutiérrez-Rojas, I.; Moreno-Sarmiento, N.; Montoya, D. [Mechanisms and regulation of enzymatic hydrolysis of cellulose in filamentous fungi: classical cases and new models]. Rev. Iberoam. Micol. 2015, 32 (1), 1–12. https://doi.org/10.1016/j.riam.2013.10.009.

36. Schuster, A.; Schmoll, M. Biology and Biotechnology of Trichoderma. Appl. Microbiol. Biotechnol. 2010, 87 (3), 787–799. https://doi.org/10.1007/s00253-010-2632-1.

37. Berka, R. M.; Grigoriev, I. V.; Otillar, R.; Salamov, A.; Grimwood, J.; Reid, I.; Ishmael, N.; John, T.; Darmond, C.; Moisan, M.-C.; Henrissat, B.; Coutinho, P. M.; Lombard, V.; Natvig, D. O.; Lindquist, E.; Schmutz, J.; Lucas, S.; Harris, P.; Powlowski, J.; Bellemare, A.; Taylor, D.; Butler, G.; de Vries, R. P.; Allijn, I. E.; van den Brink, J.; Ushinsky, S.; Storms, R.; Powell, A. J.; Paulsen, I. T.; Elbourne, L. D. H.; Baker, S. E.; Magnuson, J.; Laboissiere, S.; Clutterbuck, A. J.; Martinez, D.; Wogulis, M.; de Leon, A. L.; Rey, M. W.; Tsang, A. Comparative Genomic Analysis of the Thermophilic Biomass-Degrading Fungi Myceliophthora Thermophila and Thielavia Terrestris. Nat. Biotechnol. 2011, 29 (10), 922–927. https://doi.org/10.1038/nbt.1976.

38. Rautio, J. J.; Smit, B. A.; Wiebe, M.; Penttilä, M.; Saloheimo, M. Transcriptional Monitoring of Steady State and Effects of Anaerobic Phases in Chemostat Cultures of the Filamentous Fungus Trichoderma Reesei. BMC Genomics 2006, 7, 247. https://doi.org/10.1186/1471-2164-7-247.

39. Le, Y.; Zhang, M.; Wu, P.; Wang, H.; Ni, J. Biofuel Production from Lignocellulose via Thermophile-Based Consolidated Bioprocessing. Eng. Microbiol. 2024, 4 (4), 100174. https://doi.org/10.1016/j.engmic.2024.100174.

40. Fontes, C. M. G. A.; Gilbert, H. J. Cellulosomes: Highly Efficient Nanomachines Designed to Deconstruct Plant Cell Wall Complex Carbohydrates. Annu. Rev. Biochem. 2010, 79, 655–681. https://doi.org/10.1146/annurev-biochem-091208-085603.

41. Gruber, T. M.; Gross, C. A. Multiple Sigma Subunits and the Partitioning of Bacterial Transcription Space. Annu. Rev. Microbiol. 2003, 57, 441–466. https://doi.org/10.1146/annurev.micro.57.030502.090913.

42. Browning, D. F.; Busby, S. J. The Regulation of Bacterial Transcription Initiation. Nat. Rev. Microbiol. 2004, 2 (1), 57–65. https://doi.org/10.1038/nrmicro787.

43. Gross, C. A.; Chan, C.; Dombroski, A.; Gruber, T.; Sharp, M.; Tupy, J.; Young, B. The Functional and Regulatory Roles of Sigma Factors in Transcription. Cold Spring Harb. Symp. Quant. Biol. 1998, 63, 141–155. https://doi.org/10.1101/sqb.1998.63.141.

44. Sachdeva, P.; Misra, R.; Tyagi, A. K.; Singh, Y. The Sigma Factors of Mycobacterium Tuberculosis: Regulation of the Regulators. FEBS J. 2010, 277 (3), 605–626. https://doi.org/10.1111/j.1742-4658.2009.07479.x.

45. Barnard, A.; Wolfe, A.; Busby, S. Regulation at Complex Bacterial Promoters: How Bacteria Use Different Promoter Organizations to Produce Different Regulatory Outcomes. Curr. Opin. Microbiol. 2004, 7 (2), 102–108. https://doi.org/10.1016/j.mib.2004.02.011.

46. Kazmierczak, M. J.; Wiedmann, M.; Boor, K. J. Alternative Sigma Factors and Their Roles in Bacterial Virulence. Microbiol. Mol. Biol. Rev. MMBR 2005, 69 (4), 527–543. https://doi.org/10.1128/MMBR.69.4.527-543.2005.

47. Kazmierczak, M. J.; Mithoe, S. C.; Boor, K. J.; Wiedmann, M. Listeria Monocytogenes Sigma B Regulates Stress Response and Virulence Functions. J. Bacteriol. 2003, 185 (19), 5722–5734. https://doi.org/10.1128/JB.185.19.5722-5734.2003.

48. El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S. R.; Luciani, A.; Potter, S. C.; Qureshi, M.; Richardson, L. J.; Salazar, G. A.; Smart, A.; Sonnhammer, E L. L.; Hirsh, L.; Paladin, L.; Piovesan, D.; Tosatto, S. C. E.; Finn, R. D. The Pfam Protein Families Database in 2019. Nucleic Acids Res. 2019, 47 (D1), D427–D432. https://doi.org/10.1093/nar/gky995.

49. Helmann, J. D.; Chamberlin, M. J. Structure and Function of Bacterial Sigma Factors. Annu. Rev. Biochem. 1988, 57, 839–872. https://doi.org/10.1146/annurev.bi.57.070188.004203.

50. Souza, B. M.; Castro, T. L. de P.; Carvalho, R. D. de O.; Seyffert, N.; Silva, A.; Miyoshi, A.; Azevedo, V. σ(ECF) Factors of Gram-Positive Bacteria: A Focus on Bacillus Subtilis and the CMNR Group. Virulence 2014, 5 (5), 587–600. https://doi.org/10.4161/viru.29514.

51. Paget, M. S. Bacterial Sigma Factors and Anti-Sigma Factors: Structure, Function and Distribution. Biomolecules 2015, 5 (3), 1245–1265. https://doi.org/10.3390/biom5031245.

52. Pátek, M.; Nešvera, J. Sigma Factors and Promoters in Corynebacterium Glutamicum. J. Biotechnol. 2011, 154 (2–3), 101–113. https://doi.org/10.1016/j.jbiotec.2011.01.017.

53. Cartagena, A. J.; Banta, A. B.; Sathyan, N.; Ross, W.; Gourse, R. L.; Campbell, E. A.; Darst, S. A. Structural Basis for Transcription Activation by Crl through Tethering of σS and RNA Polymerase. Proc. Natl. Acad. Sci. U. S. A. 2019, 116 (38), 18923–18927. https://doi.org/10.1073/pnas.1910827116.

54. Liu, B.; Zuo, Y.; Steitz, T. A. Structures of E. Coli σS-Transcription Initiation Complexes Provide New Insights into Polymerase Mechanism. Proc. Natl. Acad. Sci. U. S. A. 2016, 113 (15), 4051–4056. https://doi.org/10.1073/pnas.1520555113.

55. Shi, W.; Zhou, W.; Zhang, B.; Huang, S.; Jiang, Y.; Schammel, A.; Hu, Y.; Liu, B. Structural Basis of Bacterial Σ28 -Mediated Transcription Reveals Roles of the RNA Polymerase Zinc-Binding Domain. EMBO J. 2020, 39 (14), e104389. https://doi.org/10.15252/embj.2020104389.

56. Lu, Q.; Chen, T.; Wang, J.; Wang, F.; Ye, W.; Ma, L.; Wu, S. Structural Insight into the Mechanism of Σ32-Mediated Transcription Initiation of Bacterial RNA Polymerase. Biomolecules 2023, 13 (5), 738. https://doi.org/10.3390/biom13050738.

57. Staroń, A.; Sofia, H. J.; Dietrich, S.; Ulrich, L. E.; Liesegang, H.; Mascher, T. The Third Pillar of Bacterial Signal Transduction: Classification of the Extracytoplasmic Function (ECF) Sigma Factor Protein Family. Mol. Microbiol. 2009, 74 (3), 557–581. https://doi.org/10.1111/j.1365-2958.2009.06870.x.

58. Lonetto, M. A.; Brown, K. L.; Rudd, K. E.; Buttner, M. J. Analysis of the Streptomyces Coelicolor sigE Gene Reveals the Existence of a Subfamily of Eubacterial RNA Polymerase Sigma Factors Involved in the Regulation of Extracytoplasmic Functions. Proc. Natl. Acad. Sci. U. S. A. 1994, 91 (16), 7573–7577. https://doi.org/10.1073/pnas.91.16.7573.

59. Pinto, D.; Liu, Q.; Mascher, T. ECF σ Factors with Regulatory Extensions: The One-Component Systems of the σ Universe. Mol. Microbiol. 2019, 112 (2), 399–409. https://doi.org/10.1111/mmi.14323.

60. Mascher, T. Past, Present, and Future of Extracytoplasmic Function σ Factors: Distribution and Regulatory Diversity of the Third Pillar of Bacterial Signal Transduction. Annu. Rev. Microbiol. 2023, 77, 625–644. https://doi.org/10.1146/annurev-micro-032221-024032.

61. Helmann, J. D. The Extracytoplasmic Function (ECF) Sigma Factors. Adv. Microb. Physiol. 2002, 46, 47–110. https://doi.org/10.1016/s0065-2911(02)46002-x.

62. Lane, W. J.; Darst, S. A. The Structural Basis for Promoter -35 Element Recognition by the Group IV Sigma Factors. PLoS Biol. 2006, 4 (9), e269. https://doi.org/10.1371/journal.pbio.0040269.

63. Campagne, S.; Marsh, M. E.; Capitani, G.; Vorholt, J. A.; Allain, F. H.-T. Structural Basis for -10 Promoter Element Melting by Environmentally Induced Sigma Factors. Nat. Struct. Mol. Biol. 2014, 21 (3), 269–276. https://doi.org/10.1038/nsmb.2777.

64. Fang, C.; Li, L.; Shen, L.; Shi, J.; Wang, S.; Feng, Y.; Zhang, Y. Structures and Mechanism of Transcription Initiation by Bacterial ECF Factors. Nucleic Acids Res. 2019, 47 (13), 7094–7104. https://doi.org/10.1093/nar/gkz470.

65. Li, L.; Fang, C.; Zhuang, N.; Wang, T.; Zhang, Y. Structural Basis for Transcription Initiation by Bacterial ECF σ Factors. Nat. Commun. 2019, 10 (1), 1153. https://doi.org/10.1038/s41467-019-09096-y.

66. Lin, W.; Mandal, S.; Degen, D.; Cho, M. S.; Feng, Y.; Das, K.; Ebright, R. H. Structural Basis of ECF-σ factor-Dependent Transcription Initiation. Nat. Commun. 2019, 10 (1), 710. https://doi.org/10.1038/s41467-019-08443-3.

67. Campagne, S.; Allain, F. H.-T.; Vorholt, J. A. Extra Cytoplasmic Function Sigma Factors, Recent Structural Insights into Promoter Recognition and Regulation. Curr. Opin. Struct. Biol. 2015, 30, 71–78. https://doi.org/10.1016/j.sbi.2015.01.006.

68. Rhodius, V. A.; Segall-Shapiro, T. H.; Sharon, B. D.; Ghodasara, A.; Orlova, E.; Tabakh, H.; Burkhardt, D. H.; Clancy, K.; Peterson, T. C.; Gross, C. A.; Voigt, C. A. Design of Orthogonal Genetic Switches Based on a Crosstalk Map of Σs, Anti-Σs, and Promoters. Mol. Syst. Biol. 2013, 9, 702. https://doi.org/10.1038/msb.2013.58.

69. Liu, G.; Qu, Y. Integrated Engineering of Enzymes and Microorganisms for Improving the Efficiency of Industrial Lignocellulose Deconstruction. Eng. Microbiol. 2021, 1, 100005. https://doi.org/10.1016/j.engmic.2021.100005.

70. Gurovic, M. S. V.; Viceconte, F. R.; Bidegain, M. A.; Dietrich, J. Regulation of Lignocellulose Degradation in Microorganisms. J. Appl. Microbiol. 2023, 134 (1), lxac002. https://doi.org/10.1093/jambio/lxac002.

71. Alvarez, A. F.; Georgellis, D. Environmental Adaptation and Diversification of Bacterial Two-Component Systems. Curr. Opin. Microbiol. 2023, 76, 102399. https://doi.org/10.1016/j.mib.2023.102399.

72. Stock, A. M.; Robinson, V. L.; Goudreau, P. N. Two-Component Signal Transduction. Annu. Rev. Biochem. 2000, 69, 183–215. https://doi.org/10.1146/annurev.biochem.69.1.183.

73. Buschiazzo, A.; Trajtenberg, F. Two-Component Sensing and Regulation: How Do Histidine Kinases Talk with Response Regulators at the Molecular Level? Annu. Rev. Microbiol. 2019, 73, 507–528. https://doi.org/10.1146/annurev-micro-091018-054627.

74. Francis, V. I.; Porter, S. L. Multikinase Networks: Two-Component Signaling Networks Integrating Multiple Stimuli. Annu. Rev. Microbiol. 2019, 73, 199–223. https://doi.org/10.1146/annurev-micro-020518-115846.

75. Alm, E.; Huang, K.; Arkin, A. The Evolution of Two-Component Systems in Bacteria Reveals Different Strategies for Niche Adaptation. PLoS Comput. Biol. 2006, 2 (11), e143. https://doi.org/10.1371/journal.pcbi.0020143.

76. Alvarez, A. F.; Barba-Ostria, C.; Silva-Jiménez, H.; Georgellis, D. Organization and Mode of Action of Two Component System Signaling Circuits from the Various Kingdoms of Life. Environ. Microbiol. 2016, 18 (10), 3210–3226. https://doi.org/10.1111/1462-2920.13397.

77. Capra, E. J.; Perchuk, B. S.; Skerker, J. M.; Laub, M. T. Adaptive Mutations That Prevent Crosstalk Enable the Expansion of Paralogous Signaling Protein Families. Cell 2012, 150 (1), 222–232. https://doi.org/10.1016/j.cell.2012.05.033.

78. Laub, M. T.; Goulian, M. Specificity in Two-Component Signal Transduction Pathways. Annu. Rev. Genet. 2007, 41, 121–145. https://doi.org/10.1146/annurev.genet.41.042007.170548.

79. Groban, E. S.; Clarke, E. J.; Salis, H. M.; Miller, S. M.; Voigt, C. A. Kinetic Buffering of Cross Talk between Bacterial Two-Component Sensors. J. Mol. Biol. 2009, 390 (3), 380–393. https://doi.org/10.1016/j.jmb.2009.05.007.

80. Emami, K.; Topakas, E.; Nagy, T.; Henshaw, J.; Jackson, K. A.; Nelson, K. E.; Mongodin, E. F.; Murray, J. W.; Lewis, R. J.; Gilbert, H. J. Regulation of the Xylan-Degrading Apparatus of Cellvibrio Japonicus by a Novel Two-Component System. J. Biol. Chem. 2009, 284 (2), 1086–1096. https://doi.org/10.1074/jbc.M805100200.

81. Martens, E. C.; Lowe, E. C.; Chiang, H.; Pudlo, N. A.; Wu, M.; McNulty, N. P.; Abbott, D. W.; Henrissat, B.; Gilbert, H. J.; Bolam, D. N.; Gordon, J. I. Recognition and Degradation of Plant Cell Wall Polysaccharides by Two Human Gut Symbionts. PLoS Biol. 2011, 9 (12), e1001221. https://doi.org/10.1371/journal.pbio.1001221.

82. Celik, H.; Blouzard, J.-C.; Voigt, B.; Becher, D.; Trotter, V.; Fierobe, H.-P.; Tardif, C.; Pagès, S.; de Philip, P. A Two-Component System (XydS/R) Controls the Expression of Genes Encoding CBM6-Containing Proteins in Response to Straw in Clostridium Cellulolyticum. PloS One 2013, 8 (2), e56063. https://doi.org/10.1371/journal.pone.0056063.

83. Lansky, S.; Salama, R.; Shulami, S.; Lavid, N.; Sen, S.; Schapiro, I.; Shoham, Y.; Shoham, G. Carbohydrate-Binding Capability and Functional Conformational Changes of AbnE, an Arabino-Oligosaccharide Binding Protein. J. Mol. Biol. 2020, 432 (7), 2099–2120. https://doi.org/10.1016/j.jmb.2020.01.041.

84. Görke, B.; Stülke, J. Carbon Catabolite Repression in Bacteria: Many Ways to Make the Most out of Nutrients. Nat. Rev. Microbiol. 2008, 6 (8), 613–624. https://doi.org/10.1038/nrmicro1932.

85. Novak, J. K.; Gardner, J. G. Current Models in Bacterial Hemicellulase-Encoding Gene Regulation. Appl. Microbiol. Biotechnol. 2024, 108 (1), 39. https://doi.org/10.1007/s00253-023-12977-4.

86. Feucht, B. U.; Saier, M. H. Fine Control of Adenylate Cyclase by the Phosphoenolpyruvate:Sugar Phosphotransferase Systems in Escherichia Coli and Salmonella Typhimurium. J. Bacteriol. 1980, 141 (2), 603–610. https://doi.org/10.1128/jb.141.2.603-610.1980.

87. Deutscher, J.; Küster, E.; Bergstedt, U.; Charrier, V.; Hillen, W. Protein Kinase-Dependent HPr/CcpA Interaction Links Glycolytic Activity to Carbon Catabolite Repression in Gram-Positive Bacteria. Mol. Microbiol. 1995, 15 (6), 1049–1053. https://doi.org/10.1111/j.1365-2958.1995.tb02280.x.

88. Deutscher, J.; Saier, M. H. ATP-Dependent Protein Kinase-Catalyzed Phosphorylation of a Seryl Residue in HPr, a Phosphate Carrier Protein of the Phosphotransferase System in Streptococcus Pyogenes. Proc. Natl. Acad. Sci. U. S. A. 1983, 80 (22), 6790–6794. https://doi.org/10.1073/pnas.80.22.6790.

89. Fujita, Y. Carbon Catabolite Control of the Metabolic Network in Bacillus Subtilis. Biosci. Biotechnol. Biochem. 2009, 73 (2), 245–259. https://doi.org/10.1271/bbb.80479.

90. Fujita, Y.; Miwa, Y.; Galinier, A.; Deutscher, J. Specific Recognition of the Bacillus Subtilis Gnt Cis-Acting Catabolite-Responsive Element by a Protein Complex Formed between CcpA and Seryl-Phosphorylated HPr. Mol. Microbiol. 1995, 17 (5), 953–960. https://doi.org/10.1111/j.1365-2958.1995.mmi_17050953.x.

91. Ha, S.-J.; Galazka, J. M.; Kim, S. R.; Choi, J.-H.; Yang, X.; Seo, J.-H.; Glass, N. L.; Cate, J. H. D.; Jin, Y.-S. Engineered Saccharomyces Cerevisiae Capable of Simultaneous Cellobiose and Xylose Fermentation. Proc. Natl. Acad. Sci. U. S. A. 2011, 108 (2), 504–509. https://doi.org/10.1073/pnas.1010456108.

92. Vinuselvi, P.; Lee, S. K. Engineered Escherichia Coli Capable of Co-Utilization of Cellobiose and Xylose. Enzyme Microb. Technol. 2012, 50 (1), 1–4. https://doi.org/10.1016/j.enzmictec.2011.10.001.

93. Liu, Z.; Ho, S.-H.; Sasaki, K.; den Haan, R.; Inokuma, K.; Ogino, C.; van Zyl, W. H.; Hasunuma, T.; Kondo, A. Engineering of a Novel Cellulose-Adherent Cellulolytic Saccharomyces Cerevisiae for Cellulosic Biofuel Production. Sci. Rep. 2016, 6, 24550. https://doi.org/10.1038/srep24550.

94. Lamote, B.; da Fonseca, M. J. M.; Vanderstraeten, J.; Meert, K.; Elias, M.; Briers, Y. Current Challenges in Designer Cellulosome Engineering. Appl. Microbiol. Biotechnol. 2023, 107 (9), 2755–2770. https://doi.org/10.1007/s00253-023-12474-8.

95. Xiong, W.; Reyes, L. H.; Michener, W. E.; Maness, P.-C.; Chou, K. J. Engineering Cellulolytic Bacterium Clostridium Thermocellum to Co-Ferment Cellulose- and Hemicellulose-Derived Sugars Simultaneously. Biotechnol. Bioeng. 2018, 115 (7), 1755–1763. https://doi.org/10.1002/bit.26590.

96. Rochat, T.; Bouloc, P.; Repoila, F. Gene Expression Control by Selective RNA Processing and Stabilization in Bacteria. FEMS Microbiol. Lett. 2013, 344 (2), 104–113. https://doi.org/10.1111/1574-6968.12162.

97. Xu, C.; Huang, R.; Teng, L.; Jing, X.; Hu, J.; Cui, G.; Wang, Y.; Cui, Q.; Xu, J. Cellulosome Stoichiometry in Clostridium Cellulolyticum Is Regulated by Selective RNA Processing and Stabilization. Nat. Commun. 2015, 6, 6900. https://doi.org/10.1038/ncomms7900.

98. Bhaskar, Y.; Su, X.; Xu, C.; Xu, J. Predicting Selective RNA Processing and Stabilization Operons in Clostridium Spp. Front. Microbiol. 2021, 12, 673349. https://doi.org/10.3389/fmicb.2021.673349.

99. Ortiz de Ora, L.; Muñoz-Gutiérrez, I.; Bayer, E. A.; Shoham, Y.; Lamed, R.; Borovok, I. Revisiting the Regulation of the Primary Scaffoldin Gene in Clostridium Thermocellum. Appl. Environ. Microbiol. 2017, 83 (8), e03088-16. https://doi.org/10.1128/AEM.03088-16.

100. Zou, X.; Ren, Z.; Wang, N.; Cheng, Y.; Jiang, Y.; Wang, Y.; Xu, C. Function Analysis of 5'-UTR of the Cellulosomal Xyl-Doc Cluster in Clostridium Papyrosolvens. Biotechnol. Biofuels 2018, 11, 43. https://doi.org/10.1186/s13068-018-1040-0.

101. Grondin, J. M.; Tamura, K.; Déjean, G.; Abbott, D. W.; Brumer, H. Polysaccharide Utilization Loci: Fueling Microbial Communities. J. Bacteriol. 2017, 199 (15), e00860-16. https://doi.org/10.1128/JB.00860-16.

102. Georg, J.; Hess, W. R. Cis-Antisense RNA, Another Level of Gene Regulation in Bacteria. Microbiol. Mol. Biol. Rev. MMBR 2011, 75 (2), 286–300. https://doi.org/10.1128/MMBR.00032-10.

103. Pudlo, N. A.; Urs, K.; Kumar, S. S.; German, J. B.; Mills, D. A.; Martens, E. C. Symbiotic Human Gut Bacteria with Variable Metabolic Priorities for Host Mucosal Glycans. mBio 2015, 6 (6), e01282-01215. https://doi.org/10.1128/mBio.01282-15.

104. Marcos-Torres, F. J.; Pérez, J.; Gómez-Santos, N.; Moraleda-Muñoz, A.; Muñoz-Dorado, J. In Depth Analysis of the Mechanism of Action of Metal-Dependent Sigma Factors: Characterization of CorE2 from Myxococcus Xanthus. Nucleic Acids Res. 2016, 44 (12), 5571–5584. https://doi.org/10.1093/nar/gkw150.

105. Goutam, K.; Gupta, A. K.; Gopal, B. The Fused SnoaL_2 Domain in the Mycobacterium Tuberculosis Sigma Factor σJ Modulates Promoter Recognition. Nucleic Acids Res. 2017, 45 (16), 9760–9772. https://doi.org/10.1093/nar/gkx609.

106. Asai, K.; Ootsuji, T.; Obata, K.; Matsumoto, T.; Fujita, Y.; Sadaie, Y. Regulatory Role of RsgI in sigI Expression in Bacillus Subtilis. Microbiol. Read. Engl. 2007, 153 (Pt 1), 92–101. https://doi.org/10.1099/mic.0.29239-0.

107. Gupta, N.; Gupta, A.; Kumar, S.; Mishra, R.; Singh, C.; Tripathi, A. K. Cross-Talk between Cognate and Noncognate RpoE Sigma Factors and Zn(2+)-Binding Anti-Sigma Factors Regulates Photooxidative Stress Response in Azospirillum

Brasilense. Antioxid. Redox Signal. 2014, 20 (1), 42–59. https://doi.org/10.1089/ars.2013.5314.

108. Edgar, R. J.; Xu, X.; Shirley, M.; Konings, A. F.; Martin, L. W.; Ackerley, D. F.; Lamont, I. L. Interactions between an Anti-Sigma Protein and Two Sigma Factors That Regulate the Pyoverdine Signaling Pathway in Pseudomonas Aeruginosa. BMC Microbiol. 2014, 14, 287. https://doi.org/10.1186/s12866-014-0287-2.

109. Casas-Pastor, D.; Müller, R. R.; Jaenicke, S.; Brinkrolf, K.; Becker, A.; Buttner, M. J.; Gross, C. A.; Mascher, T.; Goesmann, A.; Fritz, G. Expansion and Re-Classification of the Extracytoplasmic Function (ECF) σ Factor Family. Nucleic Acids Res. 2021, 49 (2), 986–1005. https://doi.org/10.1093/nar/gkaa1229.

110. Paget, M. S.; Molle, V.; Cohen, G.; Aharonowitz, Y.; Buttner, M. J. Defining the Disulphide Stress Response in Streptomyces Coelicolor A3(2): Identification of the sigmaR Regulon. Mol. Microbiol. 2001, 42 (4), 1007–1020. https://doi.org/10.1046/j.1365-2958.2001.02675.x.

111. Li, W.; Stevenson, C. E. M.; Burton, N.; Jakimowicz, P.; Paget, M. S. B.; Buttner, M. J.; Lawson, D. M.; Kleanthous, C. Identification and Structure of the Anti-Sigma Factor-Binding Domain of the Disulphide-Stress Regulated Sigma Factor Sigma(R) from Streptomyces Coelicolor. J. Mol. Biol. 2002, 323 (2), 225–236. https://doi.org/10.1016/s0022-2836(02)00948-8.

112. Nam, T.-W.; Ziegelhoffer, E. C.; Lemke, R. A. S.; Donohue, T. J. Proteins Needed to Activate a Transcriptional Response to the Reactive Oxygen Species Singlet Oxygen. mBio 2013, 4 (1), e00541-00512. https://doi.org/10.1128/mBio.00541-12.

113. Gómez-Santos, N.; Pérez, J.; Sánchez-Sutil, M. C.; Moraleda-Muñoz, A.; Muñoz-Dorado, J. CorE from Myxococcus Xanthus Is a Copper-Dependent RNA Polymerase Sigma Factor. PLoS Genet. 2011, 7 (6), e1002106. https://doi.org/10.1371/journal.pgen.1002106.

114. Wecke, T.; Halang, P.; Staroń, A.; Dufour, Y. S.; Donohue, T. J.; Mascher, T. Extracytoplasmic Function σ Factors of the Widely Distributed Group ECF41 Contain a Fused Regulatory Domain. MicrobiologyOpen 2012, 1 (2), 194–213. https://doi.org/10.1002/mbo3.22.

115. Liu, Q.; Pinto, D.; Mascher, T. Characterization of the Widely Distributed Novel ECF42 Group of Extracytoplasmic Function σ Factors in Streptomyces Venezuelae. J. Bacteriol. 2018, 200 (21), e00437-18. https://doi.org/10.1128/JB.00437-18.

116. Bayer-Santos, E.; Lima, L. D. P.; Ceseti, L. de M.; Ratagami, C. Y.; de Santana, E. S.; da Silva, A. M.; Farah, C. S.; Alvarez-Martinez, C. E. Xanthomonas Citri T6SS Mediates Resistance to Dictyostelium Predation and Is Regulated by an ECF σ Factor and Cognate Ser/Thr Kinase. Environ. Microbiol. 2018, 20 (4), 1562–1575. https://doi.org/10.1111/1462-2920.14085.

117. Kaczmarczyk, A.; Hochstrasser, R.; Vorholt, J. A.; Francez-Charlot, A. Complex Two-Component Signaling Regulates the General Stress Response in Alphaproteobacteria. Proc. Natl. Acad. Sci. U. S. A. 2014, 111 (48), E5196-5204. https://doi.org/10.1073/pnas.1410095111.

118. Fiebig, A.; Herrou, J.; Willett, J.; Crosson, S. General Stress Signaling in the Alphaproteobacteria. Annu. Rev. Genet. 2015, 49, 603–625. https://doi.org/10.1146/annurev-genet-112414-054813.

119. Barchinger, S. E.; Ades, S. E. Regulated Proteolysis: Control of the Escherichia Coli σ(E)-Dependent Cell Envelope Stress Response. Subcell. Biochem. 2013, 66, 129–160. https://doi.org/10.1007/978-94-007-5940-4_6.

120. Llamas, M. A.; Imperi, F.; Visca, P.; Lamont, I. L. Cell-Surface Signaling in Pseudomonas: Stress Responses, Iron Transport, and Pathogenicity. FEMS Microbiol. Rev. 2014, 38 (4), 569–597. https://doi.org/10.1111/1574-6976.12078.

121. Heinrich, J.; Wiegert, T. Regulated Intramembrane Proteolysis in the Control of Extracytoplasmic Function Sigma Factors. Res. Microbiol. 2009, 160 (9), 696–703. https://doi.org/10.1016/j.resmic.2009.08.019.

122. Hizukuri, Y.; Oda, T.; Tabata, S.; Tamura-Kawakami, K.; Oi, R.; Sato, M.; Takagi, J.; Akiyama, Y.; Nogi, T. A Structure-Based Model of Substrate Discrimination by a Noncanonical PDZ Tandem in the Intramembrane-Cleaving Protease RseP. Struct. Lond. Engl. 1993 2014, 22 (2), 326–336. https://doi.org/10.1016/j.str.2013.12.003.

123. Akiyama, K.; Mizuno, S.; Hizukuri, Y.; Mori, H.; Nogi, T.; Akiyama, Y. Roles of the Membrane-Reentrant β-Hairpin-like Loop of RseP Protease in Selective Substrate Cleavage. eLife 2015, 4, e08928. https://doi.org/10.7554/eLife.08928.

124. Wen, W.; Chen, C.; Cui, Q.; Xuan, J.; Feng, Y. The Final Proteolytic Step in Transmembrane Signaling of Multiple RsgI Anti-σ Factors in Clostridium Thermocellum. Biosci. Rep. 2025, 45 (4), 233–245. https://doi.org/10.1042/BSR20253055.

125. Nataf, Y.; Bahari, L.; Kahel-Raifer, H.; Borovok, I.; Lamed, R.; Bayer, E. A.; Sonenshein, A. L.; Shoham, Y. Clostridium Thermocellum Cellulosomal Genes Are Regulated by Extracytoplasmic Polysaccharides via Alternative Sigma Factors. Proc. Natl. Acad. Sci. U. S. A. 2010, 107 (43), 18646–18651. https://doi.org/10.1073/pnas.1012175107.

126. Kahel-Raifer, H.; Jindou, S.; Bahari, L.; Nataf, Y.; Shoham, Y.; Bayer, E. A.; Borovok, I.; Lamed, R. The Unique Set of Putative Membrane-Associated Anti-Sigma Factors in Clostridium Thermocellum Suggests a Novel Extracellular Carbohydrate-Sensing Mechanism Involved in Gene Regulation. FEMS Microbiol. Lett. 2010, 308 (1), 84–93. https://doi.org/10.1111/j.1574-6968.2010.01997.x.

127. Dong, S.; Chen, C.; Li, J.; Liu, Y.-J.; Bayer, E. A.; Lamed, R.; Mizrahi, I.; Cui, Q.; Feng, Y. Unique Fn3-like Biosensor in σI/Anti-σI Factors for Regulatory Expression of

Major Cellulosomal Scaffoldins in Pseudobacteroides Cellulosolvens. Protein Sci. Publ. Protein Soc. 2024, 33 (11), e5193. https://doi.org/10.1002/pro.5193.

128. Zuber, U.; Drzewiecki, K.; Hecker, M. Putative Sigma Factor SigI (YkoZ) of Bacillus Subtilis Is Induced by Heat Shock. J. Bacteriol. 2001, 183 (4), 1472–1475. https://doi.org/10.1128/JB.183.4.1472-1475.2001.

129. Yaniv, O.; Fichman, G.; Borovok, I.; Shoham, Y.; Bayer, E. A.; Lamed, R.; Shimon, L. J. W.; Frolow, F. Fine-Structural Variance of Family 3 Carbohydrate-Binding Modules as Extracellular Biomass-Sensing Components of Clostridium Thermocellum Anti-σI Factors. Acta Crystallogr. D Biol. Crystallogr. 2014, 70 (Pt 2), 522–534. https://doi.org/10.1107/S139900471302926X.

130. Grinberg, I. R.; Yaniv, O.; de Ora, L. O.; Muñoz-Gutiérrez, I.; Hershko, A.; Livnah, O.; Bayer, E. A.; Borovok, I.; Frolow, F.; Lamed, R.; Voronov-Goldman, M. Distinctive Ligand-Binding Specificities of Tandem PA14 Biomass-Sensory Elements from Clostridium Thermocellum and Clostridium Clariflavum. Proteins 2019, 87 (11), 917–930. https://doi.org/10.1002/prot.25753.

131. Bahari, L.; Gilad, Y.; Borovok, I.; Kahel-Raifer, H.; Dassa, B; Nataf, Y.; Shoham, Y.; Lamed, R.; Bayer, E. A. Glycoside Hydrolases as Components of Putative Carbohydrate Biosensor Proteins in Clostridium Thermocellum. J. Ind. Microbiol. Biotechnol. 2011, 38 (7), 825–832. https://doi.org/10.1007/s10295-010-0848-9.

132. Bayer, E. A.; Kenig, R.; Lamed, R. Adherence of Clostridium Thermocellum to Cellulose. J. Bacteriol. 1983, 156 (2), 818–827. https://doi.org/10.1128/jb.156.2.818-827.1983.

133. Lamed, R.; Setter, E.; Bayer, E. A. Characterization of a Cellulose-Binding, Cellulase-Containing Complex in Clostridium Thermocellum. J. Bacteriol. 1983, 156 (2), 828–836. https://doi.org/10.1128/jb.156.2.828-836.1983.

134. Artzi, L.; Bayer, E. A.; Moraïs, S. Cellulosomes: Bacterial Nanomachines for Dismantling Plant Polysaccharides. Nat. Rev. Microbiol. 2017, 15 (2), 83–95. https://doi.org/10.1038/nrmicro.2016.164.

135. Bayer, E. A.; Belaich, J.-P.; Shoham, Y.; Lamed, R. The Cellulosomes: Multienzyme Machines for Degradation of Plant Cell Wall Polysaccharides. Annu. Rev. Microbiol. 2004, 58, 521–554. https://doi.org/10.1146/annurev.micro.57.030502.091022.

136. Stahl, S. W.; Nash, M. A.; Fried, D. B.; Slutzki, M.; Barak, Y.; Bayer, E. A.; Gaub, H. E. Single-Molecule Dissection of the High-Affinity Cohesin-Dockerin Complex. Proc. Natl. Acad. Sci. U. S. A. 2012, 109 (50), 20431–20436. https://doi.org/10.1073/pnas.1211929109.

137. Pagès, S.; Bélaïch, A.; Bélaïch, J. P.; Morag, E.; Lamed, R.; Shoham, Y.; Bayer, E. A. Species-Specificity of the Cohesin-Dockerin Interaction between Clostridium Thermocellum and Clostridium Cellulolyticum: Prediction of Specificity Determinants of the Dockerin Domain. Proteins 1997, 29 (4), 517–527.

138. Smith, S. P.; Bayer, E. A. Insights into Cellulosome Assembly and Dynamics: From Dissection to Reconstruction of the Supramolecular Enzyme Complex. Curr. Opin. Struct. Biol. 2013, 23 (5), 686–694. https://doi.org/10.1016/j.sbi.2013.09.002.

139. Barba-Cedillo, V.; Montanier, C. Y. Effect of Multimodularity and Spatial Organization of Glycoside Hydrolases on Catalysis. Essays Biochem. 2023, 67 (3), 629–638. https://doi.org/10.1042/EBC20220167.

140. Morag, E.; Halevy, I.; Bayer, E. A.; Lamed, R. Isolation and Properties of a Major Cellobiohydrolase from the Cellulosome of Clostridium Thermocellum. J. Bacteriol. 1991, 173 (13), 4155–4162. https://doi.org/10.1128/jb.173.13.4155-4162.1991.

141. Wang, Y.; Li, L.; Xia, Y.; Zhang, T. Reliable and Scalable Identification and Prioritization of Putative Cellulolytic Anaerobes With Large Genome Data. Front. Bioinforma. 2022, 2, 813771. https://doi.org/10.3389/fbinf.2022.813771.

142. Raman, B.; Pan, C.; Hurst, G. B.; Rodriguez, M.; McKeown, C. K.; Lankford, P. K.; Samatova, N. F.; Mielenz, J. R. Impact of Pretreated Switchgrass and Biomass Carbohydrates on Clostridium Thermocellum ATCC 27405 Cellulosome Composition: A Quantitative Proteomic Analysis. PloS One 2009, 4 (4), e5271. https://doi.org/10.1371/journal.pone.0005271.

143. Artzi, L.; Dadosh, T.; Milrot, E.; Moraïs, S.; Levin-Zaidman, S.; Morag, E.; Bayer, E. A. Colocalization and Disposition of Cellulosomes in Clostridium Clariflavum as Revealed by Correlative Superresolution Imaging. mBio 2018, 9 (1), e00012-18. https://doi.org/10.1128/mBio.00012-18.

144. Berg Miller, M. E.; Antonopoulos, D. A.; Rincon, M. T.; Band, M.; Bari, A.; Akraiko, T.; Hernandez, A.; Thimmapuram, J.; Henrissat, B.; Coutinho, P. M.; Borovok, I.; Jindou, S.; Lamed, R.; Flint, H. J.; Bayer, E. A.; White, B. A. Diversity and Strain Specificity of Plant Cell Wall Degrading Enzymes Revealed by the Draft Genome of Ruminococcus Flavefaciens FD-1. PloS One 2009, 4 (8), e6650. https://doi.org/10.1371/journal.pone.0006650.

145. Rincon, M. T.; Cepeljnik, T.; Martin, J. C.; Lamed, R.; Barak, Y.; Bayer, E. A.; Flint, H. J. Unconventional Mode of Attachment of the Ruminococcus Flavefaciens Cellulosome to the Cell Surface. J. Bacteriol. 2005, 187 (22), 7569–7578. https://doi.org/10.1128/JB.187.22.7569-7578.2005.

146. Zhivin, O.; Dassa, B.; Moraïs, S.; Utturkar, S. M.; Brown, S. D.; Henrissat, B.; Lamed, R.; Bayer, E. A. Unique Organization and Unprecedented Diversity of the Bacteroides (Pseudobacteroides) Cellulosolvens Cellulosome System. Biotechnol. Biofuels 2017, 10, 211. https://doi.org/10.1186/s13068-017-0898-6.

147. Haimovitz, R.; Barak, Y.; Morag, E.; Voronov-Goldman, M.; Shoham, Y.; Lamed, R.; Bayer, E. A. Cohesin-Dockerin Microarray: Diverse Specificities between Two

Complementary Families of Interacting Protein Modules. Proteomics 2008, 8 (5), 968–979. https://doi.org/10.1002/pmic.200700486.

148. Shimon, L. J.; Bayer, E. A.; Morag, E.; Lamed, R.; Yaron, S.; Shoham, Y.; Frolow, F. A Cohesin Domain from Clostridium Thermocellum: The Crystal Structure Provides New Insights into Cellulosome Assembly. Struct. Lond. Engl. 1993 1997, 5 (3), 381–390. https://doi.org/10.1016/s0969-2126(97)00195-0.

149. Dragani, B.; Aceto, A. About the Role of Conserved Amino Acid Residues in the Calcium-Binding Site of Proteins. Arch. Biochem. Biophys. 1999, 368 (2), 211–213. https://doi.org/10.1006/abbi.1999.1312.

150. Lytle, B. L.; Volkman, B. F.; Westler, W. M.; Heckman, M. P.; Wu, J. H. Solution Structure of a Type I Dockerin Domain, a Novel Prokaryotic, Extracellular Calcium-Binding Domain. J. Mol. Biol. 2001, 307 (3), 745–753. https://doi.org/10.1006/jmbi.2001.4522.

151. Carvalho, A. L.; Dias, F. M. V.; Nagy, T.; Prates, J. A. M.; Proctor, M. R.; Smith, N.; Bayer, E. A.; Davies, G. J.; Ferreira, L. M. A.; Romão, M. J.; Fontes, C. M. G. A.; Gilbert, H. J. Evidence for a Dual Binding Mode of Dockerin Modules to Cohesins. Proc. Natl. Acad. Sci. U. S. A. 2007, 104 (9), 3089–3094. https://doi.org/10.1073/pnas.0611173104.

152. Nash, M. A.; Smith, S. P.; Fontes, C. M.; Bayer, E. A. Single versus Dual-Binding Conformations in Cellulosomal Cohesin-Dockerin Complexes. Curr. Opin. Struct. Biol. 2016, 40, 89–96. https://doi.org/10.1016/j.sbi.2016.08.002.

153. Carvalho, A. L.; Dias, F. M. V.; Prates, J. A. M.; Nagy, T.; Gilbert, H. J.; Davies, G. J.; Ferreira, L. M. A.; Romão, M. J.; Fontes, C. M. G. A. Cellulosome Assembly Revealed by the Crystal Structure of the Cohesin-Dockerin Complex. Proc. Natl. Acad. Sci. U. S. A. 2003, 100 (24), 13809–13814. https://doi.org/10.1073/pnas.1936124100.

154. Cameron, K.; Najmudin, S.; Alves, V. D.; Bayer, E. A.; Smith, S. P.; Bule, P.; Waller, H.; Ferreira, L. M. A.; Gilbert, H. J.; Fontes, C. M. G. A. Cell-Surface Attachment of Bacterial Multienzyme Complexes Involves Highly Dynamic Protein-Protein Anchors. J. Biol. Chem. 2015, 290 (21), 13578–13590. https://doi.org/10.1074/jbc.M114.633339.

155. Slutzki, M.; Jobby, M. K.; Chitayat, S.; Karpol, A.; Dassa, B.; Barak, Y.; Lamed, R.; Smith, S. P.; Bayer, E. A. Intramolecular Clasp of the Cellulosomal Ruminococcus Flavefaciens ScaA Dockerin Module Confers Structural Stability. FEBS Open Bio 2013, 3, 398–405. https://doi.org/10.1016/j.fob.2013.09.006.

156. Vera, A. M.; Galera-Prat, A.; Wojciechowski, M.; Różycki, B.; Laurents, D. V.; Carrión-Vázquez, M.; Cieplak, M.; Tinnefeld, P. Cohesin-Dockerin Code in Cellulosomal Dual Binding Modes and Its Allosteric Regulation by Proline Isomerization. Struct. Lond. Engl. 1993 2021, 29 (6), 587-597.e8. https://doi.org/10.1016/j.str.2021.01.006.

157. Brás, J. L. A.; Pinheiro, B. A.; Cameron, K.; Cuskin, F.; Viegas, A.; Najmudin, S.; Bule, P.; Pires, V. M. R.; Romão, M. J.; Bayer, E. A.; Spencer, H. L.; Smith, S.; Gilbert, H. J.; Alves, V. D.; Carvalho, A. L.; Fontes, C. M. G. A. Diverse Specificity of Cellulosome Attachment to the Bacterial Cell Surface. Sci. Rep. 2016, 6, 38292. https://doi.org/10.1038/srep38292.

158. Duarte, M.; Carvalho, A. L.; Ferreira, M. C.; Caires, B.; Romão, M. J.; Prates, J. A. M.; Najmudin, S.; Bayer, E. A.; Fontes, C. M.; Bule, P. Tripartite Binding Mode of Cohesin-Dockerin Complexes from Ruminococcus Flavefaciens Involving Naturally Truncated Dockerins. J. Biol. Chem. 2025, 301 (7), 110325. https://doi.org/10.1016/j.jbc.2025.110325.

159. Alves, V. D.; Fontes, C. M. G. A.; Bule, P. Cellulosomes: Highly Efficient Cellulolytic Complexes. Subcell. Biochem. 2021, 96, 323–354. https://doi.org/10.1007/978-3-030-58971-4_9.

160. Bolam, D. N.; Ciruela, A.; McQueen-Mason, S.; Simpson, P.; Williamson, M. P.; Rixon, J. E.; Boraston, A.; Hazlewood, G. P.; Gilbert, H. J. Pseudomonas Cellulose-Binding Domains Mediate Their Effects by Increasing Enzyme Substrate Proximity. Biochem. J. 1998, 331 (Pt 3) (Pt 3), 775–781. https://doi.org/10.1042/bj3310775.

161. Boraston, A. B.; Kwan, E.; Chiu, P.; Warren, R. A. J.; Kilburn, D. G. Recognition and Hydrolysis of Noncrystalline Cellulose. J. Biol. Chem. 2003, 278 (8), 6120–6127. https://doi.org/10.1074/jbc.M209554200.

162. Caspi, J.; Barak, Y.; Haimovitz, R.; Irwin, D.; Lamed, R.; Wilson, D. B.; Bayer, E. A. Effect of Linker Length and Dockerin Position on Conversion of a Thermobifida Fusca Endoglucanase to the Cellulosomal Mode. Appl. Environ. Microbiol. 2009, 75 (23), 7335–7342. https://doi.org/10.1128/AEM.01241-09.

163. Galera-Prat, A.; Pantoja-Uceda, D.; Laurents, D. V.; Carrión-Vázquez, M. Solution Conformation of a Cohesin Module and Its Scaffoldin Linker from a Prototypical Cellulosome. Arch. Biochem. Biophys. 2018, 644, 1–7. https://doi.org/10.1016/j.abb.2018.02.016.

164. Kang, S.; Barak, Y.; Lamed, R.; Bayer, E. A.; Morrison, M. The Functional Repertoire of Prokaryote Cellulosomes Includes the Serpin Superfamily of Serine Proteinase Inhibitors. Mol. Microbiol. 2006, 60 (6), 1344–1354. https://doi.org/10.1111/j.1365-2958.2006.05182.x.

165. Levy-Assaraf, M.; Voronov-Goldman, M.; Rozman Grinberg, I.; Weiserman, G.; Shimon, L. J. W.; Jindou, S.; Borovok, I.; White, B. A.; Bayer, E. A.; Lamed, R.; Frolow, F. Crystal Structure of an Uncommon Cellulosome-Related Protein Module

from Ruminococcus Flavefaciens That Resembles Papain-like Cysteine Peptidases. PloS One 2013, 8 (2), e56138. https://doi.org/10.1371/journal.pone.0056138.

166. Chen, C.; Cui, Z.; Song, X.; Liu, Y.-J.; Cui, Q.; Feng, Y. Integration of Bacterial Expansin-like Proteins into Cellulosome Promotes the Cellulose Degradation. Appl. Microbiol. Biotechnol. 2016, 100 (5), 2203–2212. https://doi.org/10.1007/s00253-015-7071-6.

167. Currie, M. A.; Adams, J. J.; Faucher, F.; Bayer, E. A.; Jia, Z.; Smith, S. P. Scaffoldin Conformation and Dynamics Revealed by a Ternary Complex from the Clostridium Thermocellum Cellulosome. J. Biol. Chem. 2012, 287 (32), 26953–26961. https://doi.org/10.1074/jbc.M112.343897.

168. Galera-Prat, A.; Moraïs, S.; Vazana, Y.; Bayer, E. A.; Carrión-Vázquez, M. The Cohesin Module Is a Major Determinant of Cellulosome Mechanical Stability. J. Biol. Chem. 2018, 293 (19), 7139–7147. https://doi.org/10.1074/jbc.RA117.000644.

169. Adams, J. J.; Pal, G.; Jia, Z.; Smith, S. P. Mechanism of Bacterial Cell-Surface Attachment Revealed by the Structure of Cellulosomal Type II Cohesin-Dockerin Complex. Proc. Natl. Acad. Sci. U. S. A. 2006, 103 (2), 305–310. https://doi.org/10.1073/pnas.0507109103.

170. Salama-Alber, O.; Jobby, M. K.; Chitayat, S.; Smith, S. P.; White, B. A.; Shimon, L. J. W.; Lamed, R.; Frolow, F.; Bayer, E. A. Atypical Cohesin-Dockerin Complex Responsible for Cell Surface Attachment of Cellulosomal Components: Binding Fidelity, Promiscuity, and Structural Buttresses. J. Biol. Chem. 2013, 288 (23), 16827–16838. https://doi.org/10.1074/jbc.M113.466672.

171. Dassa, B.; Borovok, I.; Lombard, V.; Henrissat, B.; Lamed, R.; Bayer, E. A.; Moraïs, S. Pan-Cellulosomics of Mesophilic Clostridia: Variations on a Theme. Microorganisms 2017, 5 (4), 74. https://doi.org/10.3390/microorganisms5040074.

172. Yoav, S.; Barak, Y.; Shamshoum, M.; Borovok, I.; Lamed, R.; Dassa, B.; Hadar, Y.; Morag, E.; Bayer, E. A. How Does Cellulosome Composition Influence Deconstruction of Lignocellulosic Substrates in Clostridium (Ruminiclostridium) Thermocellum DSM 1313? Biotechnol. Biofuels 2017, 10, 222. https://doi.org/10.1186/s13068-017-0909-7.

173. Navarre, W. W.; Schneewind, O. Proteolytic Cleavage and Cell Wall Anchoring at the LPXTG Motif of Surface Proteins in Gram-Positive Bacteria. Mol. Microbiol. 1994, 14 (1), 115–121. https://doi.org/10.1111/j.1365-2958.1994.tb01271.x.

174. Mesnage, S.; Fontaine, T.; Mignot, T.; Delepierre, M.; Mock, M.; Fouet, A. Bacterial SLH Domain Proteins Are Non-Covalently Anchored to the Cell Surface via a Conserved Mechanism Involving Wall Polysaccharide Pyruvylation. EMBO J. 2000, 19 (17), 4473–4484. https://doi.org/10.1093/emboj/19.17.4473.

175. Mahoney, B. J.; Takayesu, A.; Zhou, A.; Cascio, D.; Clubb, R. T. The Structure of the Clostridium Thermocellum RsgI9 Ectodomain Provides Insight into the Mechanism of Biomass Sensing. Proteins 2022, 90 (7), 1457–1467. https://doi.org/10.1002/prot.26326.

176. Takayesu, A.; Mahoney, B. J.; Goring, A. K.; Jessup, T.; Ogorzalek Loo, R. R.; Loo, J. A.; Clubb, R. T. Insight into the Autoproteolysis Mechanism of the RsgI9 Anti-σ Factor from Clostridium Thermocellum. Proteins 2024, 92 (8), 946–958. https://doi.org/10.1002/prot.26690.

177. Minor, C. M.; Takayesu, A.; Ha, S. M.; Salwinski, L.; Sawaya, M. R.; Pellegrini, M.; Clubb, R. T. A Genomic Analysis Reveals the Diversity of Cellulosome Displaying Bacteria. Front. Microbiol. 2024, 15, 1473396. https://doi.org/10.3389/fmicb.2024.1473396.

178. Boerjan, W.; Ralph, J.; Baucher, M. Lignin Biosynthesis. Annu. Rev. Plant Biol. 2003, 54, 519–546. https://doi.org/10.1146/annurev.arplant.54.031902.134938.

Chapter 2

The Structure of the *Clostridium thermocellum* RsgI9 Ectodomain Provides Insight into the Mechanism of Biomass Sensing

The work described in this chapter has been reproduced with permission from:

Mahoney BJ, **Takayesu A**, Zhou A, Cascio D, Clubb RT. The structure of the Clostridium thermocellum RsgI9 ectodomain provides insight into the mechanism of biomass sensing. Proteins. 2022 Jul;90(7):1457-1467. doi: 10.1002/prot.26326.

https://onlinelibrary.wiley.com/doi/10.1002/prot.26326

I contributed to this publication with *in vitro* biochemical carbohydrate binding assays, protease activity experiments, figure making, and analysis of the data.

2.1 Overview

Clostridium thermocellum is actively being developed as a microbial platform to produce biofuels and chemicals from renewable plant biomass. An attractive feature of this bacterium is its ability to efficiently degrade lignocellulose using surface-displayed cellulosomes, large multi-protein complexes that house different types of cellulase enzymes. *Clostridium thermocellum* tailors the enzyme composition of its cellulosome using nine membrane-embedded anti-σ factors (RsgI1-9), which are thought to sense different types of extracellular carbohydrates and then elicit distinct gene expression programs via cytoplasmic σ factors. Here we show that the RsgI9 anti-σ factor interacts with cellulose via a C-terminal bi-domain unit. A 2.0 Å crystal structure reveals that the unit is constructed from S1C peptidase and NTF2-like protein domains that contain a potential binding site for cellulose. Small-angle X-ray scattering experiments of the intact ectodomain indicate that it adopts a bi-lobed, elongated conformation. In the structure, a conserved RsgI extracellular (CRE) domain is connected to the bi-domain via a proline-rich linker, which is expected to project the carbohydrate-binding unit ~160 Å from the cell surface. The CRE and proline-rich elements are conserved in several other *C. thermocellum* anti-σ factors, suggesting that they will also form extended structures that sense carbohydrates.

2.2 Introduction

Concerns about climate change and declining petroleum supplies have created a pressing need for renewable transportation fuels [1]. Bioethanol produced from lignocellulosic plant biomass (called cellulosic ethanol) is potentially an attractive solution to this problem, as plant biomass is highly abundant and renewable. Typically, cellulosic ethanol is produced by first degrading lignocellulose into its component sugars using purified cellulase enzymes, followed by fermentation of the sugars into ethanol [2]. However, the high costs of the cellulase cocktails used in this process are prohibitive, such that relatively little cellulosic ethanol is produced industrially [3,4]. To increase the cost-efficiency of bioethanol production, much attention has been devoted to developing a consolidated bioprocessing (CBP) platform that does not require the use of purified cellulases. In CBP, a single organism would both produce the cellulases that hydrolyze the cellulose and hemicellulose polymers in plant biomass as well as ferment the resultant sugars into ethanol, or other biofuels or materials [5,6]. *Clostridium thermocellum* (also renamed as *Ruminiclostridium thermocellum, Hungateiclostridium thermocellum,* and *Acetivibrio thermocellus*) is actively being developed as a CBP platform due to its high native cellulolytic activity, ethanologenic potential, and preference to grow under thermophilic conditions [7]. The ability of *C. thermocellum* to degrade lignocellulose is particularly notable, as it is one of the most cellulolytic organisms thus far discovered [8,9]. Current research is focused toward uncovering fundamental genome-encoded properties that confer potent, tunable cellulolytic activity in *C. thermocellum*, and are an important step toward developing cost-effective microbial approaches to produce biofuels and chemicals from plant biomass.

C. thermocellum and other anaerobic bacteria within the bacterial orders Clostridiales and Bacteroidales degrade lignocellulose using cellulosomes, multi-enzyme complexes that house an array of enzymes with different substrate specificities (cellulases, hemicellulases, pectinases, esterases etc.) [10]. Cellulosome-displaying bacteria degrade lignocellulose more

efficiently than microbes that simply secrete cellulases, because enzyme colocalization within the cellulosome promotes enzyme-enzyme synergy, enzyme-proximity enhancement, and cellulose-enzyme-microbe interactions [11,12]. Cellulosomes adopt elaborate structures that are built using a series of surface-displayed scaffoldin proteins that coordinate the binding of the enzyme machinery via dockerin-cohesin domain interactions. They are assembled in modular fashion to create massive polycellulosomal structures that are readily visible by electron microscopy and can harbor >140 distinct dockerin-borne enzymes in some bacterial species [10,13]. The primary scaffoldin in the *C. thermocellum* cellulosome, CipA, contains nine type I cohesin modules that bind to cellulases harboring type I dockerin modules [7]. The scaffoldin also contains a carbohydrate-binding module (CBM) that tethers the cellulosome complex to its cellulose substrate, as well as a C-terminal type II dockerin module that anchors the complex to the cell by interacting with surface displayed proteins that contain type II cohesin domains [14,15].

To degrade various types forms of plant biomass, *C. thermocellum* tailors the enzyme composition of its cellulosome using membrane embedded anti-σ factors [16,17]. Work by the Lamed and Bayer groups has shown that ten membrane-associated anti-σ factors may control the composition of the cellulosome, nine of which share homology to the *B. subtilis* RsgI anti-σ factor (named RsgI1 to RsgI9) [18]. Based on studies of SigI1 and RsgI1, each anti-σ factor is thought to bind to a different type of biomass-derived carbohydrate on the extracellular surface, and then trigger distinct gene expression programs by releasing a cognate extracytoplasmic function (ECF) σ factor (SigI1 to SigI8) that confers promoter specificity for the RNA polymerase [17-19]. With the notable exception of the RsgI9, genes expressing each anti-σ factor are located on the same operon as a gene encoding an ECF σ factor, suggesting that their protein products are functionally linked; for example, the *rsgI1* and *sigI1* genes are expressed from the same operon suggesting Rsg1 and SigI1 proteins form a functional anti-σ/σ pair. However, our

understanding of how these regulatory systems sense carbohydrates and alter gene expression remains limited. *In vitro* binding studies have experimentally identified carbohydrates that interact with the RsgI1, RsgI2, RsgI3, RsgI4, and RsgI6 anti-σ factors, but the specific genes whose expression is ultimately regulated by these binding events is generally not known [18,20-22]. This is because at present only partial regulons have only been identified for SigI1, SigI3, and SigI6 using bioinformatics analyses and verified experimentally [17]. Finally, the molecular mechanism through which extracellular carbohydrate binding to each anti-σ factor triggers gene expression by releasing its cognate ECF σ factor remains to be determined.

The function of the RsgI9 anti-σ factor is poorly understood because it is an 'orphan'; the gene encoding RsgI9 is not located within an operon that also contains a gene that encodes for an ECF σ factor. Moreover, it is not known if RsgI9 is capable of sensing biomass as its primary sequence does not encode for a known carbohydrate-binding module (CBM) or carbohydrate-active enzyme. Herein, we report the first characterization of RsgI9's ectodomain to gain insight into its function. Using X-ray crystallography, we show that RsgI9 contains a unique C-terminal bi-domain unit that is formed from S1C peptidase and NTF2-like domains. Biochemical experiments indicate that the NTF2-like domain interacts with cellulose, while the S1C peptidase domain is not enzymatically active. In the ectodomain of RsgI9, the C-terminal bi-domain module is connected to a Conserved RsgI Extracellular (CRE) domain via a proline-rich linker. Using Small Angle X-ray Scattering (SAXS), we show that RsgI9's intact ectodomain primarily adopts an extended conformation that can be expected to project the bi-domain module up to 160 Å from the cell surface. Inspection of other RsgI proteins reveals that they also contain CRE domains and proline-rich linkers, indicating they will also adopt extended structures to sense different types of biomass.

2.3 Materials and Methods

2.3.1 Cloning and expression of RsgI9.

The nucleotide sequence encoding the C-terminal region of the *C. thermocellum* (DSM 1313) RsgI9 ectodomain (residues 387-702, $RsgI9^{S1C\text{-}CTD}$) was purchased from Twist Bioscience (San Francisco, CA). The pET-29b expression plasmid includes an N-terminal hexahistidine (6xHis) tag followed by a TEV protease recognition site. The plasmid was transformed into *E. coli* BL21 (DE3) cells (New England Biolabs) and grown at 37°C in LB media in the presence of 50 µg/mL kanamycin until OD_{600} of 0.6-0.8 was reached, followed by induction with 1 mM IPTG and overnight (12-18 hours) protein expression at 17°C. Expression pellets were resuspended in lysis buffer containing 50 mM sodium phosphate, 300 mM NaCl, 2 mM DTT, pH 8.0, and lysed via sonication in the presence of 1 mg/mL egg white lysozyme, 400 µL of protease inhibitor cocktail (Calbiochem), 2 mM phenylmethanesulfonyl fluoride (PMSF, Sigma), and 0.5 mg *S. marcescens* nuclease [23]. Lysates were clarified by centrifugation at 15,000 rpm for 45 minutes. The supernatant was applied to HisPur cobalt resin (Thermo Fisher Scientific) equilibrated in lysis buffer, and unbound proteins were washed by several column volumes of lysis buffer followed by a wash buffer (lysis buffer + 10 mM imidazole). The bound protein was eluted using an elution buffer composed of the lysis buffer with 200 mM imidazole. TEV protease (purified in-house) was added to the eluted sample followed by dialysis against imidazole-free lysis buffer. The sample was then loaded back onto cobalt resin and the cleaved protein was collected by rinsing with lysis buffer. The resulting protein was further purified by gel-filtration chromatography with a Superdex S75 size-exclusion column on an Akta FPLC system (GE Healthcare Life Sciences). Protein was concentrated via Amicon Ultra centrifugal filters (EMD Millipore), with purity confirmed to be >98% via SDS-PAGE.

The nucleotide sequence encoding the full extracellular region of RsgI9 (residues 167-707, $RsgI9^{ecto}$) was cloned out of *C. thermocellum* (ATCC 27405) genomic DNA and inserted

into a pET-28b-based vector containing N-terminal 6xHis and small ubiquitin-like modifier (SUMO) fusion tags using Gibson assembly. Growth and expression in *E. coli* BL21 (DE3) cells proceeded as described above. Purification was also carried out identically, but with ULP1 protease (purified in-house) used to cleave the 6xHis-SUMO tag from RsgI9. Expression plasmids producing the isolated Conserved RsgI Extracellular (CRE) portion of the ectodomain (residues 167-343, $RsgI9^{CRE}$) and the S1C domain (residues 387-592, $RsgI9^{S1C}$) were also prepared by cloning the relevant portions of the gene, with expression and purification carried out as above. Site-directed mutagenesis was used to create a T535S variant of the protease domain, $RsgI9^{S1C,T535S}$.

2.3.2 Crystallization, data collection, and crystal structure determination.

$RsgI9^{S1C-CTD}$ was concentrated to 10 mg/mL, and spontaneously formed small protein crystals in a solution of 20 mM HEPES, 2 mM DTT, pH 8.0. However, these crystals were too small to be suitable for diffraction. The small crystals were dissolved in the same buffer at a concentration of 8 mg/mL and used to screen in 24-well hanging drop format with increasing concentrations of glycerol and PEG chain additives, and larger crystals were grown after a week in 20 mM HEPES, 2 mM DTT, pH 7.5, and 10% PEG-3350. For data collection, the crystals were cryoprotected with reservoir solution containing 30% glycerol. Diffraction datasets were collected at the Advanced Photon Source (Argonne National Laboratory) on beamline 24-ID-C equipped with a Dectris EIGER2 X 16M detector at 100K. Data was collected at a wavelength of 0.97903 Å, detector distance of 280 mm, and 0.25° oscillations. The crystals diffracted X-rays to 2.0 Å resolution. XDS/XSCALE was used to index, integrate, and scale data in the $P2_1$ space group [24]. The asymmetric unit of the crystal contained two molecules, resulting in a Matthews coefficient of 2.68 $Å^3$/Da and 54.15% solvent content in the unit cell[25-27]. The MRage automated suite in Phenix was used to perform molecular replacement with templates of S1C domain homologs using its built-in NCBI BLAST search [28-30]. The top 5 hits each resulted in

excellent solutions (LLG > 130), with the top solution (PDB 5B6L, LLG of 162.5) used to continue building a complete structure using the AutoBuild function [31,32]. The structure was iteratively improved by manual rebuilding in Coot and automatic refinement in BUSTER, with NCS restraints enabled and TLS groups defined for the separate domains [33-36]. Complete refinement and structure statistics are reported in **Table 1**. Coordinates and structure factors have been deposited in the Protein Data Bank under the accession code 7SJY. Figure images were generated using PyMOL 2.4.1 (Schrödinger, LLC).

2.3.3 Nuclear magnetic resonance and small-angle X-ray scattering (SAXS) experiments.

A ^{15}N-labeled sample of RsgI9$^{S1C\text{-}CTD}$ was produced by following the expression and purification methods above, but with expression induced after exchanging the cultures into minimal media supplemented with $^{15}NH_4Cl$ (Cambridge Isotope Laboratories, Tewksbury, MA) [37]. Spectra were acquired using a 0.4 mM sample in NMR buffer (50 mM sodium phosphate, 200 mM NaCl, 0.03% sodium azide, pH 7.8) at 298 K on a Bruker Avance III HD 600 MHz (14.1 T) spectrometer equipped with a triple resonance cryogenic probe. A 2-D ^{15}N-^{1}H TROSY-HSQC spectrum showed a well-folded protein. The rotational correlation time (τ_c) of RsgI9$^{S1C\text{-}CTD}$ was estimated from a series of 1-D ^{15}N-TRACT experiments, with 2,048 complex points, 32 transients, 100 experiments for each spin state, and the relevant delay incremented by 4 ms [38]. The decrease in integrated backbone amide intensity was fitted to an exponential decay function, resulting in a measured transverse cross-correlated relaxation rate (η_{xy}) of 22.9 Hz, which corresponds to a τ_c value of ~ 22.1 ns, calculated via algebraic solutions [39].

Size-exclusion chromatography small-angle X-ray scattering (SEC-SAXS) was performed at the SIBYLS beamline (Advanced Light Source beamline 12.3.1, Lawrence Berkeley National Laboratory) [40]. Samples were injected at 5 mg/mL on an Agilent 1290 HPLC with a Shodex KW-803 column equilibrated in size exclusion buffer (50 mM sodium phosphate, 300 mM NaCl, pH 7.0) at a flow rate of 0.65 mL/min. X-ray scattering of eluent was continuously

collected on a Dectris PILATUS3 X 2M detector in two-second frames, with X-ray wavelength at 1.216 Å and a sample-to-detector distance of 2.0 m. Multi-angle light scattering (MALS), quasi-elastic light scattering (QELS), and refractometry data were also collected on this eluent and processed using the ASTRA software package (Wyatt) to estimate molecular weight. SAXS frames from before and after the protein elution peak were used to subtract the scattering contribution of the buffer alone, and frames corresponding to protein were merged using ScÅtter IV (www.bioisis.net). SAXS analysis (Guinier analysis and distance distribution calculations) to determine R_g and $p(r)$ function was performed using BioXTAS RAW and the ATSAS Suite [41,42]. Molecular weights were estimated using the Bayesian inference approach [43]. The experimental SAXS data of RsgI9$^{S1C\text{-}CTD}$ was compared to the theoretical scattering curve from the crystal structure (with missing N-terminal residues added using MODELLER) using the FoXS server [44-46]. For RsgI9^{CRE}, RsgI9$^{S1C\text{-}CTD}$, and RsgI9^{ecto}, *ab initio* bead modelling was performed on the output of GNOM using DAMMIF to generate 15 models in slow mode, following by averaging and refinement with DAMMIN [47,48]. DENSS was used to reconstruct electron density of the same constructs in slow mode to generate 20 density maps, followed by averaging and refinement [49]. Models were fit into maps using UCSF ChimeraX, which was also used to generate images [50].

2.3.4 Cellulose-binding and peptidase activity assays.

Binding to the microcrystalline cellulose substrate (Avicel, Sigma-Aldrich) was qualitatively assayed via SDS-PAGE pull-down assays modified from previous protocols [20,22]. Recombinant protein and control samples were prepared using 50 µg of protein in 200 µL of lysis buffer (50 mM Tris-HCl, 300 mM NaCl, pH 8.0) and mixed with 15 µg of microcrystalline cellulose (Avicel). Each sample was thoroughly vortexed and then incubated at room temperature with gentle rotation for 60 min. After incubation, samples were separated by centrifugation at 12,000 x *g* for 10 min. 10 µL of the supernatant containing unbound protein

was taken for SDS-PAGE analysis. The remaining sedimented polysaccharide was washed five times with 200 µL aliquots of lysis buffer with 0.1% Triton-200 detergent to mitigate nonspecific protein binding. After the final centrifugation, the sedimented polysaccharide and bound protein was resuspended in 200 µL SDS loading buffer and boiled for 5 min before loading for SDS-PAGE. Each assay was repeated at least three times.

The proteolytic activity of the separate RsgI9 domains was quantitatively measured with UV spectroscopy in triplicate by monitoring the hydrolysis of the modified nonspecific protease substrate azocasein (Sigma-Aldrich). 200 µM protein samples were dissolved in lysis buffer and combined in equal volume with 1% azocasein dissolved in glycine-NaOH buffer (100 mM, pH 10.0 or pH 8.0) for a final reaction volume of 500 µL. A sample containing no enzyme was used as a blank and additional control. Samples were incubated at either 37°C or 50°C with gentle rotation for 30 min. Each reaction was terminated via the addition of 750 µL of 10% trifluoroacetic acid (TFA) and subsequently chilled on ice for 30 min. Samples were centrifuged at 10,000 g for 10 min to separate precipitated enzyme after TFA addition. 125 µL samples were plated into a 96-well plate and the absorbance at 440 nm (A_{440}) was measured using a SpectraMax iD3 Multi-Mode Microplate spectrophotometer (Molecular Devices).

2.4 Results and Discussion

2.4.1 The ectodomain in the RsgI9 anti-σ factor contains a C-terminal bi-domain module

We inspected the primary sequence of the RsgI9 anti-σ factor to gain insight into its function. An alignment of the amino acid sequences of the nine RsgI anti-σ factors in *C. thermocellum* (RsgI1-9) reveals that they possess the same basic architecture (**Fig. 1A**). Each contains a cytoplasmic N-terminal region that is followed by a transmembrane helix and a larger C-terminal ectodomain that in some of the anti-σ factors has been shown to bind to carbohydrates [17]. Several regions are highly conserved amongst the proteins (**Fig. 1A**, shaded orange). They include a ~40 amino acid N-terminal segment annotated RsgI_N (Pfam: PF12791) that binds to cognate σ factors in the cytoplasm, and a ~170 residue segment that immediately follows the transmembrane helix and is presumably located on the extracellular surface, hereafter called the Conserved RsgI Extracellular (CRE) domain [51]. NMR studies indicate the CRE domain is autonomously folded, but its structure has not been determined [52]. In RsgI9 and several other RsgI proteins a proline-rich segment connects the CRE region to a C-terminal region that is varied amongst the anti-σ factors. In RsgI9, the C-terminal region contains amino acids that share homology with trypsin-like S1C domains (residues I396-H578). Residues following this segment do not display sequence homology to any protein of known structure. However, they are classified by the program DISOPRED3 as being structurally ordered and are hereafter referred to as the C-terminal domain (CTD) [53]. The DISOPRED3 prediction suggests that the S1C and CTD are closely linked, raising the possibility that they form an ordered bi-domain unit. The module may be involved in biomass sensing, as a search of microbial genomes reveals that it is only present in a few bacterial species within the genus *Acetivibrio*, several of which exhibit cellulolytic activity similar to *C. thermocellum* (**Fig. S1**) [54].

2.4.2 The bi-domain module forms a structurally ordered unit that contains peptidase- and NTF2-like domains.

To shed light onto the structure and function of the bi-domain module we determined the 2.0 Å crystal structure of a polypeptide containing the predicted S1C and CTD domains (RsgI9$^{S1C\text{-}CTD}$, residues A387-K702). RsgI9$^{S1C\text{-}CTD}$ crystallized in the $P2_1$ space group and its structure was determined by molecular replacement using the coordinates of the S1C peptidase domain within the HhoA protein as a search model (PDB: 5B6L, 34% sequence identity), followed by iterative model building of the CTD coordinates (Fig. 2A) [31]. Continuous electron density allowed modeling of the entirety of the protein, except for amino acid G502 in the β7-β8 loop. Two RsgI9$^{S1C\text{-}CTD}$ molecules are present in the asymmetric unit and are arranged in head-to-tail manner. The proteins are related by two-fold non-crystallographic symmetry and adopt similar atomic structures (backbone RMSD < 0.5 Å). Dimerization buries 907 Å^2 of protein surface area, but based on a PDBePISA analysis this interface is not biologically relevant [55]. This is consistent with size-exclusion chromatography with multi-angle light scattering detection (SEC-MALS) analyses that show that RsgI9$^{S1C\text{-}CTD}$ is monomeric; the SEC-MALS derived molecular weight is 35.3 ± 0.4 kDa, as compared to a predicted value of 34.5 kDa. Complete data collection and structure refinement statistics are shown in Table 1.

The bi-domain unit adopts a rigid structure that houses potential ligand binding sites within S1C peptidase and NTF2-like domains (Fig. 2A, B). The N-terminal S1C domain is composed of residues P389-H578 and shares structural homology with the Deg/HtrA family of serine proteases. Its structure is formed by two α-helices (αA, αB) that flank a pair of β-barrels (β1-β6 and β7-β12). In Deg/HtrA proteases these β-barrels surround a pocket housing a His-Asp-Ser active site triad that mediates catalysis [56]. However, in RsgI9 the serine residue in the triad is replaced with threonine (H434-D465-T535, RsgI9 numbering). The CTD is formed by residues F596-K702 and contains three helices (αC, αD, and αE) that are placed over a four-stranded β-sheet (β14-β17). Based on DALI and PDBeFold analyses it adopts a NTF2-like fold [28,57], which was originally identified in the Nuclear Transport Factor 2 (NTF2) protein [58].

However, the CTD shares less than 20% sequence identity with NTF2 or its structural homologs. NTF2-like folds are broadly distributed in biology and typically form cone-like structures that harbor a recessed hydrophobic cavity that mediates ligand binding [59]. In the CTD, residues that would enclose this pocket are replaced by a short loop between the αD and αE helices, generating a more open cleft that might serve as a ligand binding site (Fig. 2D, S2). A network of hydrogen-bonding and salt-bridging interactions between the S1C and NTF2 domains buries ~680 $Å^2$ of solvent exposed surface area (Fig. 2C). The short tether that links the primary sequences of the domains is at the heart of the interface (residues S579-D595). It forms strand β13 that connects the β-sheets of each module by hydrogen bonding; strand β13 pairs with strands β5 and β14 in the S1C and CTD folds, respectively. Additionally, the β1/β2 loop in the S1C domain, known as loop LA in serine protease nomenclature, packs over the αC helix of the CTD [60]. These inter-domain interactions appear to be important for the stability of the CTD, as attempts to express a polypeptide containing only this domain were unsuccessful. NMR and SAXS analyses reveal that packing of the S1C and CTD domains causes them to adopt a rigid structure in solution. In particular, the experimental molecular correlation time (τ_c) of $RsgI9^{S1C\text{-}CTD}$ was determined by NMR to be 22.1 ns, which agrees with the theoretical value for a 34.5 kDa globular protein (~20.9 ns) [61,62] (Fig. S3). SAXS experiments described below also indicate that the domains are immobilized with respect to one another. Thus, we conclude that the phylogenetically conserved bi-domain within RsgI9 forms a structurally ordered unit that houses potential ligand binding sites within its S1C and CTD domains.

2.4.3 The bi-domain unit interacts with cellulose.

The RsgI1, RsgI2, RsgI3, RsgI4, and RsgI6 anti-σ factors in *C. thermocellum* bind to carbohydrates and are thought to serve as biomass sensors [20-22]. We therefore probed whether RsgI9 could interact with cellulose using an established pull-down assay [20,22]. In these experiments polypeptides containing either the entire ectodomain ($RsgI9^{ecto}$), the bi-

domain unit (RsgI9$^{S1C-CTD}$), the S1C domain (RsgI9^{S1C}), or the CRE domain (RsgI9^{CRE}) were tested. Only peptides containing the bi-domain unit domains interact with microcrystalline cellulose (Avicel, **Fig. 3A**), as the CRE domain did not interact with any of the carbohydrates that were tested. Control experiments confirm specificity, as a known cellulase Cel48S binds to cellulose in these experiments, whereas no binding was detected for bovine serum albumin (BSA). Interestingly, significantly more RsgI9$^{S1C-CTD}$ protein binds to Avicel as compared to RsgI9^{S1C}, suggesting that the majority of affinity for cellulose originates from the CTD. Although the molecular basis of cellulose binding remains to be determined, it is possible that it is mediated by interactions with CTD's cleft. This is because the walls of this groove contain a series of surface exposed aromatic and non-polar amino acid side chains (**Fig. 3B, C**), which in other carbohydrate-binding proteins form π-stacking interactions with bound sugars [63]. Notably, the *C. thermocellum* RsgI1, RsgI2, and RsgI4 anti-σ factors each contain CBM3 carbohydrate binding domains that employ a linear strip of aromatic amino acids to interact with the pyranoside rings within cellulose [22,64]. Although the CTD and CBM3 adopt distinct folds, the CTD cleft also contains a linear strip of aromatic residues in strands β14-β16 (Y641, Y645, Y664, Y666, F676, Y679) suggesting that it may form similar interactions. A more deeply recessed anionic patch is located immediately adjacent to the cleft and occupied by a glycerol molecule in the structure of RsgI9$^{S1C-CTD}$. Although its function remains unknown, similar anionic pockets are present in carbohydrate-binding proteins and often interact with metal cations (e.g. Ca^{2+}) [65-67].

Despite exhibiting structural homology with members of the Deg/HtrA protease family, the S1C domain in RsgI9 is enzymatically inactive *in vitro*. In marked contrast to the trypsin control, neither the intact ectodomain nor the bi-domain unit exhibits proteolytic activity when assayed using azocasein as a substrate (**Fig. 4A**). The structure of RsgI9's S1C domain is similar to the *E. coli* DegS protease, which mediates the heat shock response by degrading the

RseA anti-σ factor [68]. Superimposing the coordinates of their S1C domains reveals striking similarities in the positioning of residues that form the active site in DegS. The side chains that construct the His-Asp-Ser catalytic triad in DegS overlay well with residues in the RsgI9 S1C domain, except that the serine within the triad is replaced with a threonine in RsgI9 (**Fig. 4B**). Interestingly, a T535S variant of RsgI9 is also enzymatically inactive even though it contains a full complement of triad residues. This suggests that the pocket housing the triad in RsgI9 lacks other essential features that are needed for catalysis [69]. Indeed, a detailed comparison with the structure of DegS reveals that the oxyanion hole used to stabilize reaction intermediates may not be properly formed in RsgI9 (**Fig. S4**) [70]. Interestingly, DegS and the bi-domain unit also have similar domain architectures. In DegS the S1C domain packs against a smaller C-terminal PDZ domain that regulates proteolytic activity in response to binding to unfolded peptides, while in RsgI9 the S1C domain is packed against a NTF2-like module that may bind to carbohydrates. As the S1C domain in RsgI9 is inactive, it may simply function as a spacer that properly positions the CTD and/or its pocket containing the His-Asp-Thr triad may be used to bind to a ligand whose identity remains to be determined.

2.4.4 The full-length ectodomain adopts a rod-like conformation that may extend from the cell surface.

To gain insight into the solution conformation of the ectodomain, size-exclusion chromatography coupled with small-angle X-ray scattering (SEC-SAXS) data was collected for polypeptides containing the isolated bi-domain unit and the intact ectodomain. The SEC-SAXS data for RsgI9$^{S1C\text{-}CTD}$ indicates that the bi-domain unit adopts a rigid structure, consistent with its crystal structure and NMR measurements of its molecular correlation time. This is evident from the dimensionless Kratky plot, which shows the presence of a peak near $q^*R_g = \sqrt{3}$ and normalized intensity = 1.1 for RsgI9$^{S1C\text{-}CTD}$ (**Fig. 5A**, red) [71]. Furthermore, the scattering data also fits very well to the FoXS predicted scattering curve for the RsgI9$^{S1C\text{-}CTD}$ crystal structure (χ^2

= 1.23) (**Fig. S5A**) [44], and the calculated P(*r*) function for RsgI9$^{S1C\text{-}CTD}$ is also consistent with the crystal structure (D_{max} of 82 Å versus 80 Å in the crystal structure) (**Fig. 5B**, red). Finally, *ab initio* bead model and electron density reconstructions using DAMMIN/F and DENSS, respectively, fit well to a single polypeptide from the crystal structure (**Fig. S5B**) [47,48].

The SAXS analysis indicates that the intact ectodomain adopts a bi-lobed, rod-shaped structure in solution. RsgI9^{ecto} eluted from the size exclusion chromatography step as a monomer; the theoretical molecular weight of monomeric RsgI9^{ecto} is 59.7 kDa, compatible with estimates from MALS and SAXS of 62.7 ± 0.2 kDa and 58.1 kDa, respectively. Notably, the scattering profile of RsgI9^{ecto} is indicative of an extended protein as evidenced by the up- and right-ward shift of its peak in the dimensionless Kratky plot (**Fig. 5A**, black). Furthermore, the P(*r*) distance distribution function of RsgI9^{ecto} shows a D_{max} at 160 Å with a R_g value of 45 Å, much larger than expected for a compact spherical protein of the same molecular weight (**Fig. 5B**, black) [71]. The peak centered at ~25 Å in the *p*(*r*) function of RsgI9$^{S1C\text{-}CTD}$ is also present in the corresponding plot of RsgI9^{ecto}, but the broad and nearly linear decrease toward its large D_{max} is indicative of a rod-like extension [72]. *Ab initio* electron density reconstruction using DENSS reveals that the ectodomain adopts an elongated, bi-lobed structure (**Fig. 5C**) [49]. The coordinates of the crystal structure of the bi-domain unit fit well into the major lobe of the density, while the small lobe presumably houses the CRE domain that adopts a globular structure. Density in between the lobes corresponds to the proline-rich linker that connects the CRE and S1C domains, but is less well-defined. In totality, the SAXS data strongly support the notion that the ectodomain predominantly adopts an extended configuration in which the bi-domain and CRE units do not interact with one another. Residues in the intervening proline-rich linker (residues 344-386) presumably cause RsgI9 to be elongated. Interestingly, the proline-rich linker is also present in six other *C. thermocellum* anti-σ factors (RsgI1 to RsgI6) and varies in length from 75 to 262 residues (**Fig. S6**). Thus, it appears that like RsgI9, these anti-σ factors

will also adopt elongated structures that project carbohydrate-binding modules away from the cell surface.

2.5 Conclusions

The results of our studies provide new insight into how RsgI9 and other anti-σ factors enable *C. thermocellum* to sense different types of biomass. RsgI9's ectodomain adopts an elongated structure in which two folded domains are connected by a proline-rich linker. A highly conserved CRE domain is presumably located near the cell membrane and is joined by the extended linker to an unusual bi-domain unit that interacts with cellulose. This basic architecture is likely conserved in other *C. thermocellum* anti-σ factors, as the primary sequences of RsgI1 to RsgI6 encode for CRE domains that are connected by proline-rich linkers to carbohydrate binding modules (**Fig. S5**). In order to promote new transcriptional programs, the RsgI-type anti-σ factors presumably transduce carbohydrate binding signals received on the extracellular surface to the cytoplasm, a process that releases the cognate ECF σ factor needed for RNA polymerase targeting to specific promoters. Recent structural studies have provided insight into how the intracellular RsgI_N domain engages its cognate sigma factor [19], but how these interactions are disrupted in response to carbohydrate binding on the cell surface remains unknown. The common architecture adopted by the ectodomains in *C. thermocellum* and the fact that each contains a highly conserved (CRE) domain near the membrane surface suggests that they may employ a similar mechanism to promote signaling.

2.6 Figures

Table 2.1 Crystal data collection and structure refinement statistics

	RsgI9$^{S1C\text{-}CTD}$
Data collection	
Space group	P 1 2_1 1
Cell dimensions	
a, b, c (Å)	63.63, 78.06, 81.21
α, β, γ (°)	90.00, 112.98, 90.00
Resolution (Å)	74.77-2.00 (2.05-2.00)
Wavelength (Å)	0.97903
Total observations	246720 (17477)
Unique reflections	48797 (3594)
R_{merge} (%)	4.9 (61.8)
I / σI	16.51 (2.60)
$CC_{1/2}$	99.9 (83.9)
Completeness (%)	98.4 (98.2)
Multiplicity	5.1 (4.9)
Wilson B-factor (Å^2)	40.84
Refinement	
Resolution (Å)	74.77-2.00
No. of reflections	48797
R_{work} / R_{free} (%)	19.7/21.9
No. atoms	5058
Protein	4788

	RsgI9$^{S1C\text{-}CTD}$
Ligand/ion	78
Water	192
B-factors (Å^2) (all atoms)	51.5
Protein	51.4
Ligand/ion	64.1
Water	48.1
R.m.s. deviations	
Bond lengths (Å)	0.012
Bond angles (°)	1.49
Ramachandran favored (%)	99.03
Ramachandran allowed (%)	0.97
Ramachandran outliers (%)	0.00
PDB ID	7SJY

Values in parentheses are for the highest-resolution shell.

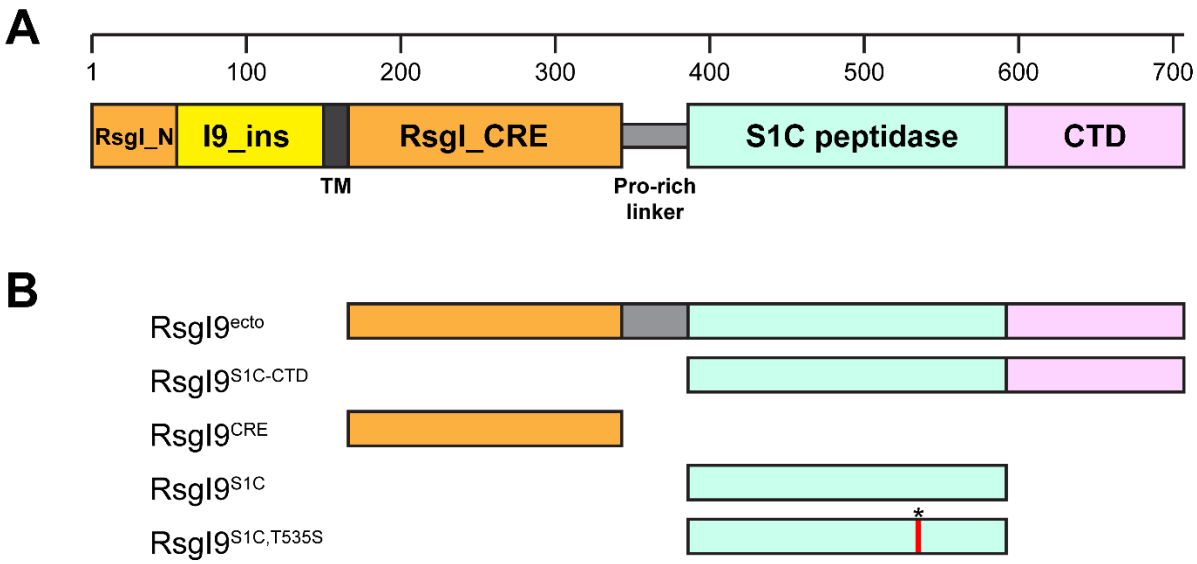


Figure 2.1 RsgI9 contains both unique and RsgI-conserved domains.

(A) RsgI9 contains conserved domains that are found in other RsgI-family proteins (orange). These include an intracellular anti-σ factor domain (RsgI_N, residues 1-55), a transmembrane helix (TM, dark gray) and an extracellular domain of unknown function (CRE or RsgI_CRE, residues 167-343). RsgI9 and many other RsgI proteins also contain a Pro-rich linker segment (light gray) that connects the RsgI_CRE to a variable C-terminal region that differs in each type of anti-σ factor. In RsgI9, the C-terminal region contains a domain that is homologous to S1C peptidase domains (mint, residues 396-578) and a C-terminal domain (CTD) of unknown structure (purple, residues 579-707). RsgI9 also contains a unique insertion immediately following RsgI_N (I9_ins, yellow) that is located in the cytoplasm. (B) Depiction of the RsgI9 polypeptide constructs used in this study. In $RsgI9^{S1C,T535S}$ an asterisk and red line indicates the location of the mutation.

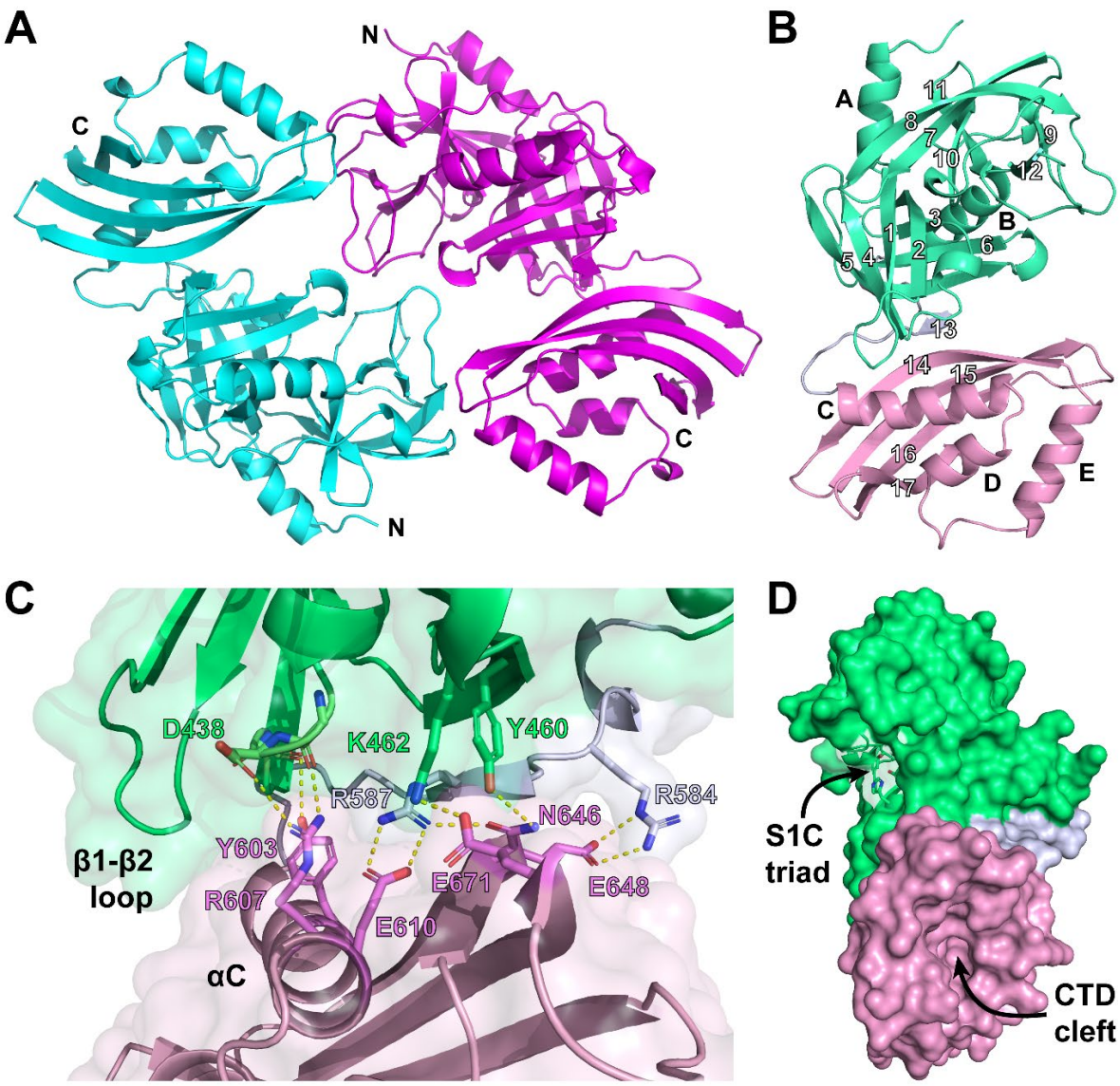


Figure 2.2 Structure of RsgI9's bi-domain unit, RsgI9$^{S1C-CTD}$.

(A) Cartoon representation of the crystal structure of RsgI9$^{S1C-CTD}$ showing the two proteins in the asymmetric unit (N- and C-termini are labeled). (B) Ribbon drawing of a single molecule of RsgI9$^{S1C-CTD}$ with its secondary structural elements labeled. The S1C and CTD domains are colored mint and pink, respectively, while the tether is colored blue-gray. (C) Close-up view of the interface between the S1C and CTD domains, coloring as in the previous panel. Hydrogen-bonding and salt bridge interactions are indicated by yellow dashed lines. (D). Surface

representation of RsgI9$^{S1C\text{-}CTD}$ showing the locations of the recessed cleft in the CTD and a pocket within the S1C domain that in homologous proteins forms a peptidase active site.

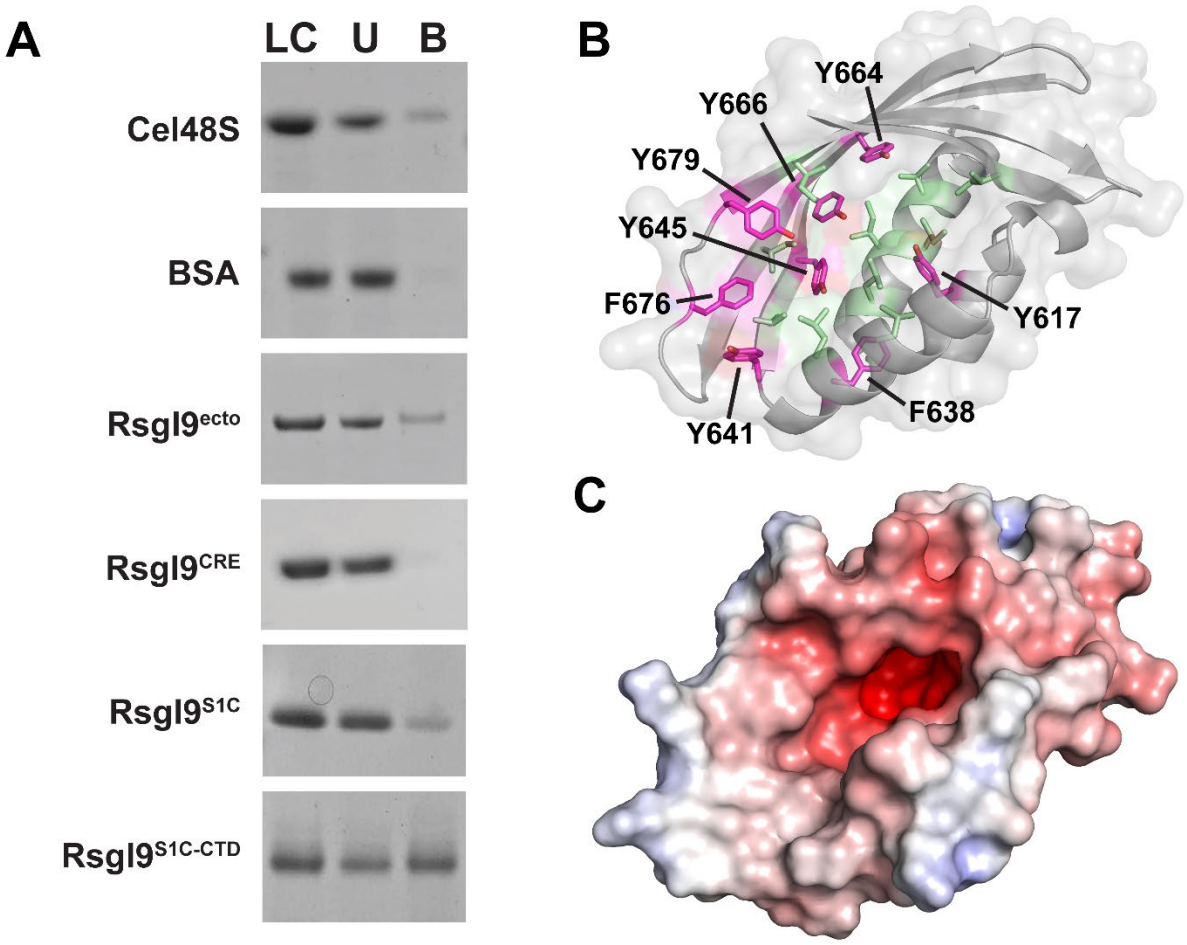


Figure 2.3 The RsgI9 ectodomain binds to crystalline cellulose.

(A) Pull-down assay that tests protein binding to microcrystalline cellulose (Avicel). In each experiment purified protein samples were incubated with Avicel, and then centrifuged to assess binding to this insoluble substrate. Lanes are labeled: LC, loading control; U, unbound protein; B, bound protein. (B) Potential carbohydrate recognition groove in the CTD exposes aromatic (magenta) and non-polar residues (light green). The aromatic sidechains form a continuous strip on one side from Y641 to Y664. (C) Solvent-excluded surface of this domain colored by surface electrostatics (from red to blue, -7.5 to +7.5 *e*) indicates a negatively charged patch at the base of the groove.

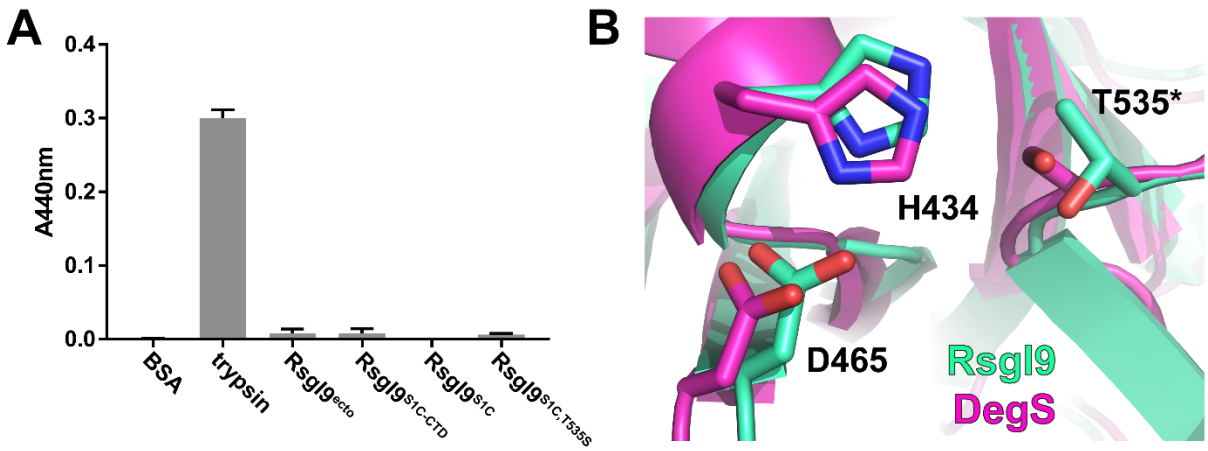


Figure 2.4 Functional studies of the protease-like domain in RsgI9 and its structural homology with DegS.

(A) Protease activity assay results. RsgI9 constructs were all inactive against a promiscuous proteolytic substrate, azocasein. (B) The catalytic triad residues as they appear in RsgI9's S1C domain (mint), including a threonine in place of the conserved serine (T535*), aligned with active form of structural homolog DegS (magenta, PDB: 4RQZ). Residues are numbered as they appear in the RsgI9 primary sequence.

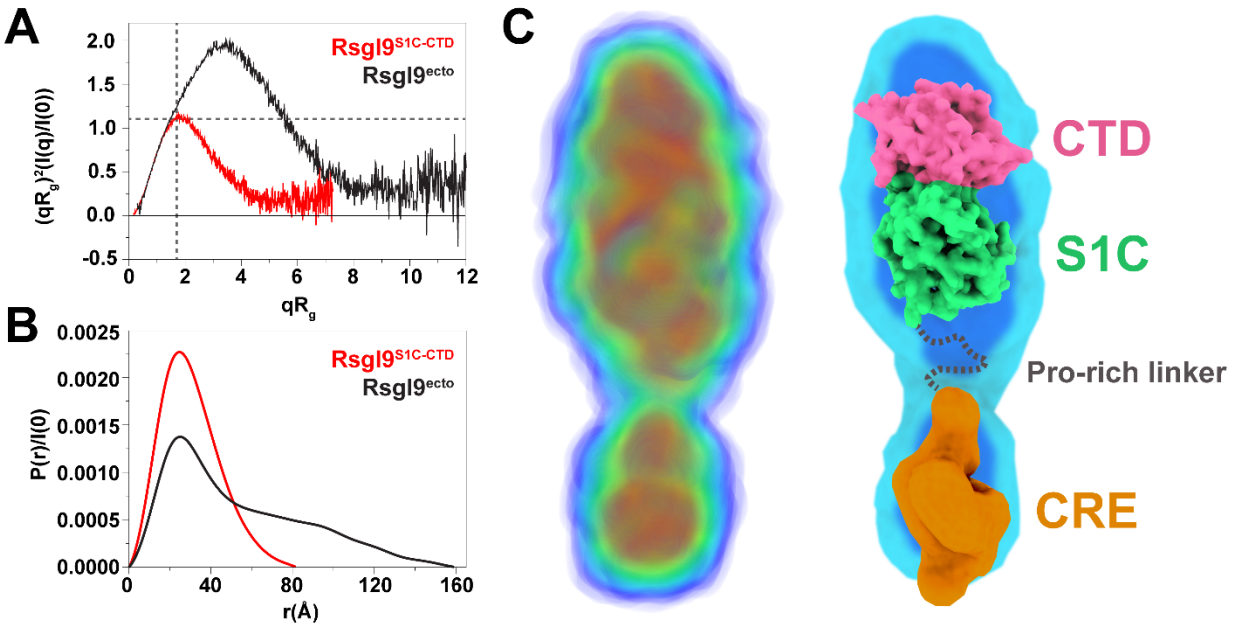


Figure 2.5 SAXS analyses show the Rsgl9 ectodomain forms an extended structure.

(A) Dimensionless Kratky plots of both the $Rsgl9^{S1C-CTD}$ bi-domain unit (red) and intact ectodomain, $Rsgl9^{ecto}$ (black). The two dashed lines indicate the theoretical peak position for a typical globular protein. The bi-domain unit behaves as a compact globular protein, while the intact ectodomain forms an extended rod-like structure. (B) Distance distribution function P(*r*) for both $Rsgl9^{S1C-CTD}$ (red, D_{max} 82 Å) and $Rsgl9^{ecto}$ (black, D_{max} 160 Å). (C) SAXS derived electron density reconstruction of the ectodomain obtained using the program DENSS [49]. The volume of electron density is colored with a gradient from blue to red to indicate regions of highest density (left). Coordinates of the crystal structure of $Rsgl9^{S1C-CTD}$ fitted into the electron density (right). A low-resolution model of the CRE domain that was obtained using the program DENSS and $Rsgl9^{CRE}$ SAXS data is also fitted into the density.

2.7 Supplementary Materials

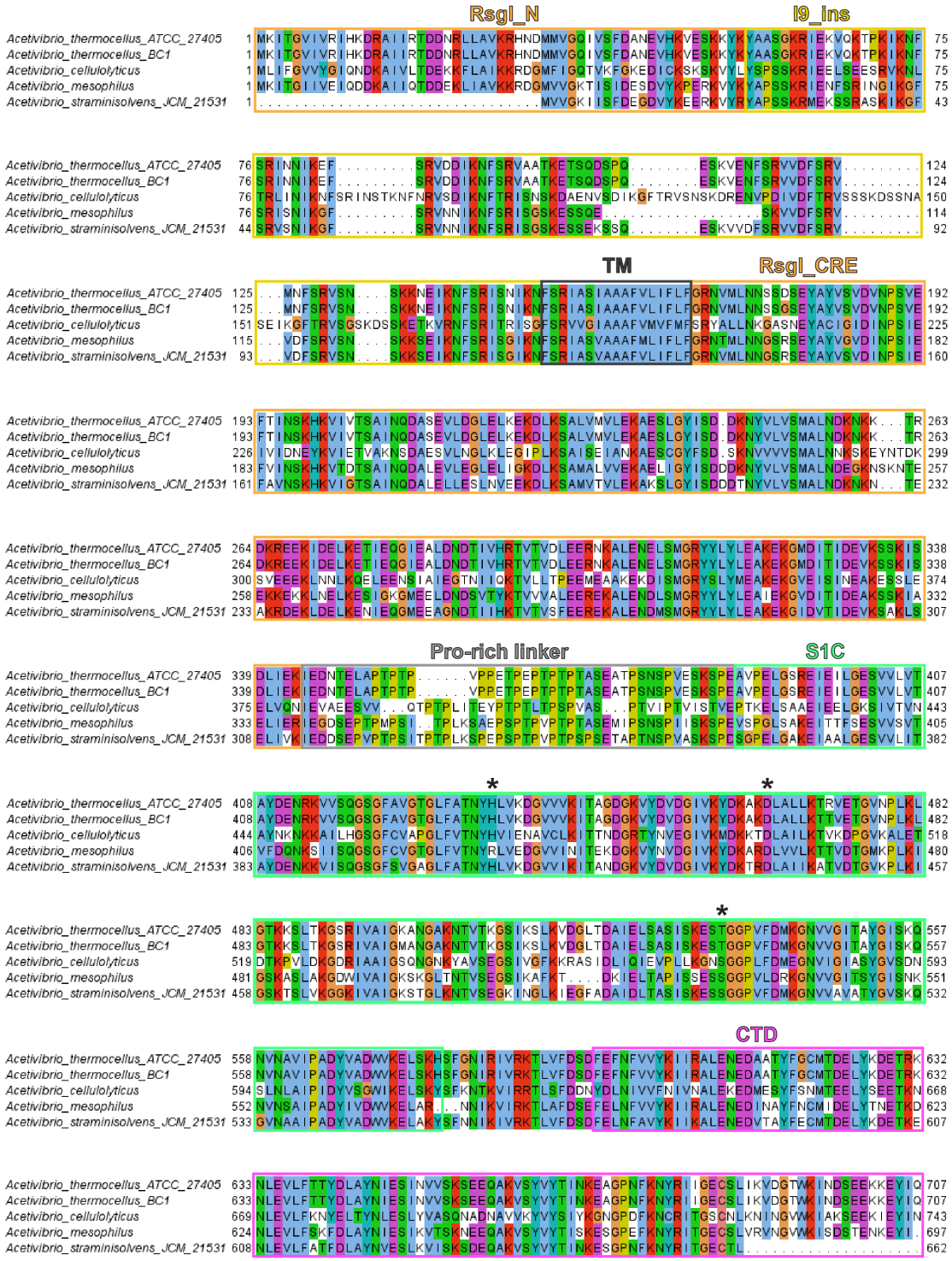


Figure S2.1 Protein sequence alignment of RsgI9 sequences in Clostridia.

Alignment of the amino acid sequences of RsgI9 in *Clostridium thermocellum* (*Acetivibrio thermocellus*) and related species using Clustal Omega [73]. The domains are boxed and labeled in the same colors, and the catalytic triad residues in the S1C peptidase domain are indicated by the asterisks.

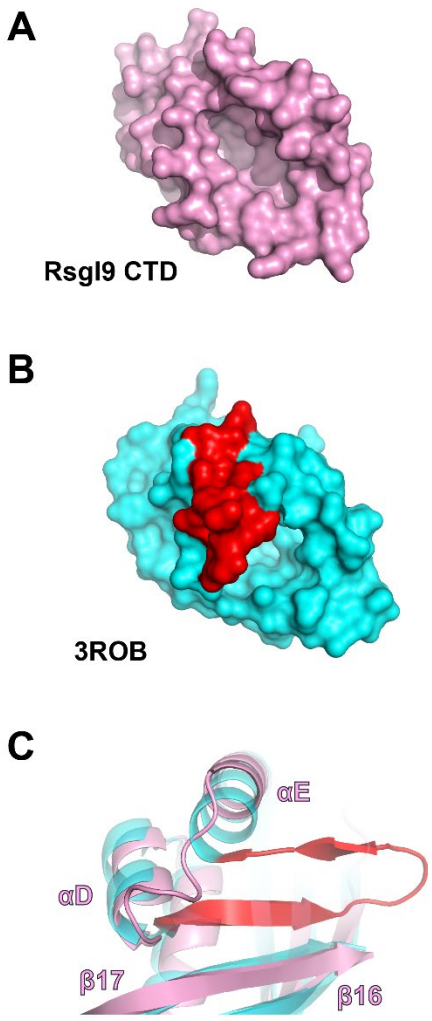


Figure S2.2 Comparison of RsgI9's CTD with a structurally similar NTF2-like protein.

(A) Surface representation of RsgI9 C-terminal domain (CTD) (pink), and (B) a NTF2-like structural homolog shown in the same orientation (cyan) (PDB accession code 3ROB). The structures are similar and have a DALI Z-score = 13.7 and PDBeFold Z-score = 7.8. (C) Cartoon figure showing the extended β-sheet found in other NTF2-like proteins (red) that is missing in RsgI9 (pink). RsgI9 secondary structural elements are labeled. The positioning of the structures is the same in each panel.

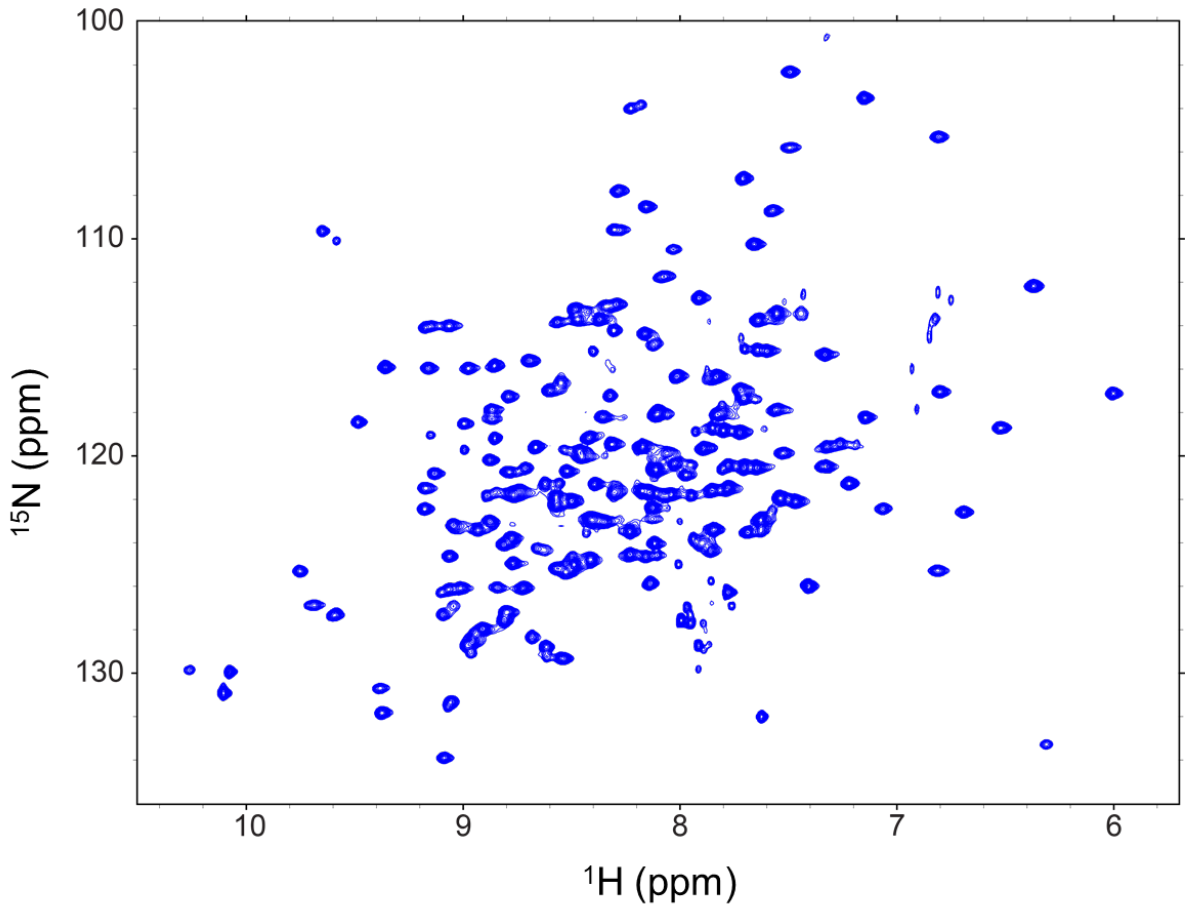


Figure S2.3 2-D NMR spectrum of RsgI9S1C-CTD.

A 2-D TROSY-HSQC NMR spectrum of ^{15}N-labeled RsgI9$^{S1C\text{-}CTD}$ was acquired prior to ^{15}N-TRACT correlation time determination. The spectrum indicates a well-folded protein with backbone amide signals well-dispersed.

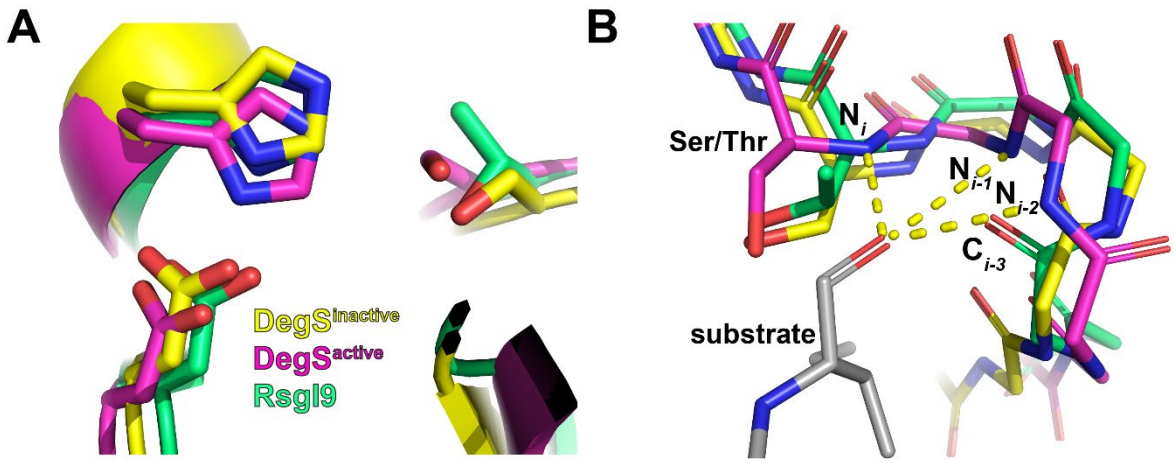


Figure S2.4 RsgI9's catalytic triad resembles inactive form of homolog DegS.

(A) Alignment of the catalytic triad residues found in RsgI9 with those in the active (magenta, PDB: 4RQZ) and inactive (yellow, PDB: 4RQY) forms of the *E. coli* DegS protease. RsgI9's configuration more closely resembles the inactive form, as the distance between the His Nε2 and the Ser/Thr hydroxyl group is too far to make a productive hydrogen bond. (B) Alignment of the oxyanion hole residues adjacent to the nucleophilic Ser/Thr, with the same structures and coloring as in panel A. Both RsgI9 and the inactive form of DegS presumably fail to stabilize the tetrahedral intermediate during proteolysis due to a deformation relative to the active conformation. Notably, the amide N_{i-2} is flipped out while the carbonyl C_{i-3} points inward.

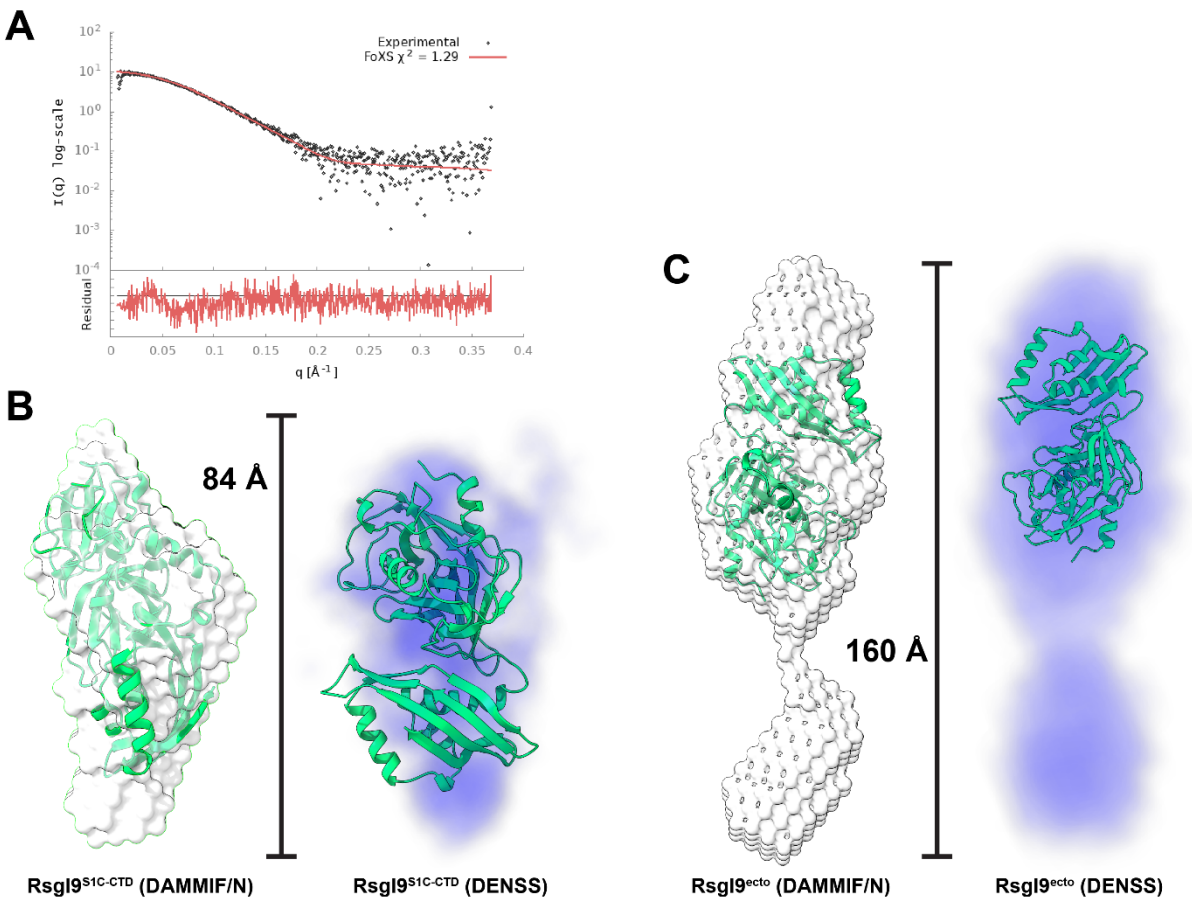


Figure S2.5 Fit of SAXS data and *ab initio* reconstructions.

(A) Fit of experimental scattering curve to that predicted for the RsgI9$^{S1C\text{-}CTD}$ crystal structure by the FoXS server (χ^2 value of 1.23). (B) *Ab initio* bead model and electron density reconstructions of RsgI9$^{S1C\text{-}CTD}$, using DAMMIF/N and DENSS, respectively. The maximum dimension of the model (84 Å) is consistent with the D_{max} observed from the data (82 Å). (C) *Ab initio* models for intact RsgI9^{ecto}. The length of the model is in agreement with the value of D_{max} derived from the SAXS data (160 Å).

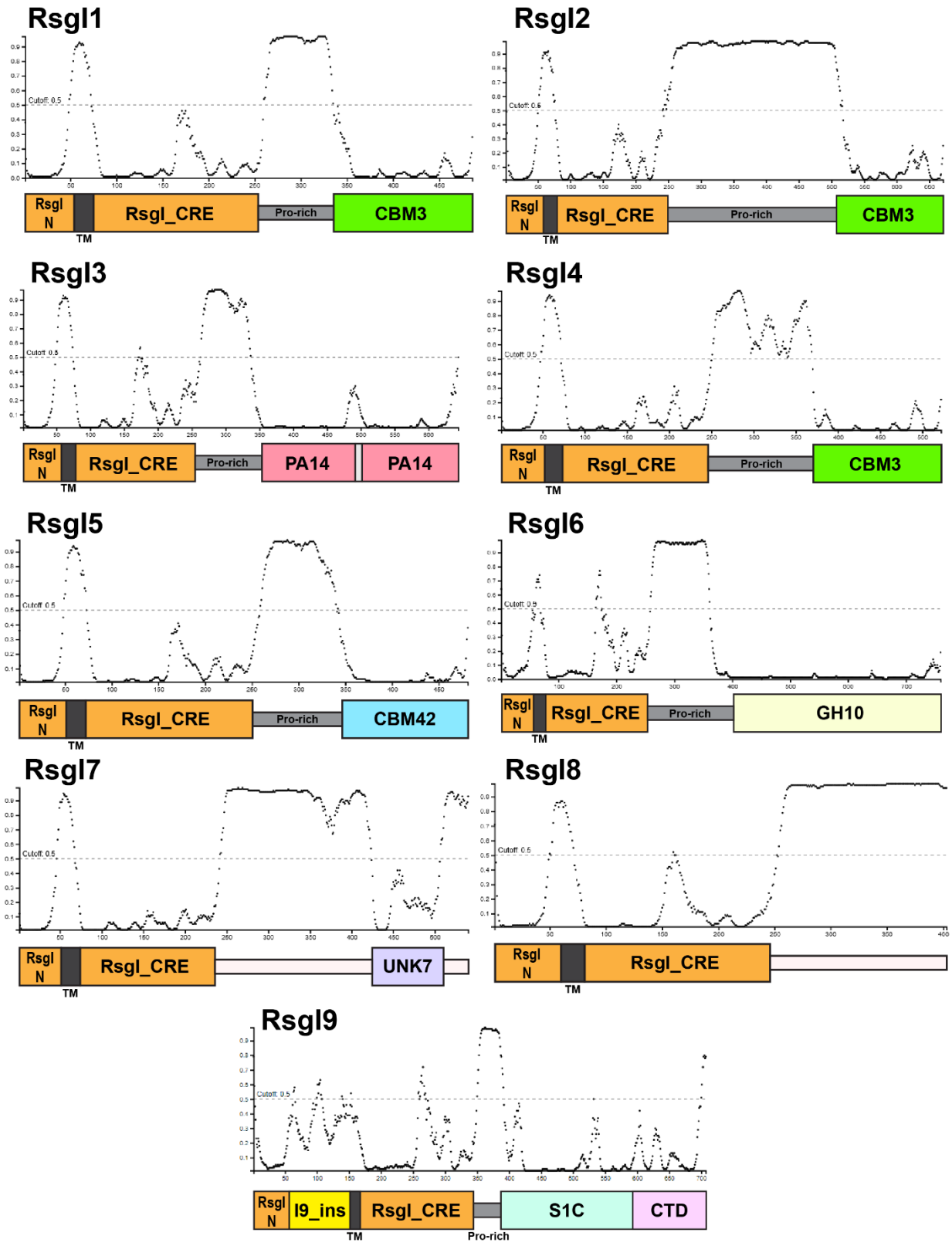


Figure S2.6 *C. thermocellum* RsgI architecture vs. DISOPRED3 disorder predictions.

The primary sequences of the nine RsgI proteins from *Clostridium thermocellum* were analyzed using DISOPRED3[53]. Values above 0.5 are predicted to be disordered, while those below are likely structured. The domain architecture of each RsgI is shown below each plot based on previous structures, UniProtKB annotation, and this work [20-22,74]. A good agreement is seen between annotated domains and predicted structured regions.

2.8 References

1. Watts N, Amann M, Ayeb-Karlsson S, et al. The Lancet Countdown on health and climate change: from 25 years of inaction to a global transformation for public health. *The Lancet.* 2018;391(10120):581-630.
2. Liu Y, Cruz-Morales P, Zargar A, et al. Biofuels for a sustainable future. *Cell.* 2021;184(6):1636-1647.
3. Sinitsyn AP, Sinitsyna OA. Bioconversion of Renewable Plant Biomass. Second-Generation Biofuels: Raw Materials, Biomass Pretreatment, Enzymes, Processes, and Cost Analysis. *Biochemistry (Moscow).* 2021;86(S1):S166-S195.
4. Viikari L, Vehmaanperä J, Koivula A. Lignocellulosic ethanol: From science to industry. *Biomass and Bioenergy.* 2012;46:13-24.
5. Olson DG, McBride JE, Joe Shaw A, Lynd LR. Recent progress in consolidated bioprocessing. *Current Opinion in Biotechnology.* 2012;23(3):396-405.
6. La Grange DC, Den Haan R, Van Zyl WH. Engineering cellulolytic ability into bioprocessing organisms. *Applied Microbiology and Biotechnology.* 2010;87(4):1195-1208.
7. Akinosho H, Yee K, Close D, Ragauskas A. The emergence of Clostridium thermocellum as a high utility candidate for consolidated bioprocessing applications. *Front Chem.* 2014;2:66.
8. Johnson EA, Sakajoh M, Halliwell G, Madia A, Demain AL. Saccharification of Complex Cellulosic Substrates by the Cellulase System from Clostridium thermocellum. *Appl Environ Microbiol.* 1982;43(5):1125-1132.
9. Lu Y, Zhang YH, Lynd LR. Enzyme-microbe synergy during cellulose hydrolysis by Clostridium thermocellum. *Proc Natl Acad Sci U S A.* 2006;103(44):16165-16169.

10. Alves VD, Fontes CMGA, Bule P. Cellulosomes: Highly Efficient Cellulolytic Complexes. In: Springer International Publishing; 2021:323-354.
11. Hu BB, Zhu MJ. Reconstitution of cellulosome: Research progress and its application in biorefinery. *Biotechnology and Applied Biochemistry.* 2019;66(5):720-730.
12. Jiang Y, Zhang X, Yuan H, et al. Research progress and the biotechnological applications of multienzyme complex. *Applied Microbiology and Biotechnology.* 2021;105(5):1759-1777.
13. Lamed R, Bayer EA. The Cellulosome of Clostridium thermocellum. In: Laskin AI, ed. *Advances in Applied Microbiology.* Vol 33. Academic Press; 1988:1-46.
14. Leibovitz E, Ohayon H, Gounon P, Beguin P. Characterization and subcellular localization of the Clostridium thermocellum scaffoldin dockerin binding protein SdbA. *J Bacteriol.* 1997;179(8):2519-2523.
15. Currie MA, Adams JJ, Faucher F, Bayer EA, Jia Z, Smith SP. Scaffoldin Conformation and Dynamics Revealed by a Ternary Complex from the Clostridium thermocellum Cellulosome. *Journal of Biological Chemistry.* 2012;287(32):26953-26961.
16. Yoav S, Barak Y, Shamshoum M, et al. How does cellulosome composition influence deconstruction of lignocellulosic substrates in Clostridium (Ruminiclostridium) thermocellum DSM 1313? *Biotechnology for Biofuels.* 2017;10(1).
17. Nataf Y, Bahari L, Kahel-Raifer H, et al. Clostridium thermocellum cellulosomal genes are regulated by extracytoplasmic polysaccharides via alternative sigma factors. *Proceedings of the National Academy of Sciences.* 2010;107(43):18646-18651.
18. Kahel-Raifer H, Jindou S, Bahari L, et al. The unique set of putative membrane-associated anti-σ factors in Clostridium thermocellum suggests a novel extracellular carbohydrate-sensing mechanism involved in gene regulation. *FEMS Microbiology Letters.* 2010;308(1):84-93.

19. Wei Z, Chen C, Liu Y-J, et al. Alternative σI/anti-σI factors represent a unique form of bacterial σ/anti-σ complex. *Nucleic Acids Research.* 2019;47(11):5988-5997.

20. Grinberg IR, Yaniv O, Ora LO, et al. Distinctive ligand-binding specificities of tandem PA14 biomass-sensory elements from Clostridium thermocellum and Clostridium clariflavum. *Proteins: Structure, Function, and Bioinformatics.* 2019;87(11):917-930.

21. Bahari L, Gilad Y, Borovok I, et al. Glycoside hydrolases as components of putative carbohydrate biosensor proteins in Clostridium thermocellum. *Journal of Industrial Microbiology & Biotechnology.* 2011;38(7):825-832.

22. Yaniv O, Fichman G, Borovok I, et al. Fine-structural variance of family 3 carbohydrate-binding modules as extracellular biomass-sensing components of Clostridium thermocellum anti-sigmaI factors. *Acta Crystallogr D Biol Crystallogr.* 2014;70(Pt 2):522-534.

23. Friedhoff P, Gimadutdinow O, Ruter T, et al. A procedure for renaturation and purification of the extracellular Serratia marcescens nuclease from genetically engineered Escherichia coli. *Protein Expr Purif.* 1994;5(1):37-43.

24. Kabsch W. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallographica Section D Biological Crystallography.* 2010;66(2):133-144.

25. Weichenberger CX, Rupp B. Ten years of probabilistic estimates of biocrystal solvent content: new insights via nonparametric kernel density estimate. *Acta Crystallogr D Biol Crystallogr.* 2014;70(Pt 6):1579-1588.

26. Kantardjieff KA, Rupp B. Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Science.* 2003;12(9):1865-1871.

27. Matthews BW. Solvent content of protein crystals. *J Mol Biol.* 1968;33(2):491-497.

28. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D Biological Crystallography.* 2004;60(12):2256-2268.

29. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. *Journal of Applied Crystallography.* 2007;40(4):658-674.

30. Liebschner D, Afonine PV, Baker ML, et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallographica Section D Structural Biology.* 2019;75(10):861-877.

31. Dong W, Wang J, Niu G, Zhao S, Liu L. Crystal structure of the zinc-bound HhoA protease from Synechocystis sp. PCC 6803. *FEBS Letters.* 2016;590(19):3435-3442.

32. Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, et al. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallographica Section D Biological Crystallography.* 2008;64(1):61-69.

33. Smart OS, Womack TO, Flensburg C, et al. Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. *Acta Crystallographica Section D Biological Crystallography.* 2012;68(4):368-380.

34. Painter J, Merritt EA. Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr D Biol Crystallogr.* 2006;62(Pt 4):439-450.

35. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. *Acta Crystallographica Section D Biological Crystallography.* 2010;66(4):486-501.

36. Bricogne G, Blanc E, Brandl M, et al. BUSTER version 2.10.4. *Global Phasing Ltd, Cambridge, UK.* 2021.

37. Marley J, Lu M, Bracken C. A method for efficient isotopic labeling of recombinant proteins. *Journal of Biomolecular NMR.* 2001;20(1):71-75.

38. Lee D, Hilty C, Wider G, Wüthrich K. Effective rotational correlation times of proteins from NMR relaxation interference. *Journal of Magnetic Resonance.* 2006;178(1):72-76.

39. Robson SA, Dağ Ç, Wu H, Ziarek JJ. TRACT revisited: an algebraic solution for determining overall rotational correlation times from cross-correlated relaxation rates. *Journal of Biomolecular NMR.* 2021.

40. Classen S, Hura GL, Holton JM, et al. Implementation and performance of SIBYLS: a dual endstation small-angle X-ray scattering and macromolecular crystallography beamline at the Advanced Light Source. *Journal of Applied Crystallography.* 2013;46(1):1-13.

41. Hopkins JB, Gillilan RE, Skou S. BioXTAS RAW: improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. *Journal of Applied Crystallography.* 2017;50(5):1545-1553.

42. Manalastas-Cantos K, Konarev PV, Hajizadeh NR, et al. ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. *Journal of Applied Crystallography.* 2021;54(1):343-355.

43. Hajizadeh NR, Franke D, Jeffries CM, Svergun DI. Consensus Bayesian assessment of protein molecular mass from solution X-ray scattering data. *Scientific Reports.* 2018;8(1).

44. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Research.* 2016;44(W1):W424-W429.

45. Schneidman-Duhovny D, Hammel M, John, Sali A. Accurate SAXS Profile Computation and its Assessment by Contrast Variation Experiments. *Biophysical Journal.* 2013;105(4):962-974.

46. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics.* 2016;54(1).

47. Svergun DI. Restoring Low Resolution Structure of Biological Macromolecules from Solution Scattering Using Simulated Annealing. *Biophysical Journal.* 1999;76(6):2879-2886.

48. Franke D, Svergun DI. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *Journal of Applied Crystallography.* 2009;42(2):342-346.

49. Grant TD. Ab initio electron density determination directly from solution scattering data. *Nature Methods.* 2018;15(3):191-193.

50. Pettersen EF, Goddard TD, Huang CC, et al. UCSF ChimeraX : Structure visualization for researchers, educators, and developers. *Protein Science.* 2021;30(1):70-82.

51. Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021. *Nucleic Acids Research.* 2021;49(D1):D412-D419.

52. Ding X, Chen C, Cui Q, Li W, Feng Y. Resonance assignments of the periplasmic domain of a cellulose-sensing trans-membrane anti-sigma factor from Clostridium thermocellum. *Biomolecular NMR Assignments.* 2015;9(2):321-324.

53. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics.* 2015;31(6):857-863.

54. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. *Nucleic Acids Research.* 2018;46(W1):W200-W204.

55. Krissinel E, Henrick K. Inference of Macromolecular Assemblies from Crystalline State. *Journal of Molecular Biology.* 2007;372(3):774-797.

56. Krojer T, Garrido-Franco M, Huber R, Ehrmann M, Clausen T. Crystal structure of DegP (HtrA) reveals a new protease-chaperone machine. *Nature.* 2002;416(6879):455-459.

57. Holm L. Using Dali for Protein Structure Comparison. In: Springer US; 2020:29-42.

58. Bullock TL, Clarkson WD, Kent HM, Stewart M. The 1.6 angstroms resolution crystal structure of nuclear transport factor 2 (NTF2). *J Mol Biol.* 1996;260(3):422-431.

59. Eberhardt RY, Chang Y, Bateman A, et al. Filling out the structural map of the NTF2-like superfamily. *BMC Bioinformatics.* 2013;14(1):327.

60. Perona JJ, Craik CS. Evolutionary Divergence of Substrate Specificity within the Chymotrypsin-like Serine Protease Fold. *Journal of Biological Chemistry.* 1997;272(48):29987-29990.

61. Cavanagh J. *Protein NMR spectroscopy : principles and practice.* 2nd ed. Amsterdam ; Boston: Academic Press; 2007.

62. Barbato G, Ikura M, Kay LE, Pastor RW, Bax A. Backbone dynamics of calmodulin studied by nitrogen-15 relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible. *Biochemistry.* 1992;31(23):5269-5278.

63. Hudson KL, Bartlett GJ, Diehl RC, et al. Carbohydrate–Aromatic Interactions in Proteins. *Journal of the American Chemical Society.* 2015;137(48):15152-15160.

64. Tormo J, Lamed R, Chirino AJ, et al. Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose. *EMBO J.* 1996;15(21):5739-5751.

65. Garron ML, Cygler M. Structural and mechanistic classification of uronic acid-containing polysaccharide lyases. *Glycobiology.* 2010;20(12):1547-1573.

66. Yang L, Connaris H, Potter JA, Taylor GL. Structural characterization of the carbohydrate-binding module of NanA sialidase, a pneumococcal virulence factor. *BMC Structural Biology.* 2015;15(1).

67. Ding N, Zhao B, Ban X, et al. Carbohydrate-Binding Module and Linker Allow Cold Adaptation and Salt Tolerance of Maltopentaose-Forming Amylase From Marine Bacterium Saccharophagus degradans 2-40 (T). *Front Microbiol.* 2021;12:708480.

68. Alba BM. DegS and YaeL participate sequentially in the cleavage of RseA to activate the sigma E-dependent extracytoplasmic stress response. *Genes & Development.* 2002;16(16):2156-2168.

69. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research.* 2018;46(D1):D624-D632.

70. Sohn J, Grant RA, Sauer RT. Allosteric Activation of DegS, a Stress Sensor PDZ Protease. *Cell.* 2007;131(3):572-583.

71. Durand D, Vives C, Cannella D, et al. NADPH oxidase activator p67(phox) behaves in solution as a multidomain protein with semi-flexible linkers. *J Struct Biol.* 2010;169(1):45-53.

72. Svergun DI, Koch MHJ. Small-angle scattering studies of biological macromolecules in solution. *Reports on Progress in Physics.* 2003;66(10):1735-1782.

73. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology.* 2011;7(1):539.

74. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. In: Humana Press; 2007:89-112.

Chapter 3

Insight into the autoproteolysis mechanism of the Rsgl9 anti-σ factor from *Clostridium thermocellum*

The work described in this chapter has been reproduced with permission from:

Takayesu A, Mahoney BJ, Goring AK, Jessup T, Ogorzalek Loo RR, Loo JA, Clubb RT. Insight into the autoproteolysis mechanism of the RsgI9 anti-σ factor from Clostridium thermocellum. Proteins. 2024 Aug;92(8):946-958. doi: 10.1002/prot.26690.

https://onlinelibrary.wiley.com/doi/10.1002/prot.26326

I was the primary author of this work. I designed and conducted experiments, analyzed data, assembled literature references, made figures and wrote the manuscript.

3.1 Overview

Clostridium thermocellum is a potential microbial platform to convert abundant plant biomass to biofuels and other renewable chemicals. It efficiently degrades lignocellulosic biomass using a surface displayed cellulosome, a megadalton sized multienzyme containing complex. The enzymatic composition and architecture of the cellulosome is controlled by several transmembrane biomass-sensing RsgI-type anti-σ factors. Recent studies suggest that these factors transduce signals from the cell surface via a Conserved RsgI Extracellular (CRE) domain (also called a periplasmic domain) that undergoes autoproteolysis through an incompletely understood mechanism. Here we report the structure of the autoproteolyzed CRE domain from the *C. thermocellum* RsgI9 anti-σ factor, revealing that the cleaved fragments forming this domain associate to form a stable α/β/α sandwich fold. Based on AlphaFold2 modeling, molecular dynamics simulations and tandem mass spectrometry, we propose that a conserved Asn-Pro bond in RsgI9 autoproteolyzes via a succinimide intermediate whose formation is promoted by a conserved hydrogen bond network holding the scissile peptide bond in a strained conformation. As other RsgI anti-σ factors share sequence homology to RsgI9, they likely autoproteolyze through a similar mechanism.

3.2 Introduction

The effects of climate change and finite petroleum supplies have created a pressing need to cost-effectively produce biofuels, chemicals, and materials from renewable and carbon-neutral sources [1]. Lignocellulosic plant biomass (LCB) is a promising feedstock to generate these biomaterials as it is the largest source of carbon in the biosphere (~450 gigatons) and can be produced sustainably [2]. Currently, the use of LCB as a feedstock is limited by its recalcitrance to degradation [3]. One approach to overcome this problem is to employ *Clostridium thermocellum* (also known as *Acetivibrio thermocellus* or *Hungateiclostridium thermocellum*) to break down biomass as this thermophilic obligate anaerobe possesses extremely potent cellulolytic activity [4, 5]. *C. thermocellum* has evolved the capacity to degrade LCB by displaying cellulosomes - massive multienzyme complexes that efficiently hydrolyze cellulose into its component sugars [6-10]. The high cellulolytic activity of the cellulosome originates from its modular architecture, which is capable of colocalizing a large a number of glycoside hydrolases (GHs) that work together to degrade cellulose and hemicellulose fibers within LCB [11]. In addition to *C. thermocellum,* a number of mesophilic and thermophilic microbial species have been shown to produce cellulosomes that confer potent cellulolytic activity [12, 13].

To degrade different types of biomass, *C. thermocellum* regulates the transcription of ~50 genes encoding GH enzymes that are incorporated into the cellulosome [14, 15]. Their expression is regulated by extracytoplasmic function (ECF) σ factors, which together with their cognate anti-σ factors are responsible for altering GH gene expression in response to the presence of different types of extracellular polysaccharides [16]. A total of nine ECF σ factors may regulate gene expression in response to different types of polysaccharides. These include eight SigI-type σ factors (σ_1 to σ_8) whose activities are modulated by at least eight RsgI-type anti-σ factors (RsgI1-8) that are co-transcribed from the same operon, and σ_{24} which is

regulated by the Rsi24c anti-σ factor [16-19]. In the signaling process for the SigI-type σ factors, each of the membrane-bound anti-σ factors are believed to bind to different types of extracellular polysaccharides [17, 20-22]. They then transduce this signal across the membrane, triggering the release of their cognate SigI-type σ factor from the cytoplasmic membrane [23]. The SigI-type σ factor then binds to the RNA polymerase complex enabling it to transcribe specific GH encoding genes, as well as genes that encode scaffoldin proteins used to construct the cellulosome [24]. Each of the RsgI anti-σ factors contains a conserved cytoplasmic N-terminal domain (NTD) that is located in the cytoplasm, where it binds to its cognate σ factor. This is followed by a single membrane embedded helix and an extracellularly located Conserved RsgI Extracellular domain (CRE domain). The CRE domain has also been referred to as either the periplasmic juxtaposition domain or SEAL domain [25, 26]. Following this segment is a non-conserved C-terminal region that frequently mediates polysaccharide binding. Biochemical studies have deduced the binding activities of several of the RsgI anti-σ factors. For example, RsgI1, RsgI2 and RsgI4 contain family 3 type carbohydrate binding modules (CBM3) known to interact with cellulose [17], RsgI3 recognizes pectin via protective antigen 14 domains, RsgI5 contains a family-10 CBM that binds to arabinose, and RsgI6 contains a family-10 GH module that degrades cellulose and xylan [17, 20, 21]. Recent studies of the RsgI-like factors from *Bacillus subtilis* and *C. thermocellum* suggest that signaling occurs via a regulated intramembrane proteolysis (RIP) mechanism in which the anti-σ factor undergoes proteolytic cleavage at two sites, called site-1 and site-2 [25-27]. Site-1 cleavage occurs within their CRE domains as a result of autoproteolysis, predisposing the anti-σ factor for signal-induced structural changes that expose site-2 for degradation by the RasP protease in *B. subtilis* or RseP in *C. thermocellum*. The second cleavage event releases the cognate σ factor from the RsgI's cytoplasmic NTD enabling it to selectively transcribe specific genes by targeting the RNA polymerase complex to their promoters.

Recently we reported the structure of the cellulose-binding ectodomain from the *C. thermocellum* RsgI9. It is an ‘orphan’ anti-σ factor [22] because unlike other RsgI factors, the gene encoding RsgI9 is not located in an operon that also contains a gene encoding a σ factor. The ectodomain in RsgI9 adopts an elongated conformation in which a C-terminal bi-domain unit is likely projected from the cell surface to engage cellulose, while its conserved CRE domain resides near the membrane to presumably affect signal transduction. Using a combination of NMR spectroscopy, AlphaFold2, molecular dynamics (MD) simulations, and mass spectrometry we show that RsgI9 exists in a pre-cleaved state on the extracellular membrane that may enable it to transduce signals caused by carbohydrate binding into gene expression changes that alter the composition of the cellulosome, and we provide evidence that is compatible with CRE domain autoproteolyzing via a conserved mechanism involving a succinimide intermediate.

3.3 Results and Discussion

3.3.1 RsgI9's CRE domain undergoes autoproteolysis.

CRE domains are conserved in a range of membrane embedded anti-σ factors and have recently been implicated in RIP-mediated signal transduction [23]. To gain insight into its function, we recombinantly produced a polypeptide encoding RsgI9's CRE domain (CRE, residues G167-K343 of RsgI9) (**Fig. 1A**). The final step of the CRE purification process involved the application of size exclusion chromatography (SEC). Interestingly, we noticed that even though the purified protein eluted from the column at a position consistent with its predicted molecular weight (~20 kDa, 177 amino acids) (**Fig. 1B**), based on SDS-PAGE analysis the material within this peak contains two polypeptide chains with masses of ~2.4 and ~17.6 kDa as confirmed by MALDI-TOF (**Fig. 1C**). To confirm the identity of the components, the two fragments were separated, digested with trypsin, subjected to reversed-phase HPLC and then analyzed via electrospray ionization tandem mass spectrometry (ESI-MS/MS) (**Fig. 1D**). This procedure revealed that the fragments corresponded to polypeptides containing residues G167-N188 and P189-K343 in CRE, which was consistent with the full CRE protein undergoing proteolysis at its N188-P189 peptide bond. Close analysis of SDS-PAGE separated proteins in *E. coli* cells overexpressing CRE did not reveal the presence of the intact polypeptide, suggesting that autoproteolysis occurs rapidly after the protein is translated.

3.3.2 NMR structure of the autoproteolyzed CRE domain.

Co-elution of the 2.4 and 17.6 kDa CRE cleavage fragments in SEC experiments suggests that they associate with one another. To investigate the nature of this interaction, a ^{13}C- and ^{15}N-labeled sample of CRE was produced and its chemical shifts were assigned using triple resonance NMR methods (**Fig. 2A**) [28]. CRE adopts a folded structure as evidenced by its well dispersed ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectrum. Notably,

both fragments within the cleaved CRE domain are structurally ordered. This is evident from the heteronuclear $^{15}N\{^{1}H\}$ nuclear Overhauser (hetNOE) data, since residues spanning both fragments within the protein exhibit values >0.6 (**Fig. 2B**). It is further substantiated by a plot of the random coil index (RCI) values predicted by TALOS-N, which reveals that residues in both the short and long fragments are structured based on their backbone chemical shifts (residues G177-I341) (**Fig. 2C**) [29].

The solution structure of the autoproteolyzed CRE domain was determined using multidimensional heteronuclear NMR data and simulated annealing approaches [30, 31]. The structure is represented by an ensemble of 20 conformers that are compatible with 2,044 structural restraints (1,668 NOE distances, 301 dihedral angles, 74 hydrogen bonds) (**Table 1)**.

Consistent with the hetNOE data (**Fig. 2B**), residues Y180-I337 are structurally ordered in the ensemble; their backbone atoms can be superimposed to the mean structure with a root mean square deviation (RMSD) of 0.57 Å (**Fig. 3A**). CRE adopts an α/β/α sandwich that is constructed from three subdomains: helices α1-3, strands β1-5 that form a central sheet, and helices α4-7 (**Fig. 3C**). The β-strands are arranged in an anti-parallel manner, with the exception of strands β4 and β5 that are parallel to one another and located at the edge of the sheet. The shorter CRE fragment (residues G167-N188) comprises strand β1 that pairs in an anti-parallel manner with strands β2 and β4 located in the longer fragment. The cleavage site residues lie within the cleaved hairpin loop between the β1 and β2 strands and are partially excluded from the solvent by the side chains of residues located within α1 (D210) and α6 (K336), as well as the side chain of Y241 located in the loop region connecting α1 and β4 (**Fig. 3B**, yellow).

The structure is similar to recently reported structures of CRE domains from the *B. subtilis* RsgI anti-σ factor, as well as the RsgI1, RsgI2 and RsgI6 proteins from *C. thermocellum* (**Fig. 4A**) [27]. These proteins share 32-41% sequence identity with RsgI9 and also undergo autoproteolysis of an Asn-Pro peptide bond (**Fig. 4B**). The structure of RsgI9 is most similar to

RsgI2, as their backbone atoms can be superimposed with an RMSD of 1.95 Å. A comparison of RsgI9 to the different CRE domain structures reveals the presence of many conserved residues that are positioned near the cleaved peptide bond (**Fig. 4B**) [26, 27]. These include amino acids within the conserved D[I/V]NPS sequence that harbor the cleaved Asn-Pro peptide bond and several residues from the longer fragment that interact with these amino acids. For example, semi-conserved polar contacts originating from the side chains N208 and D210 to the cleaved proline residue are observed, as well as contacts from L250, S310, and G312 to residues that surround the broken peptide bond (**Fig. 3D, 4B**). This conservation suggests that all CRE domains may autoproteolyze, consistent with experimental studies of RsgI from *B. subtilis* (^{Bsub}RsgI) and select RsgI proteins from *C. thermocellum* (^{Cthe}RsgI1-4, 6). A notable exception is the RsgI8 CRE domain, which replaces the Asn-Pro sequence with Asp-Ala and may not undergo autoproteolysis [26, 27].

3.3.3 A conserved hydrogen-bond network may impart strain onto the Asn-Pro peptide bond to promote its cleavage.

To gain insight into the mechanism of cleavage, we used AlphaFold2 to model the structure of an intact CRE domain that exists prior to autoproteolysis (**Fig. 5A**) [34]. The top 5 models describing the structure were accurately predicted based on their pLDDT scores of 91.8 – 93.4. The intact (AlphaFold2) and cleaved (NMR) structures are closely related, as their backbone atoms can be superimposed with an RMSD of 1.94 Å. In general, the structural differences originate from alterations in the conformations of protein loops and the presence of an extra helical turn located at the end of the α7 helix in the intact protein. However, several features in the intact protein suggest that the region containing the Asn-Pro peptide bond adopts a stressed conformation that may promote its cleavage. For an example, an analysis using the program MolProbity reveals that N188-P189 ω dihedral angle describing the geometry of the scissile peptide bond is 83.2°, instead of the canonical value of ±180° for an unstrained, trans-

peptide bond [35]. The φ and ψ torsion angles for P189 in this region also reside in the disfavored region of the Ramachandran plot (-107.3, 174.1). Finally, a CaBLAM analysis reveals several geometric outliers involving C^{α} atoms for residues V187-P189 suggesting that they adopt an energetically unfavorable conformation (**Fig. 5B**) [36]. In marked contrast, the NMR structure of the cleaved polypeptide does not exhibit this conformational stress; presumably because of the additional conformational freedom that is afforded by rupture of the N188 and P189 bond.

Molecular dynamics (MD) simulations were performed to investigate the origin of conformational strain in the CRE domain. Motions in the solvated model of the intact protein were simulated for 200 ns using the CHARMM36 all-atom force field [37]. The coordinates in the simulation stabilized after ~50 ns and thereafter exhibited only small root mean square fluctuations (RMSFs). Prominently, residues within and surrounding the β1-β2 hairpin containing the Asn-Pro peptide bond exhibited only relatively small conformational fluctuations (~0.75 Å RMSF values) and remained strained through the entirety of the simulation. In particular, the ω dihedral angle for the Asn-Pro peptide exhibited a non-ideal average value of 148.1° ± 8.4° (instead of ~±180°) with even larger transient excursions that move it as far as 55° away from planarity. In contrast, for all other residues in the protein their ω dihedral angles retain values around ±176° and thus their peptide bonds remain in a lower energy, planar configuration. Residues surrounding the scissile bond also remained in their original, unfavorable conformations. An inspection of the trajectory suggests that an extensive hydrogen bond network may hold the Asn-Pro peptide bond in a strained conformation (**Fig. 6B**). As part of this network, we also observed a long-lived water molecule that for the majority of the trajectory maintained hydrogen bonding interactions between residues D186, N188, and S190. Within the run time of the simulation, the water formed at least one hydrogen bond contact with the aforementioned residues for 94% of the simulation. In particular, the water molecule appears to

form hydrogen bonds with both the main chain amide and carbonyl of N188, which had a site occupancy of 45% during the remaining 150 ns runtime after the system had stabilized Additionally, N208 and R313 form hydrogen bond contacts with both the amide side chain and main chain carbonyl of N188. Although longer time-scale simulations are needed to fully assess the stability of this region, this network may hold the Asn-Pro peptide bond in its strained conformation. This would raise the energy of the uncleaved substrate, effectively reducing the activation energy barrier required for proteolysis. The functional importance of the network is consistent with studies of the ^{Cthe}RsgI2 CRE domain since mutation of the equivalent residues in this protein impair its autoproteolysis, specifically D186, S190, N208, and R313 [27].

3.3.4 Spontaneous cleavage may occur via a succinimide intermediate.

In principle, cleavage of the twisted scissile bond of CRE could occur via at least three distinct mechanisms: (i) direct hydrolysis of the strained peptide bond, (ii) an N→O / N→S acyl rearrangement mechanism that is caused by a nucleophilic action of a proximal residue such as a hydroxyl bearing serine residue [38, 39], or (iii) via a succinimide intermediate mechanism whereby the Asn side chain (N188) with the Asn-Pro scissile peptide bond undergoes cyclization and cleavage via the formation of a succinimide intermediate (Scheme I) [40].

Proteolysis via direct hydrolysis (i) or an acyl rearrangement (ii) seems unlikely. Direct hydrolysis seems unlikely for several reasons. First, in proteins that hydrolyze peptide bonds the water molecule is typically activated for nucleophilic attack on the scissile bond by a nearby metal or general base [41]. In the structure of RsgI9 and other CRE proteins, no metal is present near the labile peptide bond. There are numerous conserved residues near the labile bond that could potentially function to activate a water molecule for hydrolysis (e.g. S190, N208, and R313 in RsgI9). However, results obtained by the Feng and Rudner groups have shown that variants which alter these residues exhibit only mildly reduced autoproteolytic activity, suggesting that they are unlikely to function as a general base [25-27]. Second, the MD

simulation of the uncleaved AlphaFold2 CRE model revealed the presence of a long-lived water molecule that stabilizes the hydrogen-bond network and could mediate hydrolysis, but is positioned too far away from scissile peptide bond making it a poor candidate for a nucleophile [42]. However, we cannot exclude the possibility that the 200 ns simulation may have been of insufficient duration to capture the presence of a long-lived water molecule that could mediate hydrolysis. Third, the cleavage product containing C-terminal Asn residue exhibits an isoform that is consistent with a mechanism involving a succinimide intermediate (see below). The N→O / N→S acyl rearrangement mechanism (ii) also seems unlikely to be the cause of CRE domain autoproteolysis. This is the most commonly observed mechanism of protein autoproteolysis and involves an acyl rearrangement mediated by a nearby side chain (e.g. Thr, Ser, Cys, Asn) that acts as a nucleophile that attacks the carbonyl group within the scissile peptide bond [43]. Proteins that utilize this mechanism include self-cleaving proteins such as the FoxR anti-σ factor [44], GPCR autoproteolysis inducing (GAIN) domains [45], and several viral polyproteins including those from SARS-CoV-2 [46]. This acyl rearrangement is unlikely for CRE as mutagenesis of highly conserved residues near the peptide bond that could serve as potential nucleophiles in the ^{Cthe}RsgI2 protein (N208, D210, and R313) does not significantly ablate its autoproteolysis [27]. Indeed, prior work on ^{Cthe}RsgI2 demonstrated that double mutants in which more than one potential nucleophile is altered do not completely disrupt autoproteolysis, making this mechanism unlikely. Similarly, mutations of residues proximal to the autocleavage (Y182, D186, E192, V201, N208) site in the CRE of ^{Bsub}RsgI made by the Rudner group did not completely prevent autocleavage but did lead to a similar phenotype as the Δ*rsgI* mutant [25]. According to the solution structure and MD simulation, these residues are either involved within the hydrogen bond network surrounding the autocleavage site or within the core of the domain. Thus, mutations to these core residues would likely destabilize the domain such that the RsgI protein would be more susceptible to site-2 processing and activation.

We posit that the non-enzymatic cleavage of the Asn-Pro bonds in CRE domains occurs via a succinimide mechanism shown in Scheme 1. This autoproteolysis mechanism has been observed in the structurally unrelated αA-crystallin [40], aquaporin 0, FlhB and SpaS proteins [47-50]. Bond cleavage starts with the cyclization of the Asn side chain when its δ-nitrogen atom attacks its backbone carbonyl group to form a *gem*-hydroxylamine transition state that subsequently collapses into a C-terminal succinimide intermediate when the Asn-Pro bond breaks [51]. The C-terminal succinimide containing the Asn is then hydrolyzed to create a racemic mixture of cleavage product that contains either a C-terminal Asn or Asp-amide/*iso*Asn residue (Scheme 1). This reaction is typically rare in proteins as it competes with much faster Asn deamidation reaction that involves Asn cyclization [52]. However, Asn-Pro peptide bonds are prone to autoproteolysis since this competing deamidation reaction is disfavored because the proline backbone imine is a poor nucleophile [53]. If cleavage in RsgI9 occurs via a succinimide intermediate, the short peptide cleavage fragment should contain both Asn and *iso*Asn at its C-terminus (Scheme 1). LC-MS/MS was performed on a sample of purified CRE domain that had been digested with trypsin. This procedure identified 14 unique peptides corresponding to the larger fragment and 1 peptide corresponding to the shorter fragment (85.2% and 90.9% sequence coverage, respectively). Interestingly, the extracted ion chromatogram (XIC) revealed that all the tryptic peptides eluted with a standard unimodal elution profile with the notable exception of a shorter fragment containing the C-terminal Asn (or *iso*Asn) residue which exhibited a bimodal profile (**Fig. 7A**). The identity of the bimodal peptide was confirmed by LC-MS/MS as spanning residues G169-N188 (**Fig. 7B**). Notably, MS analysis of eluate from the bimodal chromatographic peak revealed that both lobes corresponded to peptides of *m/z* 1081.49. This behavior is consistent with the elution of a C-terminal Asn/*iso*Asn peptide pair. To further investigate this issue, we performed parallel reaction monitoring (PRM) on these tryptic digest products. Consistent with assignment as the autoproteolytic *iso*Asn and Asn products, both lobes of the LC peak exhibited nearly identical MS/MS spectra. This

behavior is expected given that their backbone amides are similarly susceptible to collisional fragmentation (**Fig. 7C**). It is also notable that the single long-lived water molecule, observed to stabilize the hydrogen bond network in MD simulations, is positioned to facilitate cleaving the succinimide intermediate. Poised to hydrogen bond to the main chain carbonyl of N188 (heavy atom separation 3.0 Å), the water molecule can donate a proton as the peptide bond is broken.

3.4 Conclusion

In this work, we demonstrate that the RsgI9 CRE domain undergoes autoproteolysis at a conserved Asn-Pro sequence to form a stably folded structure in which the proteolytic fragments remain associated. Coupled with prior findings that show RsgI9's C-terminal ectodomain binds cellulose, these results are consistent with it regulating the composition of the cellulosome via a RIP signal transduction mechanism [22]. In this process, polysaccharide binding to the RsgI9 domain would cause the pre-cleaved CRE domain to separate, thereby exposing it to degradation by the RasP intramembrane metalloprotease to release a yet to be identified σ factor that controls gene transcription. This work builds upon previous studies of the CRE domain by providing insight into the potential mechanism for autoproteolysis. Additionally, we have confirmed that the CRE of the orphan RsgI9 undergoes autoproteolysis, implicating its involvement in extracellular sensing and transcription regulation despite its unknown cognate σ factor. Future studies will be focused on identifying this σ factor and testing whether polysaccharide binding triggers CRE domain dissociation. Our NMR, MD and MS/MS analyses in combination with recently reported mutagenesis data provide insight into how a conserved Asn-Pro peptide bond in CRE domains is rapidly autoproteolyzed. We propose that a conserved hydrogen bond network and bound water molecule promote cleavage by maintaining the scissile peptide bond in a strained conformation that reduces the energy between the pre-cleaved substrate and transition state associated with bond breakage. This is similar to the hydrogen bond-maintained conformational strain seen in SEA domains, which have recently

been proposed Brogan et al. to function similarly to CRE [26]. There are over 1000+ CRE homologs that contain the conserved D[I/V]NPS cleavage sequence across many species of gram-positive bacteria, suggesting that this mechanism is widely employed by RsgI-type anti-σ factors to mediate signal transduction.

3.5 Materials and Methods

3.5.1 Cloning and expression of CRE

The nucleotide sequence for *C. thermocellum* DSM 1313 CRE (residues G167-K343) was cloned into a pET-29b expression plasmid which contains an N-terminal 6xHis affinity tag and followed by a Tobacco Etch Virus (TEV) protease cleavage site. The plasmid was transformed into *E. coli* BL21 (DE3) cells and cultured at 37 °C in LB media containing 50 µg/mL kanamycin until an OD_{600} of 0.6 – 0.8 was reached. Cells were then induced with 1mM isopropyl β-D-thiogalactoside (IPTG) and incubated with agitation at 17 °C for 12–17 hrs. Cell cultures were pelleted at 7k rpm at 4 °C for 25 minutes and were resuspended in lysis buffer containing 50 mM Tris-HCl and 300 mM NaCl at pH 8.0. Cells were lysed via sonication in the presence of 1 mg of lysozyme isolated from egg white and 2mM phenylmethanesulfonyl fluoride (PMSF) per liter of culture. Lysates were clarified by centrifugation at 15k rpm at 4 °C for 50 minutes. Clarified lysate was loaded onto HisPur™ Co^{2+}-NTA resin previously equilibrated in lysis buffer and incubated with agitation at 4 °C. The protein was washed with 10 CV of lysis buffer followed by a wash with lysis buffer containing 10mM imidazole and finally eluted using elution buffer (lysis buffer + 500 mM imidazole). Elutions were concentrated to 5 mL using an Amicon™ 10k MWCO Ultra centrifugal filter and then dialyzed in lysis buffer without imidazole in the presence of TEV protease overnight at 4 °C. The dialyzed and post-cleaved protein was then loaded back onto the HisPur™ Co^{2+}-NTA affinity resin and eluted with several washes of lysis buffer. Elutions were further purified using a Superdex 75 preparative grade column on an AKTA-FPLC system and concentrated again prior to storage and use. Samples from pre- and post-TEV digestion of RsgI9 CRE were separated by SDS-PAGE.

3.5.2 NMR spectroscopy and structure calculations

CRE was uniformly isotopically ^{13}C- and ^{15}N-labeled using the same expression and protein purification protocol as described above, but with expression induced after transferring to minimal M9 media supplemented with $^{15}NH_4Cl$ and [$^{13}C_6$] glucose. Protein samples for NMR were exchanged into NMR buffer (50mM sodium phosphate, 200 mM NaCl, 0.03% sodium azide, pH 6.0). NMR experiments were performed at 298 K on Bruker Avance III HD 600 MHz and Bruker Avance NEO 800 MHz spectrometers equipped with triple resonance cryogenic probes. Backbone and side chain atom assignments of CRE were determined via the following experiments: ^{15}N-HSQC, ^{13}C-HSQC, ^{15}N-TOCSY, HNCO, HN(CA)CO, HNCA, HN(CO)CA, HNCACB, HCCH-COSY, HCCH-TOCSY, HNHA, and HN(CO)CACB [28]. Data were processed using NMRPipe[54] and analyzed using XIPP NMR software [55]. Initial assignment of cross-peaks within the NOESY spectra were assigned automatically using UNIO [56] which were verified and added to manually in XIPP. $^{15}N\{^1H\}$ heteronuclear NOE data was collected in duplicate and analyzed in NMRFAM-Sparky [57]. Predicted secondary structure, RCI-S^2 values and dihedral angle restraints for structure calculations were generated using TALOS-N [58]. ^{15}N- and ^{13}C-edited NOESY spectra were acquired using a 120 ms mixing time and used to define NOE distance restraints for structural calculations made using XPLOR-NIH v3.6 [59, 60]. The final round of structural calculations included 200 structures generated, 40 of which had zero NOE violations greater than 0.5 Å or dihedral angle violations greater than 5°. A final ensemble of 20 structures based on lowest overall energy were validated via PROCHECK-NMR and selected to represent the structure of CRE, which were then deposited into the PDB (accession: 8U9O) [32]. Cartoon and ensemble representations of CRE and the AlphaFold2 model of RsgI9$^{177-350}$ were generated using MOLMOL and PyMOL [61, 62]. All backbone alignment comparisons (atoms N, Cα, C', and O) with RsgI9 CRE were done using the 'align' function in PyMOL with no pruning [62].

3.5.3 AlphaFold2 Modeling and MD Simulations

A full-length, uncleaved model of RsgI9 CRE$^{177-350}$ was generated using AlphaFold2 [34] (locally installed). 5 models were generated in structure determination. The highest ranked structure based on pLDDT score was used for further MD simulations.

GROMACS version 2021 was used for explicit-solvent molecular dynamics simulations of the uncleaved AlphaFold2 model of RsgI9 CRE (G177-L350) using the CHARMM36 all-atom force field [37]. Structures were solvated in a cubic box and embedded with an appropriate amount of Na+ and Cl- counterions for an electroneutral system. Energy minimization and equilibration of the structures were completed using a steepest descent method and subsequent NVT and NPT equilibration phases lasting 100 ps each. 200 ns simulations were performed on the equilibrated and energy minimized structures at constant temperature and pressure (300.0 K and 1.0 atm) with a time step of 2.0 fs. Global backbone RMSF over the course of the simulation reached convergence after 30 ns. The trajectory file was analyzed using VMD to visualize long-lived water molecules and changes in cleavage site side chain rotamers [63]. Fluctuations in dihedral angles over the course of the simulation were extracted from the trajectory file and visualized in XMGRACE software [64]. Hydrogen bond site occupancies were calculated using the HBonds plugin, v1.2 in VMD with a distance and angle cutoffs of 3.4 Å and 35° respectively.

3.5.4 Tandem mass spectrometry and MALDI-TOF

Full length CRE was SEC purified and buffer-exchanged into 100 mM ammonium bicarbonate, pH 8.0. A tryptic digest was performed at a ratio of 50:1 RsgI9:trypsin, overnight, using sequencing grade trypsin (Promega) before desalting using C18 resin (Empore) in the format of StageTips [65]. The resulting peptides were injected onto an Exploris Orbitrap 480 (Thermo Fisher) running the following mobile phase gradient: 3% B from 0 to 7 minutes, 3% to 35% B from 7 to 40 minutes, 35% to 80% B from 40 to 45 minutes, 80% B to 50 minutes, and 80% to 3% B from 50 to 52 minutes. Data-dependent acquisition was implemented, and the

resulting data was analyzed using Max Quant with the Andromeda search engine, default parameters. MS1 chromatograms were extracted using FreeStyle 1.8 (Thermo Fisher) and plotted and centered at their apices. Purified protein samples were analyzed by MALDI-TOF by first mixing with matrix solution (70% Acetonitrile, 30% ddH2O, 0.1% TFA, saturated α-cyano-4-hydroxycinnamic acid) and analyzed with an Applied Biosystems Voyager-DE STR MALDI-TOF in linear mode.

3.6 Figures

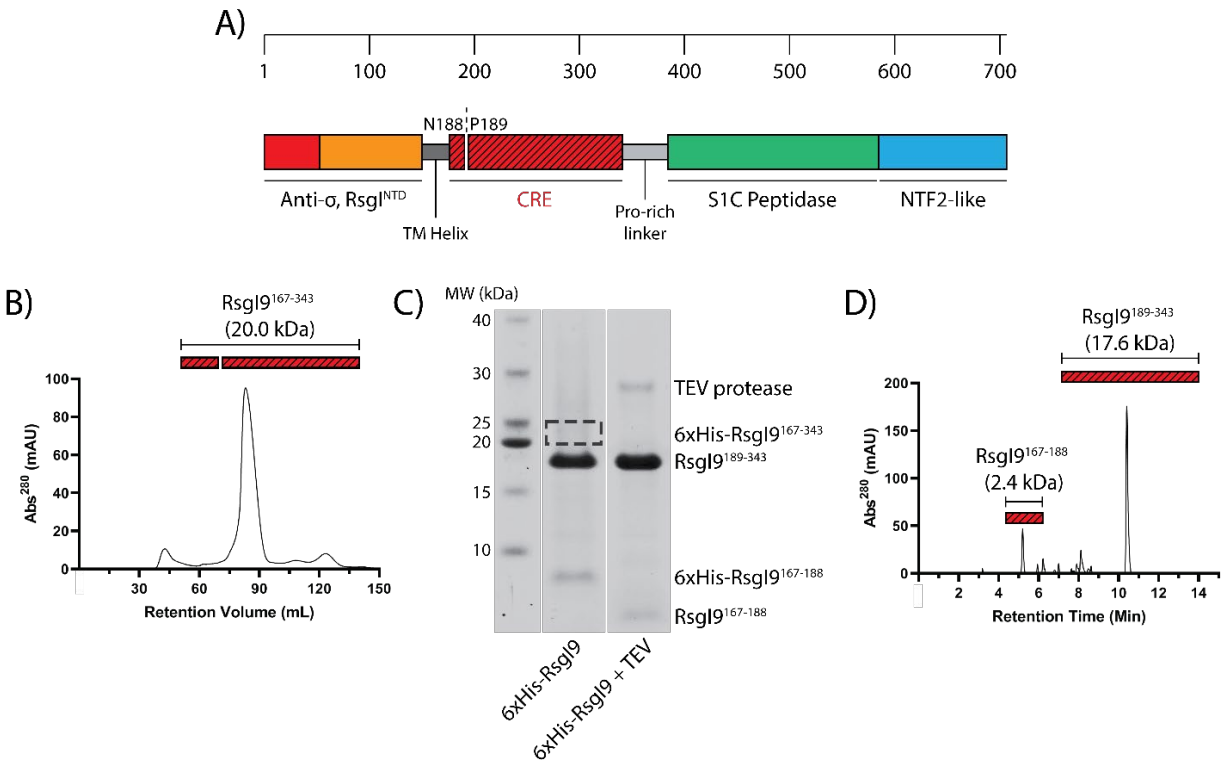


Figure 3.1 The RsgI9 anti-σ factor is proteolyzed at the Asn^{188}-Pro^{189} peptide bond.

(A) Schematic of the intact RsgI9 protein. It contains conserved domains that are found in other RsgI-family proteins (red). These include an intracellular anti-σ factor domain (NTD, residues 1–55, red), a transmembrane helix (TM, dark gray) and an extracellular domain of unknown function (CRE, residues 167–343, red, striped). RsgI9 and many other RsgI proteins also contain a Pro-rich linker segment (light gray) that connects the RsgI9 CRE to S1C peptidase (mint, residues 396–578) and NTF2-like type domains (blue, residues 579–707). RsgI9 also contains a unique insertion (orange) immediately following the NTD that is in the cytoplasm. (B) SEC chromatogram of purified CRE protein, demonstrating that it elutes as a single peak. The elution volume for CRE is consistent with the expected molecular weight of the uncleaved form. (C) SDS-PAGE of His-tagged RsgI9 CRE (167-343) digested with TEV protease. A band corresponding to a fully intact, undigested 6xHis-RsgI9 (residues 167-343) is absent in the pre-

digestion sample. (D) Reversed-phase HPLC chromatogram of purified CRE obtained from SEC. The data show that CRE consists of two polypeptides, short (residues G167 to N188, RsgI9) and long (residues 189-343, RsgI9) fragments. The identity of the polypeptides was confirmed by ESI-MS/MS.

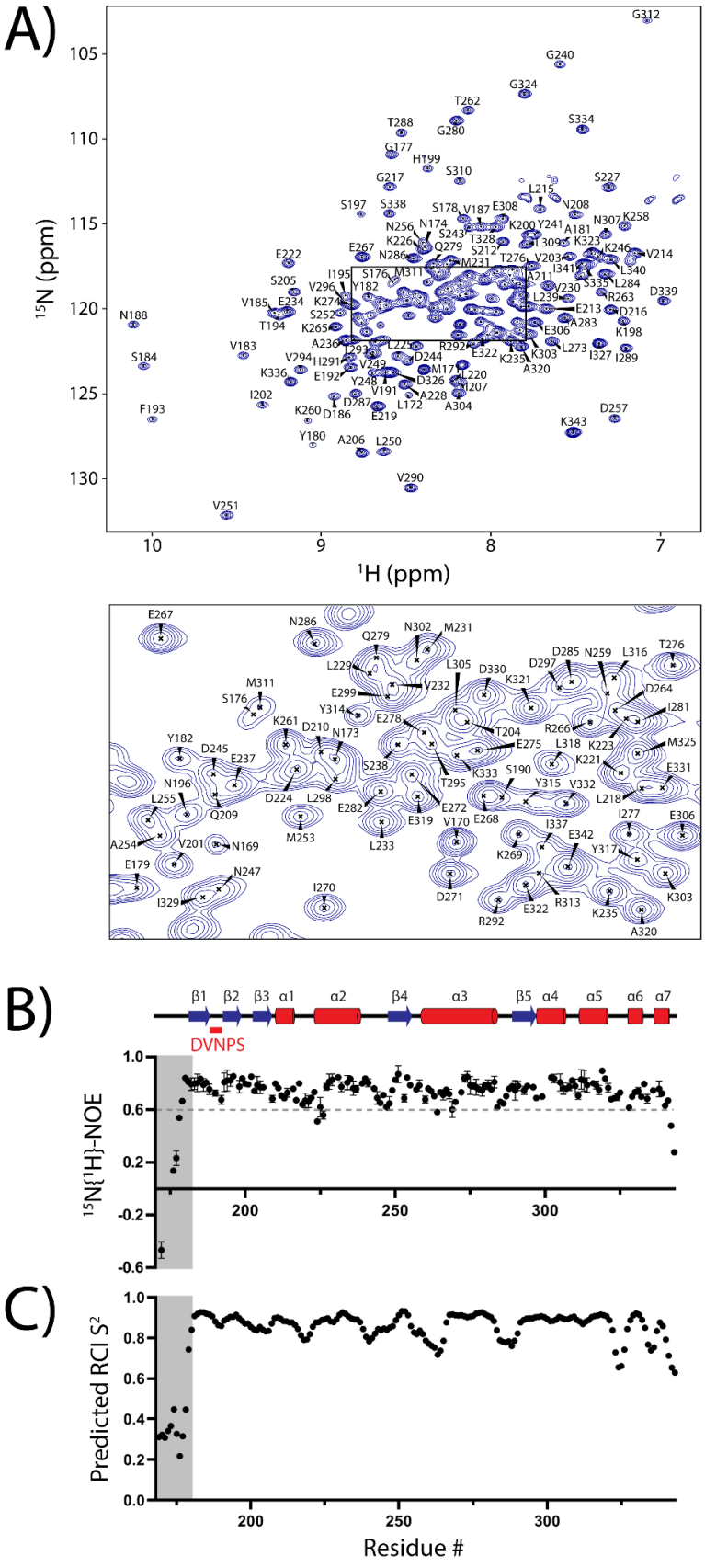

A)
^{15}N (ppm)
^{1}H (ppm)
B)
DVNPS
^{15}N{^{1}H}-NOE
C)
Predicted RCI S^2
Residue #

Figure 3.2 NMR data showing the cleaved CRE domain forms a stable structure.

(A) The ^{1}H-^{15}N HSQC spectrum of the uniformly ^{15}N-labeled RsgI9 CRE domain. Residue numbering corresponds to the full-length RsgI9 protein. (B) Plot of the steady state heteronuclear $^{15}N\{^{1}H\}$ NOE (hetNOE) values for the backbone amides in the domain plotted as a function of residue number. Error values are the maximal deviation from the average obtained from two separate measurements. A schematic of the secondary structural elements identified by NMR is depicted above. Out of a total of 176 residues containing backbone amide groups in CRE, resolved hetNOE data could be obtained for 165 residues. (C) Predicted Random Coil Index (RCI) values per residue as calculated using the program TALOS-N and chemical shift data. In panels (B) and (C) the light-gray shaded area represents the first 10 residues of the CRE protein construct and are structurally disordered.

Table 3.1 Structural statistics of the solution structure of CRE

	⟨SA⟩[a]
Root mean square deviation	
NOE Interproton distance restraints (Å)	
(1668)[b]	0.028 ± 0.002
Dihedral angles restraints (°) (301)[c]	0.550 ± 0.059
Hydrogen bond distance restraints (Å) (74)	0.027 ± 0.005
Deviation from idealized covalent geometry	
bonds (Å)	0.002 ± 0.0001
angles (°)	0.425 ± 0.011
impropers (°)	0.301 ± 0.017
PROCHECK results (%)	
most favorable region	94.8 ± 1.1
additionally allowed region	4.4 ± 1.1
generously allowed region	0.7 ± 0.4
disallowed region	0.0 ± 0.0
Coordinate Precision (Å)[d]	
Protein backbone	0.57 ± 0.07
Protein heavy atoms	1.16 ± 0.07

[a]⟨SA⟩ represents an ensemble of the 20 best structures calculated by simulated annealing. The number of terms for each restraint is given in parentheses. None of the structures exhibited distance violations greater than 0.5 Å, or dihedral angle violations greater than 5°. Residues selected for statistics and PROCHECK-NMR analysis were Y180-I337. [32]

[b]Distance restraints: 463 sequential, 261 medium (2 ≤ residue separation ≤ 4), 506 long range (>4 residues apart) and 438 intramolecular.

[c]The experimental dihedral angle restraints were as follows: 151 φ and 151 ψ angular restraints.

[d]The coordinate precision is defined as the average atomic root mean square deviation (RMSD) of the 20 individual SA structures and their mean coordinates. These values are for residues Y180-I337 of CRE. Backbone atoms refer to the N, C^{α}, and C' atoms.

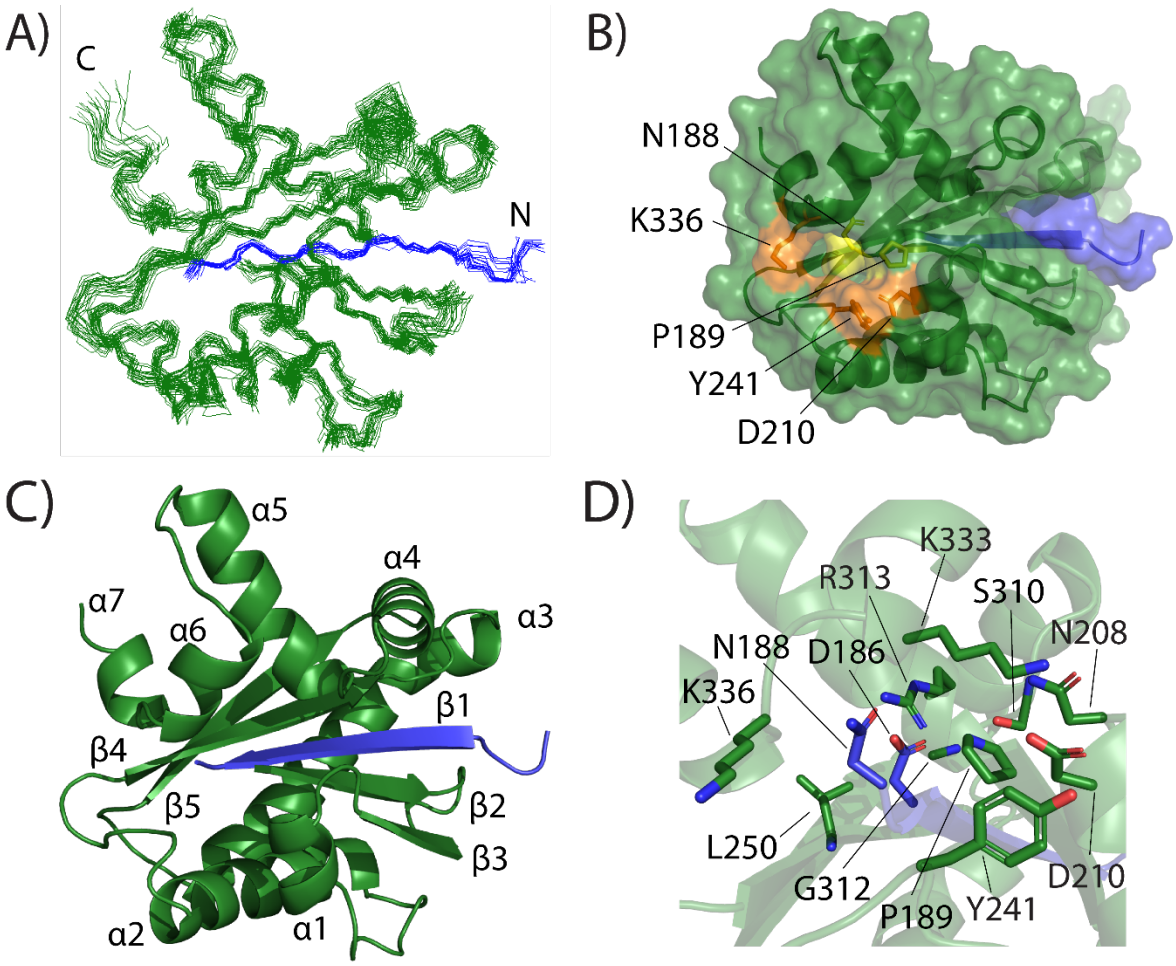


Figure 3.3 Solution structure of the Rsgl9 CRE domain.

(A) Superposition of the 20 lowest energy structures of the Rsgl9 CRE domain. Residues corresponding to the 2.4 and 17.6 kDa fragments are colored light and dark blue, respectively. Residues G167-K343 are shown with their N- and C-termini labeled. (B) Surface representation of the CRE domain structure. Cleavage site residues (yellow) and residues that partially occlude the cleavage site (orange) are colored. (C) Ribbon representation of the lowest energy structure of the CRE domain with its secondary structural elements labeled (α1-7, β1-5) (PDB accession: 8U9O) (D) Expanded view of the structure showing the cleaved peptide bond between the Asn^{188} and Pro^{189} residues. Carbon atoms are colored by fragment (G167-N188, blue and P189-K343, green). The side chains of nearby conserved residues are shown in stick format.

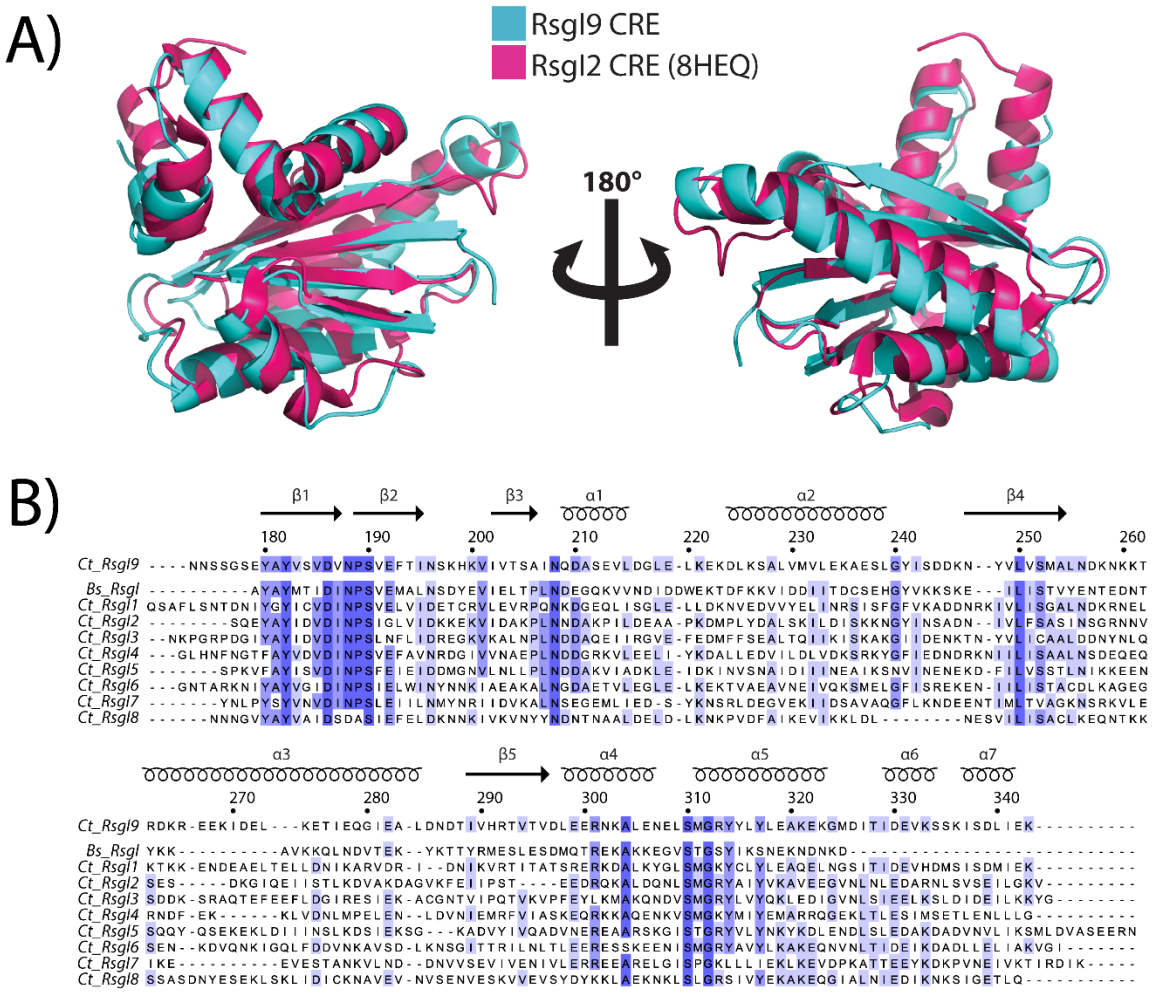


Figure 3.4 Structure and sequence comparisons with the RsgI9 CRE domain.

(A) Overlay of the lowest energy structure of the RsgI9 CRE domain (cyan, residues 167-343) with the structure of the CRE domain from the *C. thermocellum* RsgI2 protein (magenta, residues 89-248) (PDB: 8heq)[27]. The structures align with a backbone RMSD of 1.95 Å. (B) Sequence alignment of the CRE domain (top) with the other RsgI proteins in *C. thermocellum* (RsgI 1-8) and the RsgI anti-σ factor from *B. subtilis* (Bsub_RsgI). Residue numbers and secondary structures are shown above the sequences and correspond to the RsgI9 protein. Shaded residues are conserved (dark purple most conserved). Multiple sequence alignments for the RsgI proteins were generated using Clustal Omega [33].

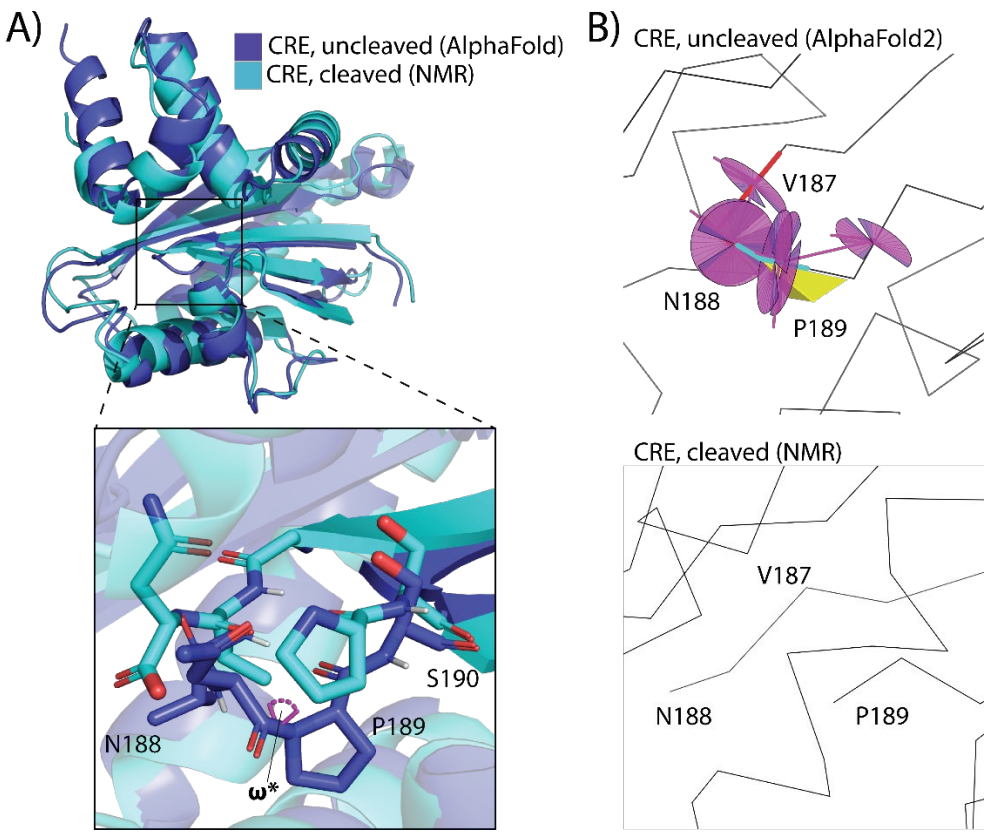


Figure 3.5 AlphaFold2 Comparison to solution-state structure of CRE.

(A) Overlay and zoom of the lowest overall energy structure calculated for RsgI9 CRE (167-343) and the AlphaFold2-predicted model of CRE (177-350). The solution-state NMR structure of CRE contains the cleavage site breakage between N188 and P189 whereas the predicted AlphaFold2 structure is continuous through that region. The two structures align with a backbone RMSD of 1.94 Å. The conserved cleavage site residues (N188/P189) are compared in the zoom-in window. The ω* denotes a strained, non-planar peptide angle (83.2°) for P189 in the uncleaved AlphaFold2 model. (B) MolProbity kinemage analysis of the CRE cleavage site residues for the uncleaved AlphaFold2 model and solution state structure of CRE. Backbone carbons are outlined in white or pale yellow. Outlier values for Ramachandran (thick cyan line), bond-length (thick red line), twisted/non-planar peptide bonds (yellow trapezoid), and C-Alpha geometry (magenta discs) are shown.

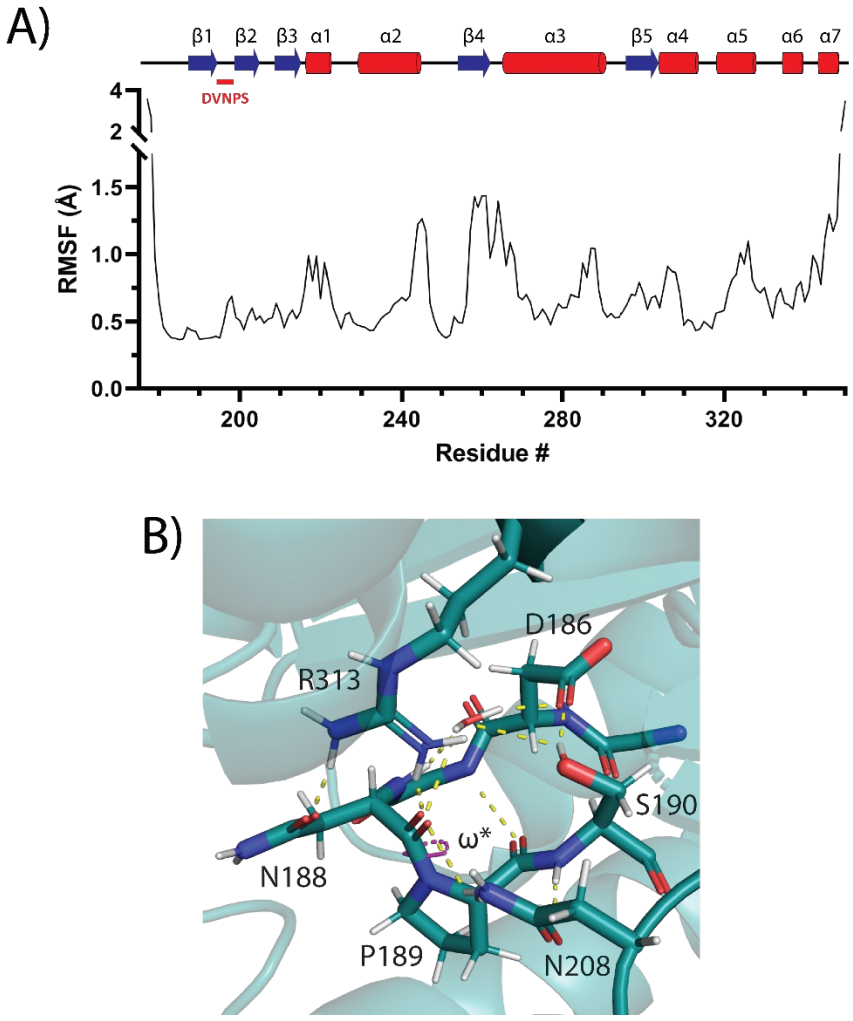
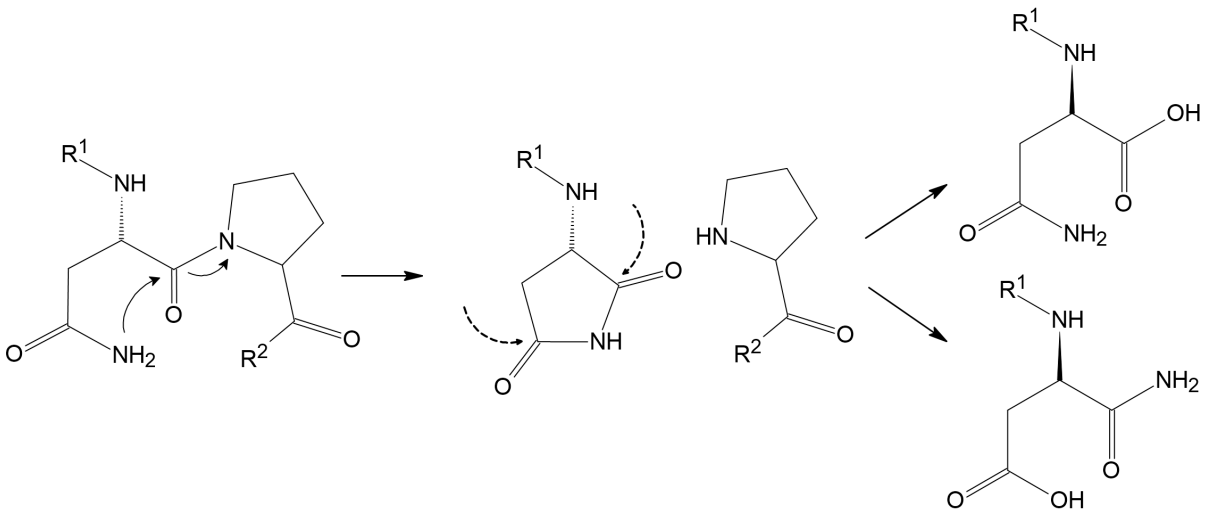


Figure 3.6 MD simulations of the AlphaFold2 model of CRE.

(A) Plot of the root mean square fluctuation (RMSF) differences of CRE backbone amide coordinates during the course of the 200 ns MD simulation. (B) Zoom in view of a representative frame from an uncleaved precursor RsgI9^{CRE} during the course of the MD simulation. The peptide bond dihedral angle for P189 (ω*, magenta) remains predominantly in a non-planar (153.2°) configuration as stabilized by the surrounding hydrogen bond network. Backbone amide and side chain hydrogens (white) are shown to illustrate peptide bond geometries. Heavy atoms are colored blue and red to indicate nitrogen and oxygen, respectively. Hydrogen bonds are denoted with dashed yellow lines.



Scheme 3.1 Mechanism of Succinimide formation from Asn-Pro Residues

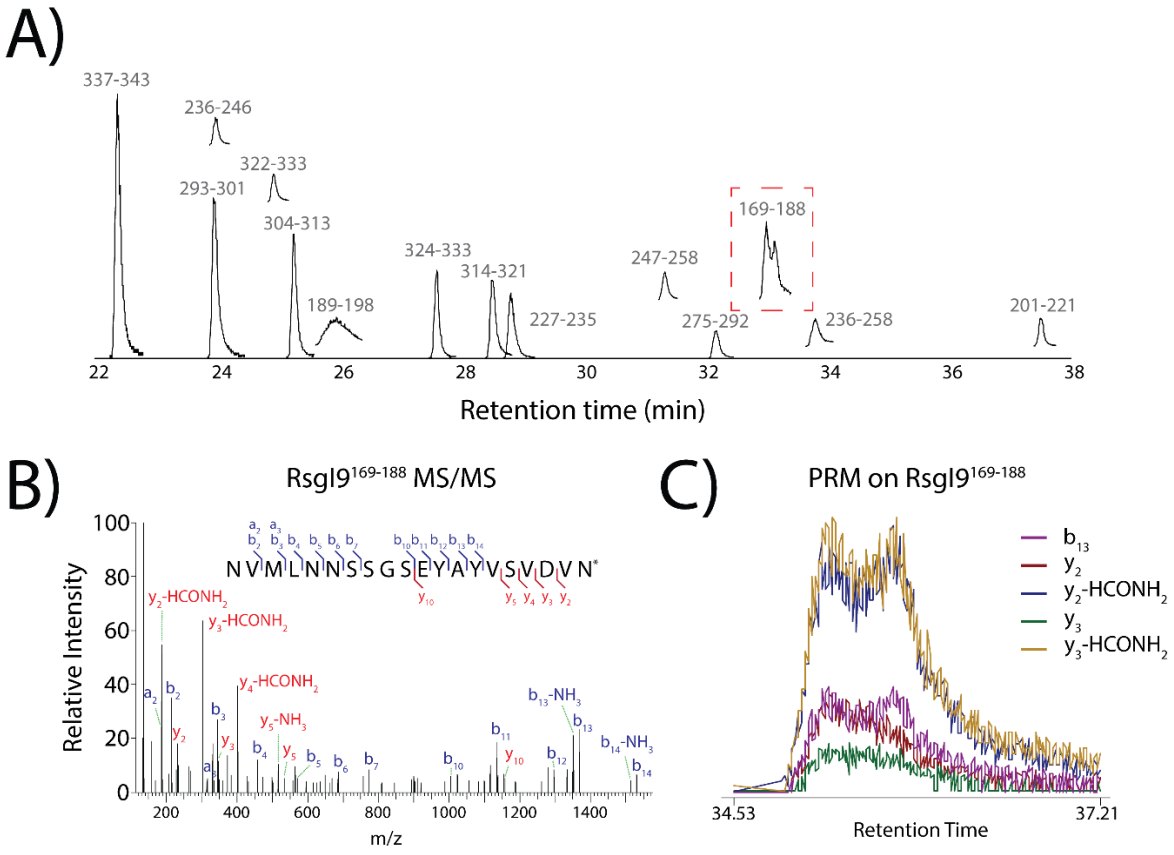


Figure 3.7 LC-MS/MS Chromatogram reveals bimodal peak for N-terminal CRE cleavage fragment.

(A) LC-MS extracted ion chromatograms of RsgI9 peptides. Numbers above each peak indicate the residue numbers of full-length RsgI9 CRE represented in each fragment. Those on the x-axis are to-scale and those above the x-axis are zoomed in to highlight their elution. Note peptide 169-188 is the only peptide with a bimodal elution. (B) Annotated MS2 spectra from peptide 169-188 confirm its identity. An asterisk (*) is used to denote either Asn/*iso*Asn (C) Parallel reaction monitoring (PRM) of the 169-188 product ions with an optimized elution gradient. The intensities of ions containing N22 residue (y_2, y_3, and dissociated products) as well as b_{13} are plotted, which shows the entire bimodal elution profile corresponds to elution peak CRE (169-188).

3.7 References

1. IEA, *World Energy Outlook 2022*, IEA, Editor. 2022, International Energy Agency (IEA): Paris.
2. Bar-On, Y.M., R. Phillips, and R. Milo, *The biomass distribution on Earth.* Proceedings of the National Academy of Sciences of the United States of America, 2018. **115**(25): p. 6506-6511.
3. Hussein N. Nassar, W.I.M.E.-a., Nour Sh. El-Gendy, *Sustainable ecofriendly recruitment of bioethanol fermentation lignocellulosic spent waste biomass for the safe reuse and discharge of petroleum production produced water via biosorption and solid biofuel production.* Journal of Hazardous Materials, 2022. **422**(15): p. 126845.
4. Johnson, E.A., et al., *Saccharification of Complex Cellulosic Substrates by the Cellulase System from Clostridium thermocellum.* Applied and Environmental Microbiology, 1982. **43**(5): p. 1125-1132.
5. Seo, H., et al., *Rewiring metabolism of Clostridium thermocellum for consolidated bioprocessing of lignocellulosic biomass poplar to produce short-chain esters.* Bioresource Technology, 2023. **384**: p. 129263.
6. Lu, Y., Y.-H.P. Zhang, and L.R. Lynd, *Enzyme-microbe synergy during cellulose hydrolysis by Clostridium thermocellum.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(44): p. 16165-16169.
7. Alves, V.D., C. Fontes, and P. Bule, *Cellulosomes: Highly Efficient Cellulolytic Complexes.* Subcell Biochem, 2021. **96**: p. 323-354.
8. Artzi, L., E.A. Bayer, and S. Morais, *Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides.* Nat Rev Microbiol, 2017. **15**(2): p. 83-95.

9. Singh, N., et al., *Enzyme systems of thermophilic anaerobic bacteria for lignocellulosic biomass conversion.* Int J Biol Macromol, 2021. **168**: p. 572-590.
10. Smith, S.P., E.A. Bayer, and M. Czjzek, *Continually emerging mechanistic complexity of the multi-enzyme cellulosome complex.* Curr Opin Struct Biol, 2017. **44**: p. 151-160.
11. Leibovitz, E., et al., *Characterization and subcellular localization of the Clostridium thermocellum scaffoldin dockerin binding protein SdbA.* Journal of Bacteriology, 1997. **179**(8): p. 2519-2523.
12. Doi, R.H., et al., *Cellulosomes from mesophilic bacteria.* Journal of Bacteriology, 2003. **185**(20): p. 5907-5914.
13. Dassa, B., et al., *Pan-Cellulosomics of Mesophilic Clostridia: Variations on a Theme.* Microorganisms, 2017. **5**(4).
14. Dassa, B., et al., *Genome-wide analysis of acetivibrio cellulolyticus provides a blueprint of an elaborate cellulosome system.* BMC genomics, 2012. **13**: p. 210.
15. Artzi, L., et al., *Clostridium clariflavum: Key Cellulosome Players Are Revealed by Proteomic Analysis.* mBio, 2015. **6**(3): p. e00411-00415.
16. Nataf, Y., et al., *Clostridium thermocellum cellulosomal genes are regulated by extracytoplasmic polysaccharides via alternative sigma factors.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(43): p. 18646-18651.
17. Kahel-Raifer, H., et al., *The unique set of putative membrane-associated anti-sigma factors in Clostridium thermocellum suggests a novel extracellular carbohydrate-sensing mechanism involved in gene regulation.* FEMS microbiology letters, 2010. **308**(1): p. 84-93.

18. Sand, A., et al., *Three cellulosomal xylanase genes in Clostridium thermocellum are regulated by both vegetative SigA (sigma(A)) and alternative SigI6 (sigma(I6)) factors.* FEBS Lett, 2015. **589**(20 Pt B): p. 3133-40.

19. Munoz-Gutierrez, I., et al., *Decoding Biomass-Sensing Regulons of Clostridium thermocellum Alternative Sigma-I Factors in a Heterologous Bacillus subtilis Host System.* PLoS One, 2016. **11**(1): p. e0146316.

20. Grinberg, I.R., et al., *Distinctive ligand-binding specificities of tandem PA14 biomass-sensory elements from Clostridium thermocellum and Clostridium clariflavum.* Proteins, 2019. **87**(11): p. 917-930.

21. Bahari, L., et al., *Glycoside hydrolases as components of putative carbohydrate biosensor proteins in Clostridium thermocellum.* Journal of Industrial Microbiology & Biotechnology, 2011. **38**(7): p. 825-832.

22. Mahoney, B.J., et al., *The structure of the Clostridium thermocellum RsgI9 ectodomain provides insight into the mechanism of biomass sensing.* Proteins, 2022. **90**(7): p. 1457-1467.

23. Marcos-Torres, F.J., et al., *Mechanisms of Action of Non-Canonical ECF Sigma Factors.* Int J Mol Sci, 2022. **23**(7).

24. Asai, K., *Anti-sigma factor-mediated cell surface stress responses in Bacillus subtilis.* Genes & Genetic Systems, 2018. **92**(5): p. 223-234.

25. Brunet, Y.R., et al., *Intrinsically disordered protein regions are required for cell wall homeostasis in Bacillus subtilis.* Genes & Development, 2022. **36**(17-18): p. 970-984.

26. Brogan, A.P., et al., *Bacterial SEAL domains undergo autoproteolysis and function in regulated intramembrane proteolysis.* Proceedings of the National Academy of Sciences of the United States of America, 2023. **120**(40): p. e2310862120.

27. Chen, C., et al., *Essential autoproteolysis of bacterial anti-σ factor RsgI for transmembrane signal transduction.* Science Advances, 2023. **9**(27): p. eadg4846.

28. *Protein NMR spectroscopy: principles and practice*. 2. ed. 2007, Amsterdam Heidelberg: Elsevier Academic Press. 885.

29. Shen, Y, et al., *TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts.* Journal of biomolecular NMR, 2009. **44**(4): p. 213-223.

30. Kay, L.E., D.A. Torchia, and A. Bax, *Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease.* Biochemistry, 1989. **28**(23): p. 8972-9.

31. Stetz, M.A., et al., *Characterization of Internal Protein Dynamics and Conformational Entropy by NMR Relaxation.* Methods Enzymol, 2019. **615**: p. 237-284.

32. Laskowski, R.A., et al., *AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR.* Journal of biomolecular NMR, 1996. **8**(4): p. 477-486.

33. Madeira, F., et al., *Search and sequence analysis tools services from EMBL-EBI in 2022.* Nucleic Acids Research, 2022. **50**(W1): p. W276-W279.

34. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.

35. Chen, V.B., et al., *MolProbity: all-atom structure validation for macromolecular crystallography.* Acta Crystallographica. Section D, Biological Crystallography, 2010. **66**(Pt 1): p. 12-21.

36. Prisant, M.G., et al., *New tools in MolProbity validation: CaBLAM for CryoEM backbone, UnDowser to rethink "waters," and NGL Viewer to recapture online 3D graphics.* Protein Sci, 2020. **29**(1): p. 315-329.

37. Huang, J. and A.D. MacKerell, *CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data.* Journal of Computational Chemistry, 2013. **34**(25): p. 2135-2145.

38. Guan, C., et al., *Activation of glycosylasparaginase. Formation of active N-terminal threonine by intramolecular autoproteolysis.* The Journal of Biological Chemistry, 1996. **271**(3): p. 1732-1737.

39. Johansson, D.G., et al., *Protein autoproteolysis: conformational strain linked to the rate of peptide cleavage by the pH dependence of the N --> O acyl shift reaction.* J Am Chem Soc, 2009. **131**(27): p. 9475-7.

40. Voorter, C.E., et al., *Spontaneous peptide bond cleavage in aging alpha-crystallin through a succinimide intermediate.* The Journal of Biological Chemistry, 1988. **263**(35): p. 19020-19023.

41. Boon, L., et al., *Protease propeptide structures, mechanisms of activation, and functions.* Critical Reviews in Biochemistry and Molecular Biology, 2020. **55**(2): p. 111-165.

42. Kemp, M.T., E.M. Lewandowski, and Y. Chen, *Low barrier hydrogen bonds in protein structure and function.* Biochimica Et Biophysica Acta. Proteins and Proteomics, 2021. **1869**(1): p. 140557.

43. Perler, F.B., M.Q. Xu, and H. Paulus, *Protein splicing and autoproteolysis mechanisms.* Current Opinion in Chemical Biology, 1997. **1**(3): p. 292-299.

44. Bastiaansen, K.C., et al., *Self-cleavage of the Pseudomonas aeruginosa Cell-surface Signaling Anti-sigma Factor FoxR Occurs through an N-O Acyl Rearrangement.* The Journal of Biological Chemistry, 2015. **290**(19): p. 12237-12246.

45. Araç, D., et al., *A novel evolutionarily conserved domain of cell-adhesion GPCRs mediates autoproteolysis.* The EMBO journal, 2012. **31**(6): p. 1364-1378.

46. Hsu, M.-F., et al., *Mechanism of the maturation process of SARS-CoV 3CL protease.* The Journal of Biological Chemistry, 2005. **280**(35): p. 31257-31266.

47. Ball, L.E., et al., *Post-translational modifications of aquaporin 0 (AQP0) in the normal human lens: spatial and temporal occurrence.* Biochemistry, 2004. **43**(30): p. 9856-9865.

48. Ferris, H.U., et al., *FlhB regulates ordered export of flagellar components via autocleavage mechanism.* The Journal of Biological Chemistry, 2005. **280**(50): p. 41236-41242.

49. Monjarás Feria, J.V., et al., *Role of autocleavage in the function of a type III secretion specificity switch protein in Salmonella enterica serovar Typhimurium.* mBio, 2015. **6**(5): p. e01459-01415.

50. Kato, K., et al., *Computational Analysis of the Mechanism of Nonenzymatic Peptide Bond Cleavage at the C-Terminal Side of an Asparagine Residue.* ACS omega, 2021. **6**(44): p. 30078-30084.

51. Robinson, N.E. and A.B. Robinson, *Molecular clocks.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(3): p. 944-949.

52. Geiger, T. and S. Clarke, *Deamidation, isomerization, and racemization at asparaginyl and aspartyl residues in peptides. Succinimide-linked reactions that contribute to protein degradation.* The Journal of Biological Chemistry, 1987. **262**(2): p. 785-794.

53. Tarelli, E. and P.H. Corran, *Ammonia cleaves polypeptides at asparagine proline bonds.* The Journal of Peptide Research: Official Journal of the American Peptide Society, 2003. **62**(6): p. 245-251.

54. Delaglio, F., et al., *NMRPipe: a multidimensional spectral processing system based on UNIX pipes.* Journal of biomolecular NMR, 1995. **6**(3): p. 277-293.

55. Garrett, D.S., M. Cai, and G.M. Clore, *XIPP: multi-dimensional NMR analysis software.* Journal of biomolecular NMR, 2020. **74**(1): p. 9-25.

56. Herrmann, T., P. Güntert, and K. Wüthrich, *Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA.* Journal of Molecular Biology, 2002. **319**(1): p. 209-227.

57. Lee, W., M. Tonelli, and J.L. Markley, *NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy.* Bioinformatics (Oxford, England), 2015. **31**(8): p. 1325-1327.

58. Shen, Y. and A. Bax, *Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks.* Journal of biomolecular NMR, 2013. **56**(3): p. 227-241.

59. Schwieters, C.D., G.A. Bermejo, and G.M. Clore, *Xplor-NIH for molecular structure determination from NMR and other data sources.* Protein Science: A Publication of the Protein Society, 2018. **27**(1): p. 26-40.

60. Schwieters, C.D., et al., *The Xplor-NIH NMR molecular structure determination package.* Journal of Magnetic Resonance (San Diego, Calif.: 1997), 2003. **160**(1): p. 65-73.

61. Koradi, R., M. Billeter, and K. Wüthrich, *MOLMOL: a program for display and analysis of macromolecular structures.* Journal of Molecular Graphics, 1996. **14**(1): p. 51-55, 29-32.

62. Lam, W.W.T. and S.W.I. Siu, *PyMOL mControl: Manipulating molecular visualization with mobile devices.* Biochemistry and Molecular Biology Education: A Bimonthly Publication of the International Union of Biochemistry and Molecular Biology, 2017. **45**(1): p. 76-83.

63. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* Journal of Molecular Graphics, 1996. **14**(1): p. 33-38, 27-28.

64. Turner, P.J., *XMGRACE*. 2005, Center for Coastal and Land-Margin Research, Oregon Graduate Institute of Science and Technology: Beaverton, OR.

65. Rappsilber, J., M. Mann, and Y. Ishihama, *Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips.* Nature Protocols, 2007. **2**(8): p. 1896-1906.

Chapter 4

A Genomic Analysis Reveals the Diversity of Cellulosome Displaying Bacteria

The work described in this chapter has been reproduced with permission from:

Minor CM & **Takayesu A**, Ha SM, Salwinski L, Sawaya MR, Pellegrini M, Clubb RT. A genomic analysis reveals the diversity of cellulosome displaying bacteria. Front Microbiol. 2024 Oct 30;15:1473396. doi: 10.3389/fmicb.2024.1473396.

I was a co-primary author of this work along with Christine Minor. I assembled literature references, made figures, analyzed data and wrote the manuscript.

4.1 Overview

Several species of cellulolytic bacteria display cellulosomes, massive multi-cellulase containing complexes that degrade lignocellulosic plant biomass (LCB). A greater understanding of cellulosome structure and enzyme content could facilitate the development of new microbial-based methods to produce renewable chemicals and materials. To identify novel cellulosome displaying microbes we searched 305,693 sequenced bacterial genomes for genes encoding cellulosome proteins; dockerin-fused glycohydrolases (DocGHs) and cohesin domain containing scaffoldins. This analysis identified 33 bacterial species with the genomic capacity to produce cellulosomes, including 10 species not previously reported to produce these complexes, such as *Acetivibrio mesophilus*. Cellulosome producing bacteria primarily originate from the *Acetivibrio, Ruminococcus, Ruminiclostridium*, and *Clostridium* genera. A rigorous analysis of their enzyme, scaffoldin, dockerin, and cohesin content reveals phylogenetically conserved features. Based on the presence of a high number of genes encoding both scaffoldins and dockerin-fused GHs, the cellulosomes in *Acetivibrio* and *Ruminococcus* bacteria possess complex architectures that are populated with a large number of distinct LCB degrading GH enzymes. Their complex cellulosomes are distinguishable by their mechanism of attachment to the cell wall, the structures of their primary scaffoldins, and by how they are transcriptionally regulated. In contrast, bacteria in the *Ruminiclostridium* and *Clostridium* genera produce ‘simple’ cellulosomes that are constructed from only few types of scaffoldins that based on their distinct complement of GH enzymes are predicted to exhibit high and low cellulolytic activity, respectively. Collectively, the results of this study reveal conserved and divergent architectural features in bacterial cellulosomes that could be useful in guiding ongoing efforts to harness their cellulolytic activities for bio-based chemical and materials production.

4.2 Introduction

Lignocellulosic plant biomass (LCB) is the largest source of carbon in the biosphere and a promising feedstock for the production of renewable materials, biofuels, and chemicals [1]. LCB's utility is limited by its recalcitrance to hydrolysis which makes it costly to degrade at an industrial scale [2,3]. LCB consists of crystalline cellulose (32-47% dry weight) and hemicellulose (19-27%) encased within an aromatic polymer web of lignin (5-24%) [4]. Several species of highly cellulolytic anaerobic bacteria have garnered significant interest as potential tools to efficiently deconstruct LCB into its components sugars for use in bio-commodity production [5]. These microbes display massive multi-cellulase containing complexes called cellulosomes that degrade LCB's cellulose and hemicellulose components [6-11]. A deeper understanding of the structural diversity of these complexes in bacteria could facilitate their usage in industrial applications.

Cellulosomes were first discovered in *Acetobivrio thermocellus* (formerly known as *Clostridium thermocellum* and *Hungateiclostridium thermocellum*) as a "cellulose-binding factor", and subsequently have been found in a range of anaerobic eubacteria [7,12]. These multi-cellulase complexes are constructed from scaffolding proteins (called scaffoldins) that coordinate the binding of an array of glycoside hydrolases (GHs) [6-10]. Each scaffoldin contains one or more cohesin domains that bind noncovalently to dockerin domains that are genetically fused to the GHs (called DocGH enzymes). In addition, scaffoldins can harbor carbohydrate binding module (CBM) domains, cell wall (e.g. S-layer homology (SLH) domains) binding modules, and dockerin domains that interact with other scaffoldins on the cell surface. For example, the ScaA scaffoldin in *A. thermocellus* contains multiple cohesin domains that bind DocGH enzymes, an internal CBM type-3 domain (CBM3) that binds cellulose, and a C-terminal dockerin domain that enables it to associate with a series of cell wall associated scaffoldins (ScaB, ScaC, ScaD, and ScaF) [13]. Three major types of GHs function synergistically to

degrade cellulose: endoglucanases, exoglucanases, and β-glucosidases [14,15]. Endoglucanases hydrolyze internal β-(1,4)-glycosidic bonds in cellulose (e.g. GH7, GH12), creating reducing and non-reducing ends that are further hydrolyzed by exoglucanases (e.g. GH5, GH9). The resulting cellodextrin carbohydrate oligomers are then degraded into glucose by β-glucosidases. The carbohydrate substrate specificities of cellulosomal DocGH enzymes vary, but members of the GH5, GH10, GH11, GH43, and GH48 families are frequently present in cellulosome producing bacteria [7]. Other types of carbohydrate active enzymes (CAZymes) are also fused to dockerin domains enabling their incorporation into cellulosomes, including polysaccharide lyases (PLs) and carbohydrate esterases (CEs). Collectively, the diversity of DocGH enzymes, CAZymes, and CBM modules within the cellulosome enable bacteria to degrade LCB more efficiently than microbes that simply secrete GHs, because enzyme colocalization by the cellulosome promotes enzyme-enzyme synergy, enzyme-proximity enhancement, and cellulose-enzyme-microbe interactions [16-18].

Several species of mesophilic and thermophilic anaerobic bacteria display cellulosomes that vary in their complexity and composition. These include complex, simple, and cell-free cellulosomes that differ in both the number and types of cohesin containing scaffoldins they possess [6,7,19,20]. Three types of cohesins (Coh1, Coh2, and Coh3) and dockerin (Doc1, Doc2, and Doc3) domains have been identified based on their primary sequences [9]. Biochemical experiments have shown that these domains typically interact with one another in a species- and type-specific manner (e.g. Doc1 binds to Coh1 domains, but not with Coh2 or Coh3 domains within the same species) [21,22], however, there are several exceptions [23-25]. It has also been noted that there are two distinct conformations Doc1 modules can bind to Coh1 modules, further expanding the possible cellulosome architectures [26,27]. Complex cellulosomes, typified by the one present in *Acetivibrio cellulolyticus*, contain a primary scaffoldin that harbors several Coh1 modules for DocGH binding and a C-terminal Doc2 that

enables it to interact with Coh2 modules presented in cell wall associated anchoring scaffoldins (**Fig. 1**, right) [28]. In many cases, microbes containing this type of primary scaffoldin also possess adaptor scaffoldins that harbor both cohesin and dockerin domains that are believed to expand both the number and types of DocGH proteins that are incorporated into the cellulosome (e.g. *Acetivibrio clariflavus* [25], *Pseudobacteroides cellulosolvens* [29], *A. cellulolyticus* [30], and *Acetivibrio alkalicellulosi* [24]). Other bacterial species typified by *Clostridium acetobutylicum* produce simple cellulosomes that contain a singular multi-cohesin primary scaffoldin that houses an N-terminal CBM3 and interspersed X2 domains (e.g. *Clostridium cellulovorans, Ruminiclostridium cellulolyticum, Clostridium josui*) (**Fig. 1**, left) [6]. The mechanism of cell surface attachment by simple cellulosomes is poorly understood but is likely mediated by the N-terminal CBM3 present on their primary scaffoldins as seen for *R. cellulolyticum* [31]. Lastly, many microbes possess cell-free cellulosomes composed of multi-cohesin scaffoldins bound with DocGH enzymes which are secreted into the environment to degrade LCB (e.g. *A. clariflavus*) [25,32].

To gain a comprehensive understanding of cellulosome displaying bacteria that could have applications in LCB degradation, we analyzed >305K complete and draft microbial genomes for genes encoding cellulosomal proteins. This analysis revealed a total of 33 bacterial species have the capacity to produce cellulosomes, including 10 species previously not reported in the literature. These microbes produce simple or complex cellulosomes that are populated with either small or large numbers of DocGH enzymes that are known to degrade LCB. The majority of cellulosome producing bacteria are members of the *Acetivibrio*, *Clostridium*, *Ruminiclostridium* and *Ruminococcus* genera and exhibit phylogenetically conserved properties when their scaffoldins, enzymes, and mechanisms of cellulosomal gene regulation are compared.

4.3 Results

4.3.1 Identification of cellulosome displaying bacteria

To identify bacteria that display cellulosomes we searched a total of 305,693 prokaryotic genome assemblies within the NCBI Reference Sequence Database (RefSeq), which contains both complete and draft genomes (**Fig. 2**) [34]. The genome of *Acetivibrio alkalicellulosi* is not deposited in the RefSeq database, because it has been suppressed due to contamination (**Table S2**). However, it was also included in our analysis because it has been reported to produce a cellulosome [24]. Initially, the program HMMER [35] was used to search the RefSeq-annotated protein-encoding genes for dockerin, cohesin, and GH domains using hidden Markov models (HMMs) obtained from the Conserved Domains Database (CDD) [36], Protein family database (Pfam) [37], and PROSITE database (see methods for the profile identifiers) [38]. A total of 255 genomes were carried forward for further analysis as they harbored at least one gene encoding a multi-cohesin containing scaffoldin (≥ 2 cohesin domains) and one gene encoding a DocGH protein. For cases where multiple genome sequences were available for the same bacterial species, only the representative genome within the ProGenomes database that contained the fewest number of contigs was analyzed [39]. An exception was made for the three sequenced genomes from *A. thermocellus* (ATCC 27405, DSM 1313, and AD2), as it is a prototypical cellulosome producing organism. After eliminating redundancies, a total of 139 distinct microbial genomes were retained for a more extensive and computationally demanding analysis using the program InterProScan (v. 5.59-91.0) [40]. This program defines the domain complement in each translated protein encoding gene by accessing several databases, including: Phobius (v 1.01) [41], SUPERFAMILY (v. 1.75) [42], ProSiteProfiles (v. 2022_01) [43], CDD (v. 3.18) [44], and Pfam (v. 35.0) [45]. Protein sequences were also analyzed using SignalP (v. 6.0) [46], Deep TMHMM (v. 1.0.19) [47], and PSORTb (v.3.0) [48]. The analysis by InterProScan revealed a total of 37 bacterial species containing genes encoding putative

cellulosomes (their genomes have at least one multi-cohesin and one DocGH encoding gene). Of these, 33 species likely produce conventional cellulosomes that are related to those in *C. acetobutylicum* and *A. cellulolyticus* (**Fig. 1**), whereas 4 species may produce non-conventional cellulosomes (described below).

4.3.2 Classification of bacteria based on their scaffoldin and DocGH enzyme content

To gain insight into the structure and composition of each bacterium's cellulosome, we systematically classified their scaffoldins into eight categories based on their domain content (**Fig. 3**). The eight scaffoldin types and their functions in cellulosome assembly are demonstrated for *C. acetobutylicum* and *A. cellulolyticus*, which produce simple and complex cellulosomes, respectively (**Fig. 1**) [30,50]. These types include: (1) "simple primary" scaffoldins that are produced by *C. acetobutylicum* and other mesophiles which contain an N-terminal CBM3 and multiple cohesin domains (either Coh1 or Coh2) that are often interspersed with X2 domains [7], (2) "single-cohesin" containing proteins of unknown function, (3) "complex primary" scaffoldins found in *A. cellulolyticus* and related species that harbor multiple Coh1 modules, an internal CBM3, and a single C-terminal Doc2 domain that enables it to bind to cell wall anchoring scaffoldins [30], (4) "SLH-anchoring" scaffoldins that contain at least one cohesin module paired with a SLH-domain, (5) "nonSLH-anchoring" scaffoldins that contain at least one cohesin module and a known cell wall interacting domain/motif (e.g. LPxTG sorting signal, Lysin motif, C-terminal TM-helix, or Cu-Amine Oxidase-like domains), (6) "Monovalent adaptor" scaffoldins that contain a dockerin and a single cohesin module which have been proposed to facilitate type-switching between different types of cohesins and dockerin-fused enzymes [7], (7) "Polyvalent adaptor" scaffoldins that contain a dockerin and several cohesins that suggest that more elaborate cellulosome architectures can be constructed by increasing the number of binding sites for DocGH enzyme proteins and/or scaffoldins [7], and (8) "cell-free" scaffoldins

that are presumably secreted to degrade LCB as they contain several cohesin domains that are capable of binding to DocGH enzymes [32].

Next, we closely examined the enzyme composition of each bacterial species. The Carbohydrate-Active Enzyme database (CAZy) is a well-curated resource that classifies glycoside hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), and glycosyl transferases (GTs) into families based on experimentally published data and publicly available sequences.[51] GHs, PLs, and CEs are of particular interest in cellulosome producing bacteria since their activities are directly linked to polymer breakdown. These enzymes are listed in Table S1 and collectively referred to as CAZymes. Many of these CAZymes are often fused to a dockerin domain to facilitate their incorporation cellulosomes, while enzymes not fused to dockerins are presumably secreted by the bacterium. Using the criteria outlined in Table S1 we identified dockerin proteins fused GHs with activity against cellulose (DocGH-Cell), hemicellulose (DocGH-Hemi), and oligosaccharides (DocGH-Oligo), as well their corresponding non-dockerin fused GHs (referred to as FreeGH-Cell, FreeGH-Hemi, and FreeGH-Oligo, respectively) (Table S4). DocGH-Cell and DocGH-Hemi enzymes are of particular interest as they contribute significantly to cellulolytic activity against LCB and are distinguished as DocGH-LCB enzymes (DocGH-Cell plus DocGH-Hemi). A similar category was defined for non-dockerin fused GHs with LCB activity (FreeGH-LCB). Based on their DocGH-LCB count there are two broad types of organisms that display cellulosomes, “high DocGH-LCB” microbes that contain genes encoding a large number of these enzymes (22 to 70) and “low DocGH-LCB” microbes that have fewer DocGH-LCBs (1 to 10) (**Fig. 4**).

4.3.3 Cellulosomes in *Ruminococcus* species.

Our genomic screen detected all multi-cohesin cellulosome bacteria previously documented to produce cellulosomes, as well as several new species. However, we detected fewer cohesin domains containing proteins than previously reported in the literature for three

genomes from two bacterial species, *R. champanellensis* (18P13) and *R. flavefaciens* (strains 17, 007c) [52,53]. The genome for *R. flavefaciens* strain FD-1 previously described in the literature as containing a cellulosome was not analyzed by us because its sequenced genome is suppressed in the RefSeq database and because this species is already represented by several sequenced genomes [53]. For the remaining three genomes, there are two main reasons why their cohesin modules were undercounted. First, each genome contains a large number of contigs that lead to sequencing truncations in their cohesin-containing scaffoldin genes (**Table S2** lists the sequencing statistics for the genomes analyzed in this study). These truncations led to nine abbreviated cohesin-containing genes in *R. flavefaciens* strain 17, two in *R. flavefaciens* strain 007c, and four in *R. champanellensis* lowering the number of detectable cohesins. For example, sequencing truncations in *R. flavefaciens* strain 17 occur in several of its scaffoldin genes (ScaB, ScaE, ScaG, ScaI, and orf02408), that previously were identified by sequencing a single contig from this microbe [54-56].

A second reason for the cohesin undercount is that the primary sequences of these modules in *Ruminococcus* species are highly divergent and thus not detectable using the HMM profiles employed by InterProScan (Coh1: cd08548, Coh2: cd08547, Coh3: cd08759, Coh: PF00963). For example, the protein encoded by the *scaA* gene in *R. flavefaciens* strain 007c has been reported to contain four cohesins and a C-terminal dockerin domain when its primary sequence was searched using BLAST and an unknown query sequence [53]. Instead, three of the modules are annotated as members of the CBM2/3 superfamily (SSF49384) and the fourth is not assigned to a protein family (**Table S3**). Furthermore, in other *Ruminococcus* scaffoldins previously reported cohesins are identified by InterProScan as G3DSA:2.60.40.680 superfamily members (CATH-3D) and not cohesins [57]. To investigate this issue, we used AlphaFold2 to predict the atomic structures of scaffoldins within the *R. champanellensis* and *R. flavefaciens* genomes that contained cohesins that could not be identified by InterProScan. The predicted

structures of these ‘undetected’ domains were compared to experimentally determined cohesin structures (Coh1 (PDB:1OHZ), Coh2 (PDB:2BM3) and Coh3 (PDB:2ZF9) cohesins) and their similarity determined by calculating a template-modeling (TM) score [58]. In all organisms, AlphaFold2-based predictions identified additional cohesin domains within scaffoldins that could not be detected by InterProScan (**Table S3**). Given that we under-counted the cohesin domains in the three Ruminococcus species above, we extended the structural analysis to three additional *R. flavefaciens* strains (AE3010, SAb67, YL228) that were identified through our analysis. There is no literature reported cohesin domains for these strains. In *R. flavefaciens* strain AE3010, only three cohesin containing proteins (7 cohesin domains) were detected using InterProScan but based on AlphaFold2 a total of 18 domains (across 11 proteins) with cohesin folds are detected (summarized in **Table S3, Figure S2**). Collectively, these results suggest that more complete genomic sequencing of *Ruminococcus* genomes is needed to fully define their scaffoldin complement and it highlights the utility of employing structure-based approaches to identify their cohesins.

4.4 Discussion

Harnessing the potent cellulolytic activity of cellulosome producing microbes could lead to improved methods to convert abundant LCB into renewable chemicals and materials. To gain insight into their structures and distribution in nature, we searched 305,693 prokaryotic genomes for genes that encode cellulosome components – at least one multi-cohesin containing protein and one DocGH enzyme. This analysis identified 33 bacterial species that likely produce conventional cellulosomes that resemble those present in *C. acetobutylicum* and *A. cellulolyticus*, as well as 4 species that produce scaffoldin-containing structures that could bind DocGH enzymes and other dockerin-fusion proteins. The cellulosomes within the 33 species can be classified as having either complex or simple structures following the convention established by Bayer and colleagues (**Fig. 1, Tables 1,2**) [7]. With only two exceptions, these microbes originate from four genera of gram-positive bacteria: *Acetivibrio, Ruminococcus*, *Ruminiclostridium*, and *Clostridium* (**Fig. 5**)**.** Based on their predicted complement of scaffoldin proteins, ten species produce complex cellulosomes that are related to the one in *A. cellulolyticus,* while the remaining 23 species may produce less complex (simple) structures that resemble *C. acetobutylicum*'s cellulosome (**Fig. 1**). We expect the cellulosomes in these bacteria to exhibit varying levels of cellulolytic activity based on their DocGH-LCB profiles (**Fig. 4, Table S4**). As documented in **Table 1**, all bacteria with the genetic capacity to produce complex cellulosomes also contain a large number of genes that encode DocGH-LCB proteins (called "high DocGH-LCB" microbes) and cellulose binding CBM modules suggesting that they are cellulolytic. In contrast, the genomes of bacteria that display simple cellulosomes can either have high or low numbers of DocGH-LCB encoding genes, implying differences in their cellulolytic activities (**Table 2**). To the best of our knowledge, over a quarter of the cellulosome producing species discovered in this search (10 total) have not been previously described in the literature. Notably, although *R. cellobioparum* (*subsp. termitidis*) [6] and *R. bromii* [59] have

previously been reported to produce cellulosomes, the sequenced genomes for these microbes lack genes for a multi-cohesin containing protein and therefore were not classified by us as containing a bona fide cellulosome. Below we summarize these findings.

4.4.1 Complex cellulosome producing bacteria.

A number of bacterial species have been shown to display multi-scaffoldin containing complexes that have been referred to as “complex”/ “highly-structured” cellulosomes [6,7]. Here we broadly define a complex cellulosome as containing at least two scaffoldins, a cell wall associated anchoring scaffoldin that contains a cohesin domain that could potentially bind to a dockerin domain located within a second multi-cohesin containing scaffoldin (**Fig. 1**). Using this definition complex cellulosomes contain either a complex primary scaffoldin (multi-cohesin, dockerin domain, and internal CBM3 domain containing) or a polyvalent adaptor (multi-cohesin and dockerin domain containing) that has the potential to bind via cohesin-dockerin interactions to the cell-surface by interacting with an anchoring scaffoldin (either a SLH- or nonSLH-anchoring scaffoldin) (**Fig. 1**). Based on this definition several *Acetivibrio* and *Ruminococcus* bacterial species produce complex cellulosomes (Table 1). *Acetivibrio* bacteria produce “classical” complex cellulosomes related the prototypical cellulosome from *A. thermocellus* in which DocGH enzymes bind to a complex primary scaffoldin that is tethered to the cell-surface via dockerin-cohesin interactions with a SLH-anchoring scaffoldin (**Fig. 1**), whereas in *Ruminococcus* bacteria the enzymes bind to a polyvalent adaptor scaffoldin that form dockerin-cohesin interactions with a nonSLH-anchoring protein. In all instances, the microbes also produce a high number of DocGH-LCB enzymes (>20) suggesting that the primary function of their complex cellulosomes is to degrade LCB (**Fig. 4**).

4.4.2 “Classical” complex cellulosomes in Acetivibrio species and *P. cellulosolvens*

Related complex cellulosomes are produced by 7 species of *Acetivibrio* bacteria (*A. alkalicellulosi* [24], *A. cellulolyticus* [23], *A. clariflavus* [61], *A. mesophilus* [62], *A. saccincola* [63], *A. straminisolvens* [64], *A. thermocellus* [12]) and *P. cellulosolvens* [29]. In the *Acetivibrio* bacteria, their complex primary scaffoldins contain multiple Coh1 domains that bind to DocGH enzymes through Doc1-Coh1 interactions and to the cell-surface through Doc2-Coh2 interactions with an SLH anchoring scaffoldin [65]. *P. cellulosolvens* is the single exception, as it originates from the *Pseudobacteroides* genus and the roles of the dockerin-cohesin interactions are reversed (i.e. it uses Doc2/Coh2 and Doc1/Coh1 interactions to mediate DocGH and scaffoldin-scaffoldin binding, respectively) [66]. In all of these bacteria the primary complex scaffoldin follow a similar domain arrangement, it contains multiple cohesins, a C-terminal dockerin module, and an internal CBM3 module that presumably enables each microbe to adhere to cellulose. These bacteria also possess Coh2-containing SLH-anchoring scaffoldins that bind to the bacterium's peptidoglycan to coordinate the binding of the complex primary scaffoldin via its Doc2 domain. The lone exception is *A. straminisolvens*, which lacks a primary complex scaffoldin, but nevertheless contains a large number of other types of scaffoldins, including two SLH-anchoring scaffoldins. All of these bacteria are “high DocGH-LCB” producers and are characterized by the presence of a large number of accessory scaffoldins that increase the number of enzymes that can be incorporated into the cellulosome (**Table 1**). These accessory scaffoldins include monovalent adaptors, polyvalent adaptors, SLH anchoring, nonSLH anchoring, and cell-free scaffolding (**Fig. 1**). Both types of adaptor scaffoldins act to increase the size and complexity of the cellulosomes by extending the existing structure and allowing for type-switching within type specific cohesin-dockerin interactions [7]. Aside from traditional SLH anchoring proteins, these bacteria have cohesin-containing nonSLH-anchoring scaffoldins that contain Cu-Amine Oxidase-like domains associated with secondary cell wall polymers [30]. The nonSLH-anchoring scaffoldins contain either Coh1 or Coh2 modules, suggesting they respectively facilitate either individual DocGH enzyme or complex primary

scaffoldin binding to the cell surface. While not scaffoldins per se, *Acetivibrio* complex cellulosome producers also contain a large number of genes encoding CBM3-SLH fusion proteins that may function to tether the microbe to cellulose [60]. The genomes of these microbes also contain a large number of genes encoding LCB active GHs that are fused to cellulose-binding CBM modules (**Table 1**, CBM(cell)-GH-LCB) as compared to other types of cellulosome producers. The cellulosomes in *Acetivibrio* bacteria are also unique because unlike other microbes, they uniformly produce multi-cohesin-containing cell-free scaffoldins that presumably form higher-order structures containing DocGH-LCB enzymes that are secreted into the environment to degrade LCB [25,67].

P. cellulosolvens produces a complex cellulosome that is most closely related to those found in *Acetivibrio* species, as it contains SLH-anchoring scaffoldins that coordinate the binding of a complex primary scaffoldin bearing a C-terminal Doc1. Moreover, as compared to *Acetivibrio* and *Ruminococcus* bacteria, it is phylogenetically more closely related to *Acetivibrio* bacteria based on its 16S rRNA sequence (**Fig. 5**). It stands out as producing the most complex cellulosome, since its genome encodes genes for an astounding 79 cohesin domains that are distributed between 33 scaffoldins: three primary, ten monovalent adaptor, three polyvalent adaptor, eight SLH-anchoring, three nonSLH-anchoring, three cell-free scaffoldins, and three single cohesin domain containing proteins. *P. cellulosolvens* is also unique because it has genes encoding both a complex primary scaffoldin, as well as two primary scaffoldins that are typically found in simple cellulosomes within *Ruminiclostridium* and *Clostridium* bacteria (previously referred ScaM1 and ScaM2 in *P. cellulosolvens*) [29]. As noted previously, the usage of the cohesin and dockerin domains in *P. cellulosolvens* is reversed as compared to other bacteria in the *Acetivibrio* category.

Another conserved feature in *Acetivibrio* bacteria and *P. cellulosolvens* is the manner in which they control the expression of the DocGH enzymes and scaffoldin proteins to construct

their cellulosome. Microbes have been shown to alter the complement of their DocGH enzymes when different types of LCB substrates are encountered [61,68] by regulating gene expression using either two-component systems [69,70], selective RNA transcript stabilization [71,72], or transmembrane biomass-sensing RsgI-type anti-σ factors that regulate σ^{I}-factors [73-75]. Our analysis suggests that in *P. cellulosolvens* and all *Acetivibrio* complex cellulosome producing bacteria, RsgI-type anti-σ factors are used to dictate cellulosomal gene expression, as each species contains genes encoding 7 to 14 of these factors (**Table 1**).

4.4.3 Complex cellulosomes in Ruminococcus species

Previous studies have shown that *Ruminococcus flavefaciens* and *Ruminococcus champanellensis* produce cellulosomes, which like their *Acetivibrio* counterparts contain an array of primary, cell wall anchoring, and adaptor scaffoldins [7]. Prior studies highlighted several unique features in *Ruminococcus* cellulosomes. First, cell surface attachment in *Ruminococcus* cellulosomes is often mediated by an anchoring scaffoldin that is covalently linked to the cell wall by a sortase enzyme instead of by a SLH-anchoring scaffoldin as observed in *Acetivibrio* species [53,76]. Second, the primary scaffoldins in *Ruminococcus* species lacks a characteristic internal CBM3 that can mediate direct attachment to cellulosic substrates. Third, unlike *Acetivibrio* bacteria, these species contain a unique monovalent adaptor scaffoldin called ScaC which is often used as the genomic signature to identify *Ruminococcus* cellulosomes [56,77]. Lastly, it has been noted in the literature that the scaffoldins in the *Ruminococcus* cellulosomes contain dockerin and cohesin modules with divergent primary sequences (frequently Doc3- and Coh3-types) [54,78].

The scaffoldin proteins were identified in *R. champanellensis* (18P13) and *R. flavefaciens* (strains 17 and 007c) before their genomes were sequenced [53]. Surprisingly, we identified fewer cohesin domains and scaffoldin proteins encoded in these genomes than previously reported. In many cases this occurred because their sequenced genomes contain a

large number of sequence contigs that caused scaffoldin gene truncations (**Table S2**). In other instances, even if an intact, non-truncated scaffoldin encoding gene was present, it was not possible to detect the full complement of cohesins within the translated protein product using the sequence profiles employed by InterProScan (Coh1 (cd08548), Coh2 (cd08547), Coh3 (cd08759), and PFAM-identified cohesin (PF00963)). Indeed, only when AlphaFold2 was used to predict the atomic structures of these proteins was the full complement of previously reported cohesin modules identified. For example, an InterProScan analysis of translated genes in *R. flavefaciens* (strain 007c) identified a single scaffoldin, whereas 10 scaffoldins have been reported in the literature [53]. However, when AlphaFold2 was employed, 9 of these 10 scaffoldins were detected. Similar results were obtained when AlphaFold2 was applied to other ruminococcal genomes documented to contain genes for cellulosomes (summarized in **Table S3**). Interestingly, even though Coh3 domains are a signature feature of ruminococcal cellulosomes, InterProScan did not identify these domains in *R. champanellensis* (18P13) or *R. flavefaciens* (using the cd14255 profile for a Coh3 domain). We conclude that the sequence signatures employed by InterProScan are not sufficiently robust to identify Coh3 cohesins, consistent with the findings reported by Flint and colleagues who have subdivided *R. flavefaciens'* cohesins into 6 different groups based on sequence homology [54]. Our results also demonstrate the utility of using AlphaFold2 structure predictions to identify cohesins with divergent primary sequences.

4.4.4 Acetivibrio mesophilus may produce a "classical" complex cellulosome

Our analysis identified a previously unrecognized bacterium as a "classical" complex cellulosome producer, *Acetivibrio mesophilus* (N2K1) (formerly known as *Hungateiclostridium mesophilum*) [79]. This gram-positive anaerobic bacterium was first isolated from a mesophile consortium in a biogas fermenter fed with maize silage [62]. It has an optimal growth temperature of 45°C and expresses two hemicellulases that have been biochemically

characterized, but the presence of a cellulosome has not been reported to the best of our knowledge [80]. Its cellulosome is likely cellulolytic because its genome contains 40 genes encoding DocGH-LCBs. Based on our analysis, *A. mesophilus*' cellulosome is strikingly similar to the archetypal cellulosome produced by *A. thermocellus* (**Fig. 6**). Specifically, both species produce a "classical" complex primary scaffoldin (ScaA-like: WP_128706406.1) that contains nine Coh1 modules, an internal CBM3, and a C-terminal Doc2 module for cell surface attachment via interactions with a SLH-anchoring scaffoldin (ScaF-like: WP_128705811.1). *A. mesophilus* has an additional three genes for SLH-anchoring scaffoldins (ScaB-like: WP_235832675.1, ScaC-like: PROKKA_02165, and ScaD-like: WP_069196093.1) that are related to *A. thermocellus*' ScaB, ScaC, and ScaD scaffoldins, as well as one nonSLH-anchoring scaffoldin (ScaG-like: WP_235832552.1) that is similar to ScaG. *A. mesophilus* also produces an additional scaffoldin that closely resembles scaffoldins found in *C. alkalicellulosi* (ScaO2) [24] and *A. cellulolyticus* (ScaO) [30]. This scaffoldin (WP_128706311.1) contains a C-terminal Coh1-Doc1 bi-domain unit, three Fibronectin-type III (FN3) repeats, as well as S8-peptidase-like and galactose oxidase-like domains of unknown function. Given the presence of a C-terminal Doc1 domain in this scaffoldin, it is tempting to speculate that it binds to the primary scaffoldin and/or the ScaD-like and ScaG-like scaffoldins that contain complementary Coh1 modules.

4.4.5 Bacteria that produce simple cellulosomes

We identified 23 species of anaerobic mesophilic bacteria that display less complex cellulosomes and fewer types of scaffoldins. The cellulosomes in these microbes always contain a simple primary scaffoldin that houses several Coh1 domains and an N-terminal CBM3 module (**Fig. 1**). In all cases the gene encoding the primary scaffoldin protein (*cipA*) is located within a *cipA* operon that also contains genes for DocGH enzymes. The vast majority of DocGHs in these microbes contain complementary Doc1 modules for direct interaction with the simple

primary scaffoldin. The lone exception is *Clostridium sp. HBUAS56017*, whose simple primary scaffoldin lacks a CBM.

Simple cellulosome producing bacteria originate from the *Ruminiclostridium* and *Clostridium* genera (**Figs. 5**). The *Ruminiclostridium* bacteria produce simple cellulosomes that likely have high cellulolytic activity as their genomes contain a large number of DocGH-LCB genes (*R. hungatei* [81], *R. papyrosolvens* [82], *R. josui* [83], *R. herbifermentans* [84], *R. cellulolyticum* [85], and *R. sufflavum* [86]) (**Fig. 4**). In contrast, nearly all of the clostridial species are low DocGH-LCB producers suggesting their simple cellulosomes have limited cellulolytic activity (*C. felsineum* (previously known as *C. roseum*) [87], *C. acetobutylicum* [50,88,89], *C. bornimense* [90], *C. cibarium* [91], *C. puniceum* [6], *C. saccharoperbutylacetonicum* [92], and several additional *Clostridium* sp. (CT7, HBUAS56017, DSM 8431, NJ4, TW13, YIM B02555)). This notion is consistent with experimental data, as the simple cellulosomes produced by high number DocGH-LCB bacteria, *R. cellulolyticum, R. papyrosolvens, and R. herbifermentans,* are shown to have potent cellulolytic activity [82,84,93,94], whereas the simple cellulosomes in *C. acetobutylicum* and *C. saccharoperbutylacetonicum* that have a low number DocGH-LCB genes are less cellulolytic [50,88,89,92]. The only two exceptions to the idea that simple cellulosomes in clostridial bacteria are less cellulolytic are from *C. cellulovorans* [95] and *Clostridium* sp. BNL1100 [96]. These species are presumably cellulolytic as their genomes contain a high number of DocGH-LCB encoding genes. However, it is noteworthy that based on its 16S rRNA sequence, *Clostridium* sp. BNL1100 can be classified as a member of the *Ruminiclostridium* genus (**Fig. 5**). Finally, there are two “non-clostridial” species with genomes encoding simple cellulosomes and a low number of DocGH-LCB enzymes, *Herbinix luporum* [97] and *Inconstantimicrobium porci* [98]. Based on their 16S rRNA sequence both species are most closely related to clostridial bacteria.

The mechanism(s) through which simple cellulosomes are attached to the cell surface remains incompletely understood. This is because their primary simple scaffoldins lack obvious cell wall binding modules or dockerin domains that could mediate their attachment to anchoring scaffoldins (either SLH-anchoring or nonSLH-anchoring scaffoldins) (**Fig. 1**). Cell wall attachment by simple primary scaffoldin may be mediated by their N-terminal CBM3 domain as supported by recent Western blot data demonstrating cell surface binding by a CBM3a module [31]. In addition, all of the simple primary scaffoldins contain X2 domains that have been implicated in both cell wall attachment [99] and cellular interactions with cellulose [31,100]. Notably, two bacterial species contain genes for more than one primary simple scaffoldin, *R. herbifermentans* and *C. cellulovorans* (**Table 2**). *R. herbifermentans'* genome has genes encoding four simple primary scaffoldins that are located in a single *cipA operon* [84], with each scaffoldin possessing 5 to 14 Coh1 modules. In the case of *C. cellulovorans*, genes for 3 primary scaffoldin proteins are present (WP_010073402.1, WP_010073403.1, and WP_013291799.1). Of these, only one scaffoldin (WP_013291799.1, CbpA) is encoded by a gene located within a *cipA* gene cluster [6].

Some species of simple cellulosome producing bacteria harbor genes encoding unique accessory scaffoldins with large numbers of FN3 domains that have been proposed to disrupt crystalline polysaccharide structures or solubilize large protein complexes [101,102]. Simple cellulosome displaying bacteria containing FN3 scaffoldins include: *R. cellulolyticum, R. herbifermentans, R. hungatei, C. josui, R. sufflavum, C. felsineum, C. pasteurianum,* and *Clostridium* sp. CT7. *C. pasteurianum* and *Clostridium* sp. CT7 are notable as they each contain a scaffoldin with 30 FN3 repeats that is capped by a C-terminal Coh2-Doc2 module whose binding partners are unknown, as no other proteins in these microbes contain Doc2 or Coh2 modules. A limited number of complex cellulosome displaying bacteria also contain scaffoldins with FN3 domains, but typically only 1-3 copies of this module are present. Finally, three simple

cellulosome producing bacteria produce cell-free scaffoldins that may be secreted as they lack dockerin domains and cell wall binding modules (*R. sufflavum*, *Clostridium* sp. HBUAS56017, and *C. bornimense*).

4.4.6 Novel and previously uncharacterized simple cellulosome displaying bacteria

We identified 9 new bacterial species that based on their genome sequences produce simple cellulosomes: *Clostridium cibarium*, *Clostridium pasteurianum, Clostridium* sp. CT7, *Clostridium* sp. DSM 8431, *Clostridium* sp. HBUAS56017, *Clostridium* sp. NJ4, *Clostridium* sp. TW13, *Clostridium* sp. YIM B02555, and *I. porci*. **Figure S1** shows their predicted scaffoldins, while **Table S4** enumerates their dockerin, cohesin, and enzyme content. They are all low DocGH producers that phylogenetically cluster with clostridial bacteria (**Fig. 5**). Each contains at least one simple primary scaffoldin that is a hallmark of simple cellulosome producers – defined as a scaffoldin that contains two or more cohesin domains and an N-terminal CMB3 (a lone exception is *Clostridium* sp. CT7 that lacks an N-terminal CBM3). As with other simple cellulosome producers, their primary scaffoldins contain X2 and Coh1 modules and each microbe almost exclusively produces DocGH enzymes containing Doc1 dockerins. Several of these microorganisms also produce a single scaffoldin that encodes 1-2 cohesins which are either Coh1- or Coh2-type (*Clostridium* sp. NJ4*, Clostridium* sp. YIM B02555*, Clostridium* sp. HBUAS56017*, Clostridium* sp. DSM 8431*, Clostridium* sp. CT7, *Clostridium cibarium*).

Our analysis predicts for the first time the scaffoldin and enzyme composition in 4 microbes previously noted to produce cellulosomes: *R. hungatei*, *C. bornimense*, *C. felsineum*, and *H. luporum* (**Table S4**). All have the genomic capacity to produce simple primary scaffoldins that are the core of a simple cellulosome (**Fig. 1**). *R. hungatei* (DSM 14427) is of particular interest as it is the only one in this group whose genome contains a high number of DocGH-LCB encoding genes, as well as genes encoding 3 accessory scaffoldins that contain 7-9 FN3 modules and a C-terminal cohesin domain. Based on their primary sequences, 2 of these

scaffoldins contain Coh2 domains whose binding partners are not known because *R. hungatei*'s genome lacks genes that encode for Doc2 containing proteins. Finally, the genomes of *H. luporum* and *C. cibarium* encode for proteins that may function as nonSLH-anchoring scaffoldins, as they contain a single Coh2 domain and a C-terminal transmembrane helix that may be embedded in the bilayer. However, the binding partners for these scaffoldins also remain unclear, since only in *C. cibarium* are genes encoding complementary Doc2 containing proteins identifiable.

4.4.7 "Scaffoldin-containing" microbes

We used broad search criteria to identify cellulosome producing bacteria – any genome that contained a gene for at least one multi-cohesin and one DocGH protein. Four microbial genomes barely satisfied these criteria and are unlikely to produce conventional cellulosomes because their largest scaffoldin contains only two cohesin modules. These include three species from the gram-positive *Bacillota* phylum whose members are known to display cellulosomes, *Iocasia fonsfrigidae* [103]*, Lachnoclostridium* sp. MSJ-17*,* and *Paenibacillus guangzhouensis* [104], as well as *Limihaloglobus sulfuriphilus* [105] which is a member of the rare Planctomycetes–Verrucomicrobia–Chlamydiae (PVC) superphylum. *Lachnoclostridium* sp. MSJ-17 encodes four cohesin-containing scaffoldins, three proteins that contain a single cohesin, and one larger scaffoldin that contains two cohesins and an N-terminal Doc1 domains. Two of the single cohesin-containing scaffoldins may be cell-associated as they contain C-terminal transmembrane helices (WP_216523161.1 and WP_216522914.1). Collectively, these scaffoldins could bind as many as 39 distinct dockerin-containing proteins, but the microbe is presumably non-cellulolytic as its genome encodes only a single DocGH-LCB. *I. fonsfrigidae* was isolated from deep sea sediment and has the potential to produce two large scaffoldins. One of them contains two cohesins, an FN3 module, and a CBM3 domain that could mediate cellulose binding. The second scaffoldin is also sizable (662 amino acids) but is predicted to

contain only a single C-terminal cohesin domain. Interestingly, this microbe's genome contains only a single gene encoding one DocGH-LCB enzyme. The gram-positive soil bacterium *P. guangzhouensis* is the most impressive of the scaffoldin-containing microbes as its genome encodes 35 cohesin-containing proteins. A total of 28 of these proteins are monovalent adaptors that contain a single dockerin-cohesin domain pair and in many instances, they possess an additional N-terminal GH enzyme that could degrade hemicellulose or oligosaccharide polymers (GH2, GH20, GH29, GH31, and GH43). There are also five SLH-anchoring scaffoldins to which these proteins could dock onto, potentially creating an enzyme rich surface that would be architecturally distinct from conventional cellulosomes [24]. Fascinatingly, in this microbe, nearly all of its dockerin domains are located within its cohesin-containing scaffoldins (only one dockerin is located in a non-scaffoldin protein). This raises the possibility that the full-complement of its dockerin proteins were not detected by InterProScan and/or its scaffoldins act to opportunistically scavenge dockerin-fusion proteins that are produced by other microbes. Notably, *P. guangzhouensis'* genome also contains a large number of genes that encode for CAZymes, but most of these are not of the family-type that is known to degrade LCB. The genome in the PVC bacterium *L. sulfuriphilus* contains genes encoding a single Coh2-Doc1-Coh1 scaffoldin-like protein and 13 dockerin-fusion proteins. It is presumably non-cellulolytic as it contains only a single DocGH-LCB and a limited number of CAZymes. Notably, the dockerin proteins in *L. sulfuriphilus* are fused to domains not commonly found in cellulosome producing bacteria, including FAD/NAD-binding, aspartic peptidase-like domains, and HdrA-like domains.

4.4.8 Phylogenetic variation in dockerins and CAZymes

An examination of their dockerin-fusion proteins provides insight into both the numbers and types of proteins that are incorporated into cellulosomes (**Fig. 7**). In general, the genomes of *Acetivibrio* and *Ruminococcus* bacteria produce complex cellulosomes containing a high number of DocGH-LCB genes suggesting that these structures are cellulolytic (**Fig. 5**). In

contrast, *Ruminiclostridium* and *Clostridium* bacteria that have the genomic capacity to produce simple cellulosomes contain either high or low numbers of DocGH-LCB genes, respectively. Interestingly, only bacteria whose genomes contain a large and diverse set of DocGH-LCBs (*Ruminiclostridium*, *Acetivibrio* and *Ruminococcus*) also possess a high number of genes that encode for other types of dockerin-fusion proteins (**Fig. 7**). For example, on average high DocGH-LCB producers contain ~40 genes encoding DocGH-LCBs and an impressive ~80 genes that encode other types of dockerin-fusion proteins. This number is much smaller in the low DocGH-LCB producers (*Clostridium*), as they only contain on average ~2 and ~3 genes that encode DocGH-LCBs and other types of dockerin-fusions, respectively. There is significant variability amongst the high DocGH-LCB producing microbes, as *P. cellulosolvens* (classified by us as an *Acetivibrio*-type) contains a total of 206 dockerin-fusion genes (of which 70 are DocGH-LCB genes), whereas *R. hungatei* (*Ruminiclostridium*-type) contains only 40 (20 DocGH-LCB genes). Interestingly, our analysis reveals that bacteria frequently supplement their DocGH-LCBs with a similar set of dockerin-fused accessory proteins that may facilitate LCB degradation. These include dockerin-fusion proteins containing: CBMs that bind carbohydrates (Doc-CBM, orange), carbohydrate active hydrolases such as pectin lyases and carbohydrate esterases (Doc-CAZymes, red), cohesin domains that are part of scaffoldins that construct the cellulosome (Doc-scaffoldin, yellow), and proteins with other functions (Doc-other, gray) (**Fig. 7**). This enrichment supports the idea that the primary function of the cellulosomes in *Acetivibrio, Ruminococcus*, and *Ruminiclostridium* bacteria is to degrade LCB or related carbohydrate polymers. Interestingly, perhaps to compensate for their deficiency in DocGH-LCBs and complementary proteins, the genomes of some clostridial bacteria contain larger numbers of genes encoding carbohydrate active hydrolases that are freely secreted (**Fig. 7**).

4.5 Conclusion

This comparative genomic analysis identified 33 bacterial species with the capacity to produce cellulosomes, of which 10 species have not been previously reported. With only a few exceptions their cellulosomes can be classified into four phylogenetically conserved families: *Acetivibrio-, Ruminococcus-, Ruminiclostridium-* and *Clostridium*-type cellulosomes. *Acetivibrio*-type cellulosomes, including the one found in *P. cellulosolvens*, are complex and can be populated with a high number distinct LCB active enzymes. They are characterized by the presence of a conserved dockerin-containing primary scaffoldin, SLH-anchoring scaffoldins that tether the cellulosome to the cell surface, cell-free scaffoldins that are presumably secreted to form multi-enzyme complexes, and CBM3-SLH proteins that may enable microbial tethering to LCB. Dockerin-fused enzymes bind to the scaffoldins via Doc1-Coh1 interactions, whereas Doc2-Coh2 interactions mediate scaffoldin-scaffoldin interactions (the exception is *P. cellulosolvens* in which these interactions are reversed). Only microbes that harbor *Acetivibrio*-type cellulosomes use a dedicated suite of polysaccharide-sensing RsgI transmembrane receptors to regulate its composition. *Ruminococcus* bacteria (*R. champanellensis* or *R. flavefaciens*) can also produce complex cellulosomes that contain a high number of distinct DocGH-LCB enzymes. However, they are unique because their anchoring scaffoldins are attached to the cell wall via sortase enzymes instead of SLH domains, and many of their dockerin-fusion proteins contain Doc3 modules. It was challenging to define the components of these structures from genomic sequence data as their cohesins frequently have divergent primary sequences that could not be detected using InterProScan. Indeed, we detected no Coh3-type modules based on their primary sequence, and only when AlphaFold2 was employed to predict their structures were several cohesins identified. *Ruminiclostridium* and *Clostridium* bacteria produce ‘simple’ cellulosomes that contain only a limited number of scaffoldins. They are further distinguished by the presence of a primary scaffoldin that contains an N-terminal CBM3, X2, and multiple cohesin domains. Their genomes encode only a limited number of scaffoldins and their primary scaffoldins adhere to the microbial surface through a poorly

understand mechanism as they lack obvious domains that are capable of binding to the cell wall. These simple cellulosomes can be subdivided further by the number of distinct DocGH-LCB enzymes they house, with *Ruminiclostridium* and *Clostridium* genomes typically encoding for high and low numbers of DocGH-LCB enzymes, respectively. Finally, several species of simple cellulosome displaying bacteria are unique as they contain scaffoldins harboring up to 30 FN3 repeats that have been proposed to disrupt crystalline polysaccharide structures and/or solubilize large protein complexes. Collectively, the results of this study provide broad insight into the structural diversity of bacterial cellulosomes, and they reveal conserved architectural features that may be useful in guiding ongoing engineering efforts to produce bio-based chemicals and materials from plant biomass.

4.6 Materials and Methods

4.6.1 Genome-based search to identify cellulosome displaying bacteria.

The retrieval of cellulosome-displaying bacteria consists of two components: i) a prescan phase to select for genomes that potentially contain cellulosomes, and ii) an in-depth scan of each of these filtered genomes. For the prescan phase, the RefSeq database [106] (July 10, 2023 release) was used, which contained 138,491 organisms with 371,291,248 records. From these, we retrieved only prokaryotic genomes (305,693) and performed a Hidden Markov model (HMM) search (hmmer.org) using dockerin (cd14256, cd14254, cd14255), cohesin (cd08548, cd08547, cd08759), and GHs (PF00150, PF03537, PF01670, PF12891, PF02015, PF02011, PF01270, PF00759) domain profiles obtained from the CDD and PFAM database [44,45]. Genomes were targeted for the next in-depth phase only if they met the following criteria: i) they must have at least one dockerin-fused GH protein, and ii) they must have at least one protein with two or more cohesins.

In the in-depth scanning phase, the 139 genomes from the prescan phase were re-annotated using Prokka (v. 1.14.6) [49], and all the locus tags from NCBI's annotations were mapped using BLAST (v. 2.13.0+) [107] to maintain consistency with the existing gene naming convention. After re-annotation, InterProScan (v. 5.59-91.0) [108] was performed, which includes the following member databases: Phobius (v. 1.01) [41], SUPERFAMILY (v. 1.75) [109], ProSiteProfiles (v. 2022_01) [43], SMART (v. 7.1) [110], CDD (v. 3.18) [44], PRINTS (v. 42.0) [111], and Pfam (v. 35.0) [45]. Additionally, three databases were added to capture signal peptide regions (SignalP v. 6.0) [46], transmembrane regions (DeepTMHMM v. 1.0.19) [47], and subcellular localization information (Psortb v. 3.0) [48].

After the in-depth genome analysis, we applied the same criteria as before (i) ≥1 DocGH fusion protein and ii) at least one protein with ≥2 cohesin domains) and identified 37 genomes

that met our criteria. 132 genomes did not pass our criteria using the InterProScan data because the initial HMMER search was not conducted with stringent threshold values, which inflated the initial number of domains each genome had. However, this genome-based analysis procedure did not identify several bacteria which have previously been shown to display cellulosomes (e.g. *Ruminococcus flavefaciens*-007c, -FD-1, -007c, and *Ruminococcus champanellensis*-18P13), presumably because RefSeq uses stringent criteria for inclusion of genomes in their database and these genomes were of insufficient quality. These species genomes were manually added into our analysis for completeness' sake. A custom python (v. 3.11.5) script was used to analyze the domain composition in the scaffoldins and dockerin-fused proteins in the 37 bacterial species that produce cellulosomes. The domain identifiers are listed used for this analysis are listed in **Table S5**.

4.6.2 AlphaFold2 analysis to identify cohesin-containing proteins

We observed that 34 protein sequences (from 3 organisms: *R. flavefaciens 17, R. flavefaciens 007c,* and *R. champanellensis*) are reported in the literature to possess sequence homology to cohesin domains but we could not identify them as such using the HMM profiles provided by InterProScan [53]. Similarly, our analysis showed significantly lower numbers of scaffoldin proteins for additional *R. flavefaciens* strains (AE3010, SAb67, YL228) that we suspected also encode cohesin domains that are too divergent to be identified by the HMM profiles. Recognizing that structure is better conserved than sequence, we used AlphaFold2 [112] to build atomic models for any protein across these six organisms that contained at least one of the following domains: Coh1 (cd08548), Coh2 (cd08547), Coh3 (cd08759), general Coh (PF00963), CBM2/3 Superfamily (IPR008965), CATH3D entry (G3DSA:2.60.40.680). This resulted in the predictions of 90 putative cohesin-containing proteins. We requested three models to be built for each sequence and used only the model with highest pLDDT score for further structural analysis. For multi-domain proteins, we divided the model into individual

domains using the program UniDoc [113], producing 311 domains. The structure of each domain was compared to four reference models: Coh1 (PDB ID: 1OHZ), Coh2 (PDB ID: 1TYJ), Coh3 (PDB ID: 2FZ9), and CBM3 (PDB ID: 6UFW). Structural comparisons were performed with the program TMalign [58]. TM-scores for all comparisons were normalized to 140 residues to simplify comparison. Query domains with TM-scores greater than 0.50 were considered as matches to the reference model. If a domain scored higher than 0.50 for multiple reference models, then we attribute the domain identity to match the highest scoring reference model.

4.7 Figures

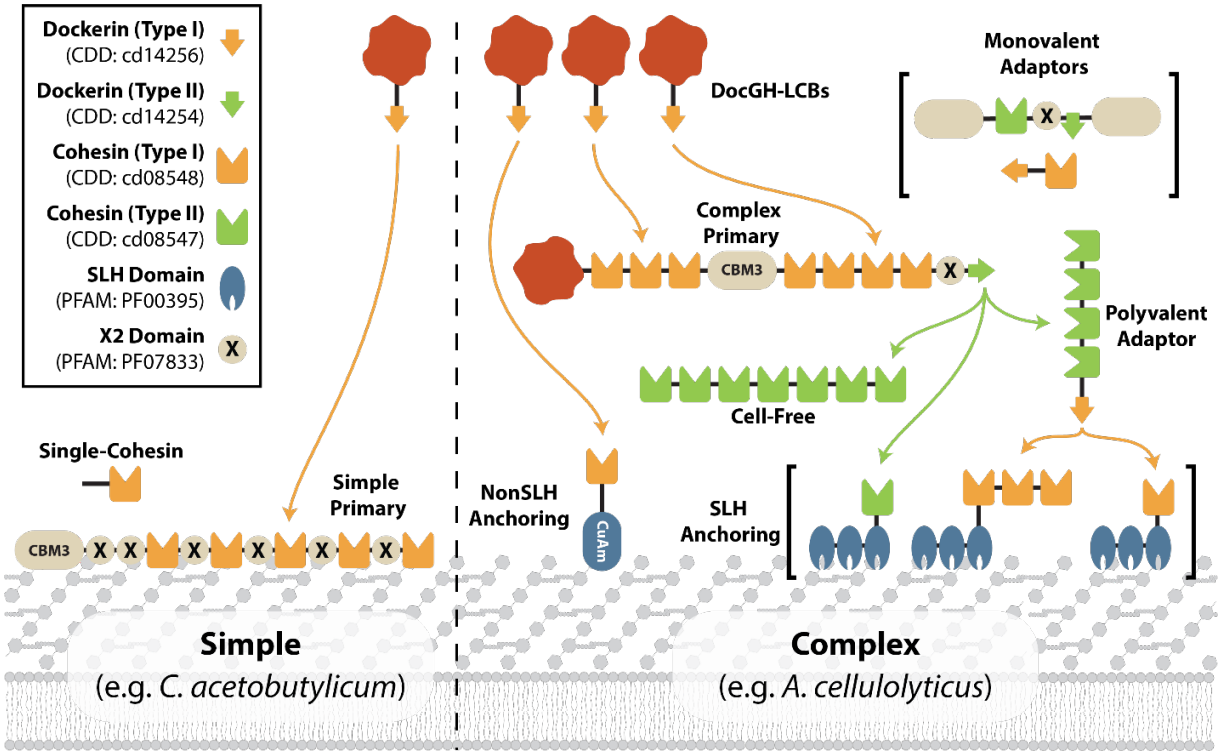


Figure 4.1 Representative scaffoldin compositions in simple and complex cellulosomes

Cartoon showing the different types of scaffoldins that are found in either simple (e.g. *C. acetobutylicum*) [33] or complex (e.g. *A. cellulolyticus*) [23,30] cellulosomes. Simple cellulosomes typically contain a simple primary scaffoldin and a small number of accessory scaffoldins. Complex cellulosomes contain a multi-cohesin complex primary scaffoldin with an associated anchoring scaffoldin, either a SLH or nonSLH-anchoring scaffoldin. Many complex cellulosomes also contain adaptor scaffoldins (polyvalent or monovalent) and cell-free scaffoldins that are presumably secreted.

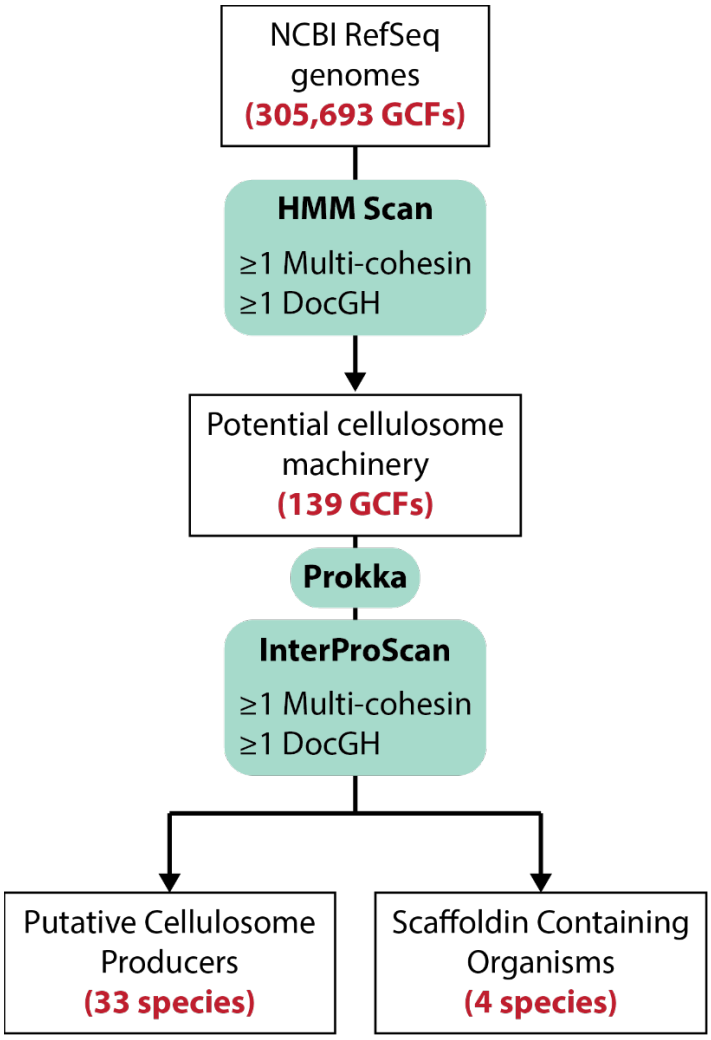


Figure 4.2 Strategy used to identify cellulosome producing microbes

More than 305k sequenced bacterial genomes within the NCBI Reference Sequence database (RefSeq). Translated protein-encoding genes were then initially searched using HMM profiles to identify genomes that contained cohesin and DocGH enzymes. A subset of those genomes were then submitted for *de novo* annotation via Prokka (v. 1.14.16) [49] followed by a more stringent analysis of these protein sequences using InterProScan (v. 5.59-9.10) [40] which identified 33 species that likely produce conventional cellulosomes based on the types of cohesin-containing proteins they possess. A total of 4 species contains multi-cohesin scaffoldins that are unlikely to assemble into conventional cellulosomes (Scaffoldin Containing Organisms).

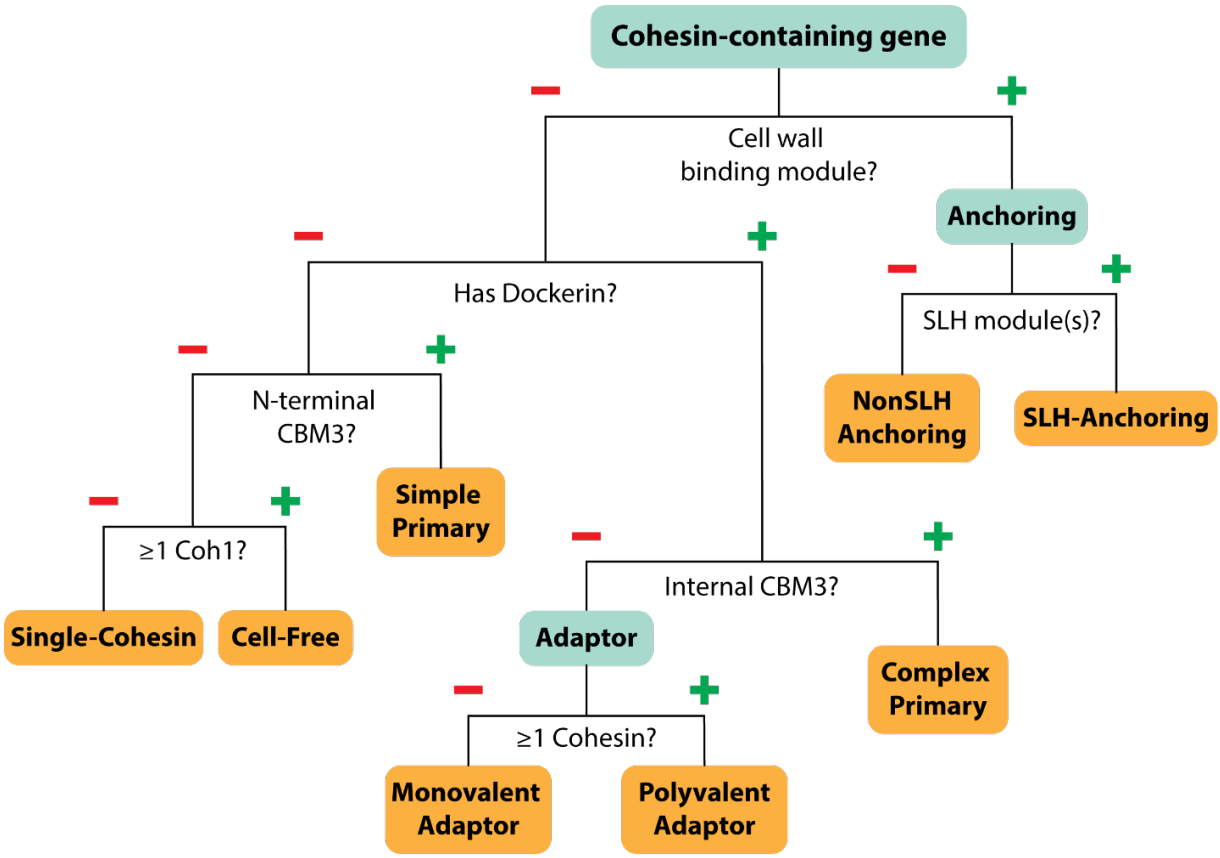


Figure 4.3 Scaffoldin categorization

Decision tree showing how each cohesin-containing protein was classified. Each cohesin-containing protein was classified into one of eight categories based on its type and abundance (Coh1, Coh2, Coh3), as well as the presence of additional domains such as dockerin, CBM3, and cell wall binding domains.

Table 4.1 Complex cellulosome producing bacteria

		Organism (GCF/Accession)	Temperature	Isolation Location	Scaffoldin										Dockerin		CAZY Profile				
					Simple Primary	Complex Primary	Monovalent Adaptor	Polyvalent Adaptor	SLH-Anchoring	NonSLH Anchoring	Cell-Free	Free Cohesin	Total Cohesins	Total Scaffoldin	Total Dockerin	DocGH-LCB	Free GH-LCB	Free CAZyme	CBM(cell)-GH-LCB	CBM(cell)-SLH	Rsgl
Complex (High Doc)	Acetivibrio	*Acetivibrio alkalicellulosi* (2509887034)	Mesophile	Soda lake	-	1	5	6	4	1	2	2	46	22	140	50	16	62	35	6	7
		Acetivibrio cellulolyticus (GCF_000179595.2)	Mesophile	Sewage Sludge	-	1	3	1	5	1	3	2	41	16	141	52	7	57	24	2	14
		Acetivibrio clariflavus (GCF_000237085.1)	Thermophile	Methanogenic sludge	-	1	1	2	4	1	4	-	51	13	81	34	9	44	19	2	10
		Acetivibrio mesophilus (GCF_004102745.1)	Thermophile (45°C)	Methanogenic sludge	-	1	1	-	4	1	1	1	30	9	74	41	9	45	26	4	9
		Acetivibrio saccincola (GCF_002844395.1)	Thermophile	Methanogenic sludge	-	1	1	-	5	2	3	1	31	14	65	34	10	45	24	2	10
		Acetivibrio straminisolvens (GCF_000521465.1)	Thermophile	-	-	-	1	-	2	2	3	3	18	12	75	30	34	78	18	1	8
		Acetivibrio thermocellus (GCF_000015865.1)	Thermophile	Sewage digester sludge	-	1	-	-	4	1	1	1	29	8	73	39	8	31	24	3	9
		Pseudobacteroides cellulosolvens (GCF_001262605.1)	Thermophile	Compost	2	1	10	3	8	3	3	3	79	33	206	70	29	101	43	7	14
	Ruminococcus	*Ruminococcus champanellensis* (GCF_000210095.1)	Mesophile	Human feces	-	-	4/6/6	-/1/3	-	-/1/1	-/1/-	2/3/2	7/18/21	7/11/12	65	24	8	30	11	-	-
		Ruminococcus flavefaciens	Mesophile	Rumen																	
		007c (GCF_000571935.1)			-	-	1/6/6	-/1/2	-	-/-/1	-	-/2/1	1/13/16	1/9/10	154	35	16	29	16	-	-
		17 (GCF_000247525.1)			-	-	1/5/5	-/1/2	-	-/-/1	-	-/2/3	1/8/21	1/8/11	135	25	13	31	10	-	-
		AE3010 (GCF_000526795.1)			-	-	1/5/-	-/1/-	-	-/3/-	1/1/-	1/3/-	7/18/-	3/13/-	161	48	11	23	22	-	-
		SAb67 (GCF_900116855.1)			-	-	2/3/-	-/3/-	-	-/1/-	1/-/-	1/4/-	10/17/-	4/11/-	121	33	11	33	17	-	-
		YL228 (GCF_900119075.1)			-	-	1/3/-	-/2/-	-	-/1/-	1/-/-	1/3/-	7/19/-	3/9/-	142	43	13	23	21	-	-

Tabulation of cellulosomal components for the complex cellulosome producing organisms identified in this study. The counts for the number of each scaffoldin type are reported (as defined in **Fig. 3**). Also reported are the number per organism for the following: individual cohesin domains (Total Cohesin), scaffoldin proteins (Total Scaffoldin), all dockerin-containing proteins (Total Dockerin), LCB-active dockerin-fused GH enzymes (DocGH-LCB), non-dockerin-fused LCB-active GH enzymes (FreeGH-LCB), non-dockerin-fused GH, CL and PE enzymes (including LCB active) (Free CAZyme) [60], LCB-active GH enzymes that are fused to a cellulose-binding CBM modules (CBM(cell)-GH-LCB), SLH-domains fused to a cellulose-binding CBM module (CBM(cell)-SLH), and transmembrane RsgI-like proteins (RsgI). A list of LCB active enzymes is provided in **Table S1**. For the *Ruminococcus* species, three numbers are reported for each scaffoldin type. They are the values obtained when InterProScan (left) or AlphaFold2 (middle) was used to identify their cohesin modules and the scaffoldin number reported previously in the literature (right). Bolded organisms indicate those which have not been previously reported as capable of producing a cellulosome.

Table 4.2 Simple cellulosome producing bacteria

	Organism (GCF/Accession)	Temperature	Isolation Location	Scaffoldin										Dockerin		CAZY Profile				RsgI
				Simple Primary	Complex Primary	Monovalent Adaptor	Polyvalent Adaptor	SLH-Anchoring	NonSLH-Anchoring	Cell-Free	Free Cohesin	Total Cohesins	Total Scaffoldin	Total Dockerin	DocGH-LCB	Free GH-LCB	Free CAZyme	CBM(cell)-GH-LCB	CBM(cell)-SLH	
Simple (High-Doc)	*Clostridium cellulovorans* (GCF_000145275.1)	Mesophile	Wood fermenter	3	-	-	-	-	1	-	7	18	11	43	27	31	84	16	-	1
	Clostridium sp. BNL1100 (GCF_000244875.1)	Mesophile	Corn stover	1	-	-	-	-	-	-	1	7	2	73	35	21	37	18	3	1
	Ruminiclostridium cellulolyticum (GCF_000022065.1)	Mesophile	Compost	1	-	-	-	-	-	-	2	10	3	62	31	16	35	17	4	1
	Ruminiclostridium herbifermentans (GCF_005473905.2)	Mesophile	Biogas reactor with maize silage	4	-	-	-	-	-	-	5	37	9	80	40	7	36	18	1	-
	Ruminiclostridium hungatei (GCF_002051585.1)	Mesophile	Soil	1	-	-	-	-	-	-	4	9	5	40	22	18	48	12	2	-
	Ruminiclostridium josui (GCF_000526495.1)	Mesophile	Compost	1	-	-	-	-	-	-	2	8	3	73	34	17	38	18	2	-
	Ruminiclostridium papyrosolvens (GCF_000175795.2)	Mesophile	Paper mill	1	-	1	-	-	-	-	1	8	3	72	33	16	53	19	4	1
	Ruminiclostridium sufflavum (GCF_003208175.1)	Mesophile	Methanogenic sludge	1	-	-	-	-	-	1	3	14	5	54	26	5	32	11	2	-
Simple (Low-Doc)	*Clostridium acetobutylicum* (GCF_000008765.1)	Mesophile	Cornmeal	1	-	-	-	-	-	-	1	6	2	10	9	22	84	5	-	1
	Clostridium bornimense (GCF_000577895.1)	Mesophile	Biogas reactor with maize silage	1	-	-	-	-	-	1	-	8	2	4	3	6	45	2	-	1
	Clostridium cibarium (GCF_014836335.1)	Mesophile	Chicken feces	1	-	-	-	-	1	-	-	5	2	17	8	27	104	9	-	2
	Clostridium felsineum (GCF_002006355.2)	Mesophile	Retting plant material	1	-	2	-	-	-	-	1	9	4	13	9	29	126	6	-	-
	Clostridium pasteurianum (GCF_001705235.1)	Mesophile	Propionate anaerobic reactor	1	-	1	-	-	-	-	1	6	3	10	8	16	112	7	-	-
	Clostridium puniceum (GCF_002006345.1)	Mesophile	Rotting potatoes	1	-	-	-	-	-	-	1	3	2	12	10	24	142	7	-	2
	Clostridium saccharoperbutylacetonicum (GCF_000340885.1)	Mesophile	Soil	1	-	-	-	-	-	-	1	3	2	8	7	31	165	7	-	2
	Clostridium sp. CT7 (GCF_002796935.1)	Mesophile	Soil	1	-	1	-	-	-	-	1	6	3	10	8	16	112	7	-	-
	Clostridium sp. DSM 8431 (GCF_900116755.1)	Mesophile	Bison rumen	1	-	-	-	-	1	-	-	7	2	11	10	11	71	6	-	1
	Clostridium sp. HBUAS56017 (GCF_007115085.1)	Mesophile	Fermented food	-	-	-	-	-	1	1	1	5	3	16	8	27	108	9	-	2
	Clostridium sp. NJ4 (GCF_003014985.1)	-	Soil	1	-	-	-	-	-	-	1	6	2	10	9	22	84	5	-	1
	Clostridium sp. TW13 (GCF_024345225.1)	-	Wastewater sludge	1	-	-	-	-	-	-	-	3	1	3	3	12	105	1	-	-
	Clostridium sp. YIM B02555 (GCF_022320955.1)	-	Root	1	-	-	-	-	-	-	1	3	2	8	7	27	162	6	-	2
	Herbinix luporum (GCF_012838165.1)	Thermophile	Thermophile biogas plant	1	-	-	-	-	1	-	-	6	2	7	4	20	29	14	-	-
	Inconstantimicrobium porci (GCF_029219665.1)	Mesophile	Pig feces	1	-	-	-	-	-	-	-	3	1	5	4	2	46	2	-	1
Scaffoldin Containing (Low Doc)	*Iocasia fonsfrigidae* (GCF_017751145.1)	Mesophile	Deep-sea sediment	-	-	-	-	-	-	1	1	3	2	1	1	8	45	3	-	-
	Lachnoclostridium sp. MSJ 17 (GCF_018918205.1)	-	Primate feces	-	-	-	1	-	-	-	4	6	5	39	1	-	18	-	-	-
	Limihaloglobus sulfuriphilus (GCF_001999965.1)	Mesophile	Sediment of evaporation pond	-	-	-	1	-	-	-	-	2	1	14	1	10	81	-	-	-
	Paenibacillus guangzhouensis (GCF_009363075.1)	Mesophile	Soil	-	-	28	-	5	-	-	2	36	35	29	1	10	189	5	3	-

Tabulation of the components within the simple cellulosome producing bacteria that respectively contain either high (shaded green) or low numbers of genes encoding DocGH-LCB (shaded orange) enzymes. The table also shows data for four species that contain scaffoldins but are unlikely to produce conventional cellulosomes (shaded red). The counts for the number of each scaffoldin type are reported (as defined in Fig. 1). The category definitions are identical to those presented in Table 1.

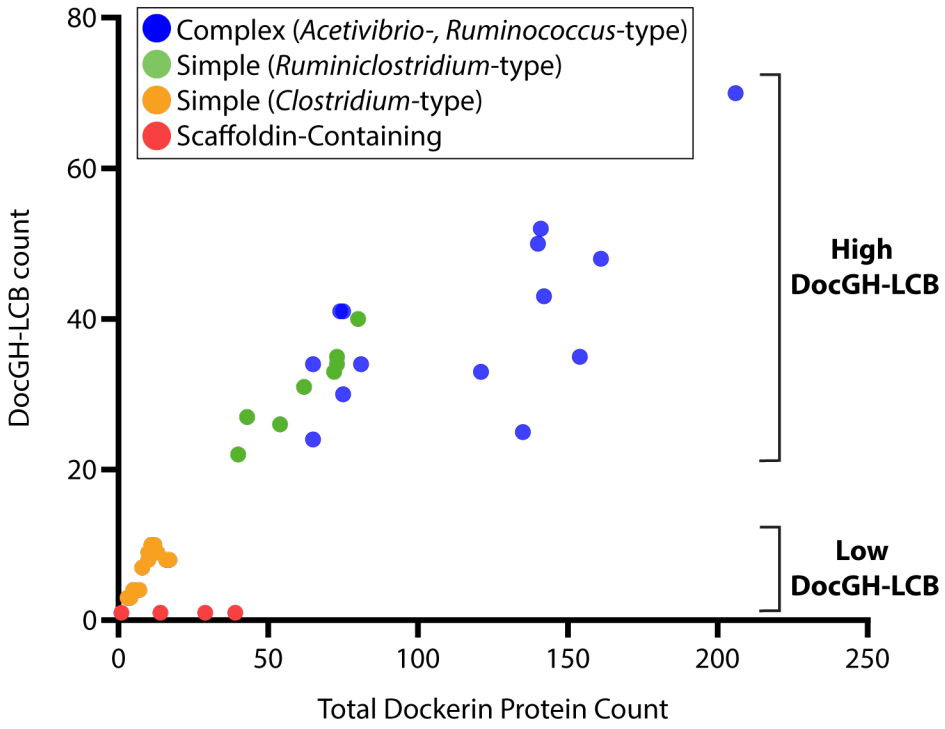


Figure 4.4 Dockerin-protein and DocGH-LCB distribution

Correlation between total dockerin protein count and DocGH-LCB protein count (R2= 0.81, p< 0.05). Reported here are the number of genes encoding dockerin-fusion and DocGH-LCB proteins for all microbes listed in Tables 1 and 2. Classes of microbes that are shown are indicated in the key and correspond to: Complex *Acetivibrio* and *Ruminococcus* cellulosomes (blue), Simple *Ruminiclostridium* cellulosomes (green), Simple *Clostridium* cellulosomes (orange), and bacteria containing scaffoldin encoding genes that are unlikely to produce conventional cellulosomes (red).

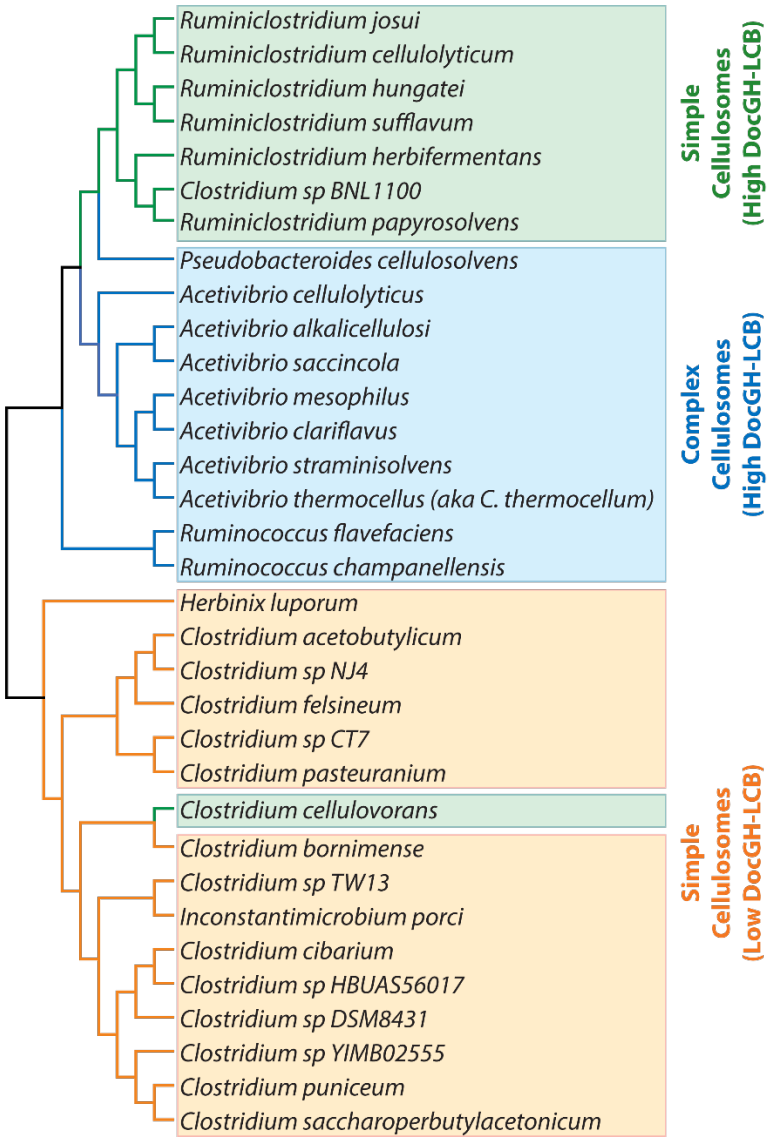


Figure 4.5 Taxonomic diversity of cellulosome producing bacteria

Phylogenetic tree of the bacteria listed in Tables 1-2 (excluding the “scaffolding-containing” organisms). Bacteria are colored according to their predicted cellulosome type and number of genes encoding DocGH-LCB enzymes: Complex cellulosomes with high numbers of DocGH-LCB genes (blue); Simple cellulosome producing bacteria with high DocGH-LCB gene counts (green); Simple cellulosome producing bacteria with a low number of DocGH-LCB genes (orange). 16S rRNA sequences were aligned with ClustalW and the tree was constructed using MEGA11.

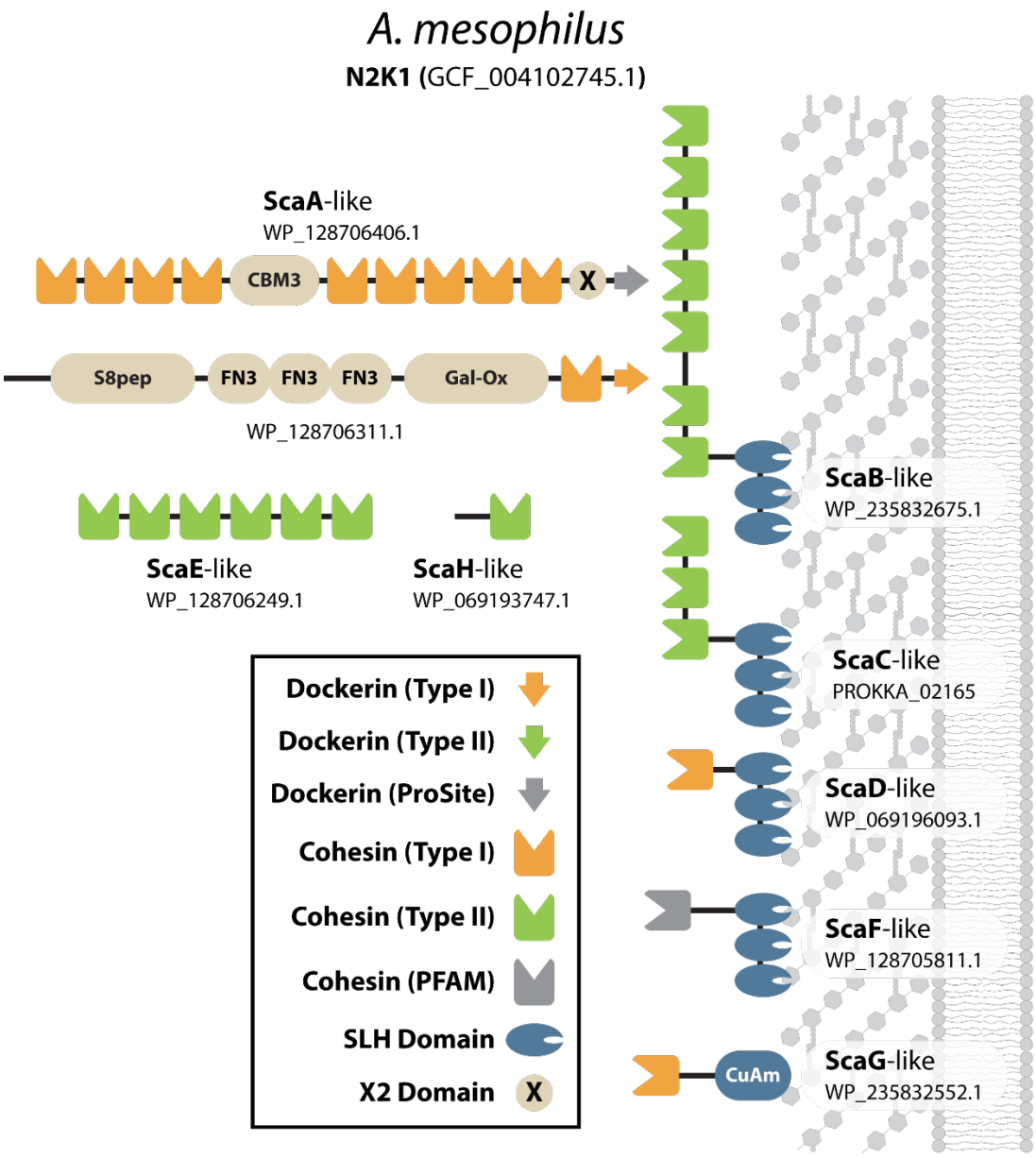


Figure 4.6 Scaffoldin proteins in *A. mesophilus* (N2K1)

Cartoon representation of all of the cohesin-containing proteins in *A. mesophilus* (N2K1). The domains they possess are defined in the key and also include: Fibronectin Type-III like (FN3), CBM type-3 domain (CBM3), Copper Amine Oxidase-Like (CuAm), S8-subtilase/peptidase (S8pep), and Galactose Oxidase-like (Gal-Ox) domains. Proteins were named based on their sequence homology to scaffoldins present in *A. thermocellus*.

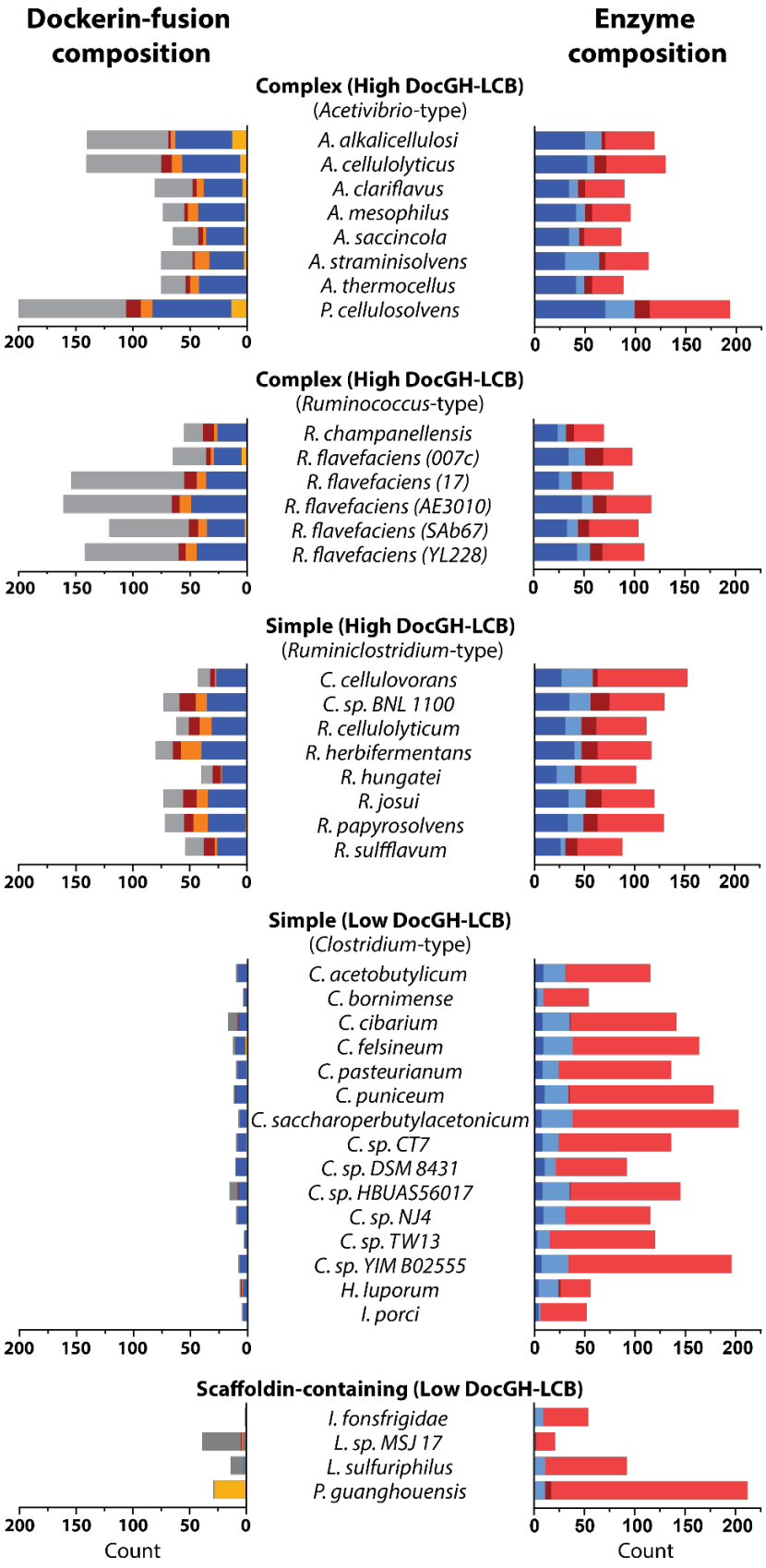

Dockerin-fusion composition
Enzyme composition
Complex (High DocGH-LCB)
(Acetivibrio-type)
A. alkalicellulosi
A. cellulolyticus
A. clariflavus
A. mesophilus
A. saccincola
A. straminisolvens
A. thermocellus
P. cellulosolvens
Complex (High DocGH-LCB)
(Ruminococcus-type)
R. champanellensis
R. flavefaciens (007c)
R. flavefaciens (17)
R. flavefaciens (AE3010)
R. flavefaciens (SAb67)
R. flavefaciens (YL228)
Simple (High DocGH-LCB)
(Ruminiclostridium-type)
C. cellulovorans
C. sp. BNL 1100
R. cellulolyticum
R. herbifermentans
R. hungatei
R. josui
R. papyrosolvens
R. sulfflavum
Simple (Low DocGH-LCB)
(Clostridium-type)
C. acetobutylicum
C. bornimense
C. cibarium
C. felsineum
C. pasteurianum
C. puniceum
C. saccharoperbutylacetonicum
C. sp. CT7
C. sp. DSM 8431
C. sp. HBUAS56017
C. sp. NJ4
C. sp. TW13
C. sp. YIM B02555
H. luporum
I. porci
Scaffoldin-containing (Low DocGH-LCB)
I. fonsfrigidae
L. sp. MSJ 17
L. sulfuriphilus
P. guanghouensis
200 150 100 50 0
0 50 100 150 200
Count
Count

Figure 4.7 Domain and enzyme composition in dockerin-fused proteins

Plots show the types of domains that are fused to dockerin proteins in microbes that produce *Acetivirbio*-type complex cellulosomes, *Ruminococcus*-type complex cellulosomes, *Ruminiclostridium*-type simple cellulosomes, *Clostridium*-type simple cellulosomes, and other bacteria with scaffoldin-containing proteins that are unlikely to form conventional cellulosomes (see Tables 1-2 for a complete list). (Left) Stacked bar plot representation of the number of dockerin-fusion proteins based on the types of domains they contain. Co-occurring domain type is color coded according to the following: Yellow: Cohesin, Dark Blue: Glycoside Hydrolase with LCB activity, Orange: Carbohydrate Binding Module, Red: Carbohydrate Active Enzymes, Grey: other. (Right) Stacked bar plot representation of cellulosomal enzyme composition for each cellulosome producing microbe, including; Dark Blue: DocGH-LCB, Light Blue: Free GH-LCB, Red: Dockerin fused CAZyme, and Light Red: Free CAZyme.

4.8 Supplementary Material

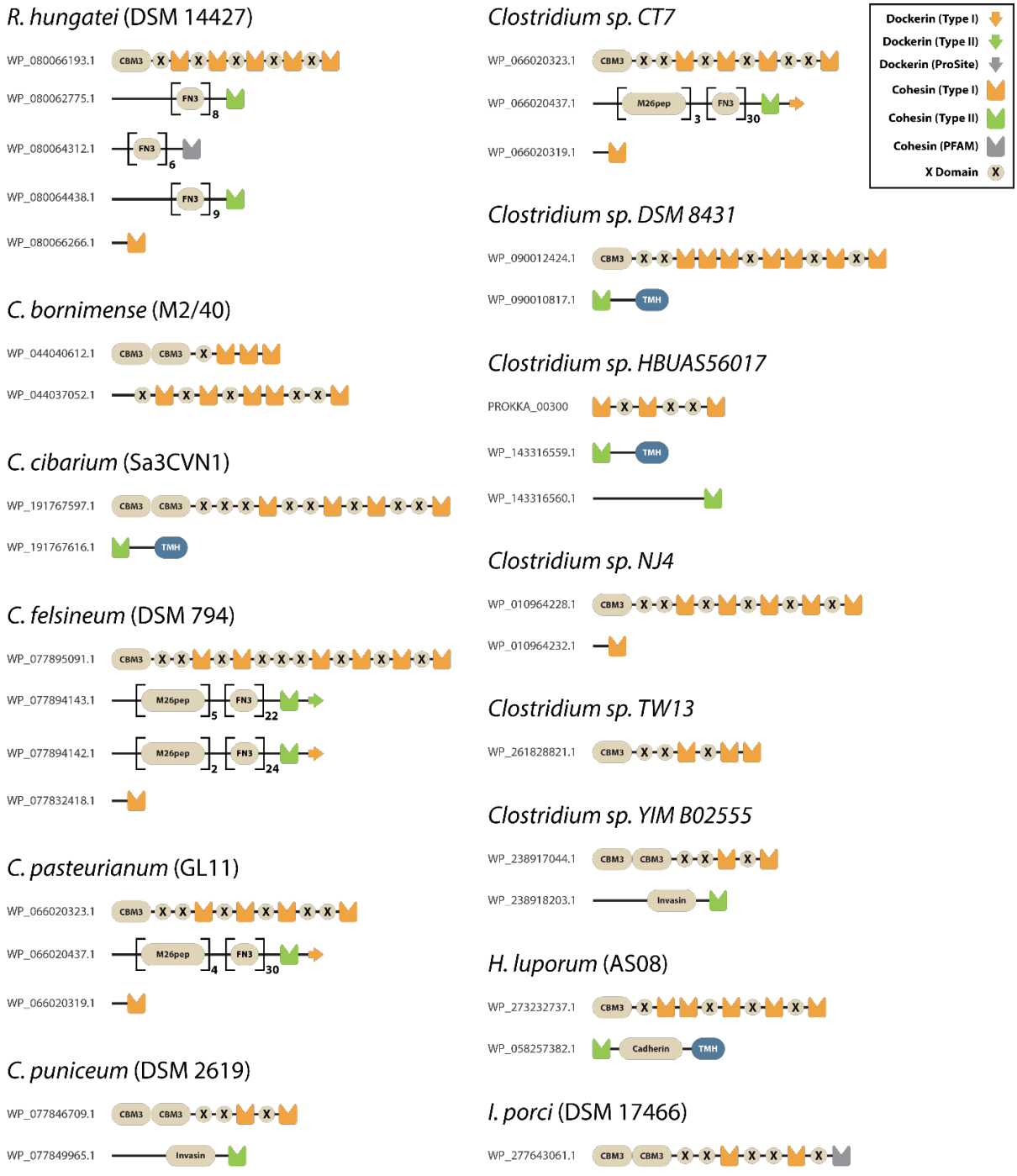


Figure S4.1 Schematic representation of unreported scaffoldins from known cellulosome producers

Cartoon representation showing the cohesin-containing proteins present within simple cellulosome producing organisms whose scaffoldin compositions have not been reported previously. Novel organisms identified in this study include: *Clostridium cibarium* (Sa3CVN1), *Clostridium pasteurianum* (GL11) ,*Clostridium sp. CT7*, *Clostridium sp. DSM 8431*, *Clostridium sp. HBUAS56017*, *Clostridium sp. NJ4*, *Clostridium sp. TW13*, *Clostridium sp. YIM B02555*, *Inconstantimicrobium porci* (DSM 17466). Domains are colored according to the upper right key with additional domains including: FN3 (beige)- Fibronectin type III (SM00060), CBM3 (beige)- Carbohydrate binding module family 3 (PF00942), M26pep (beige) - M26 Metallo-endopeptidase (PF05342), TMH (navy) - Transmembrane Helix, Invasin (beige)- Invasin/intimin cell-adhesion fragment (SSF49373), and Cadherin (beige) - Cadherin-like (PF12733).

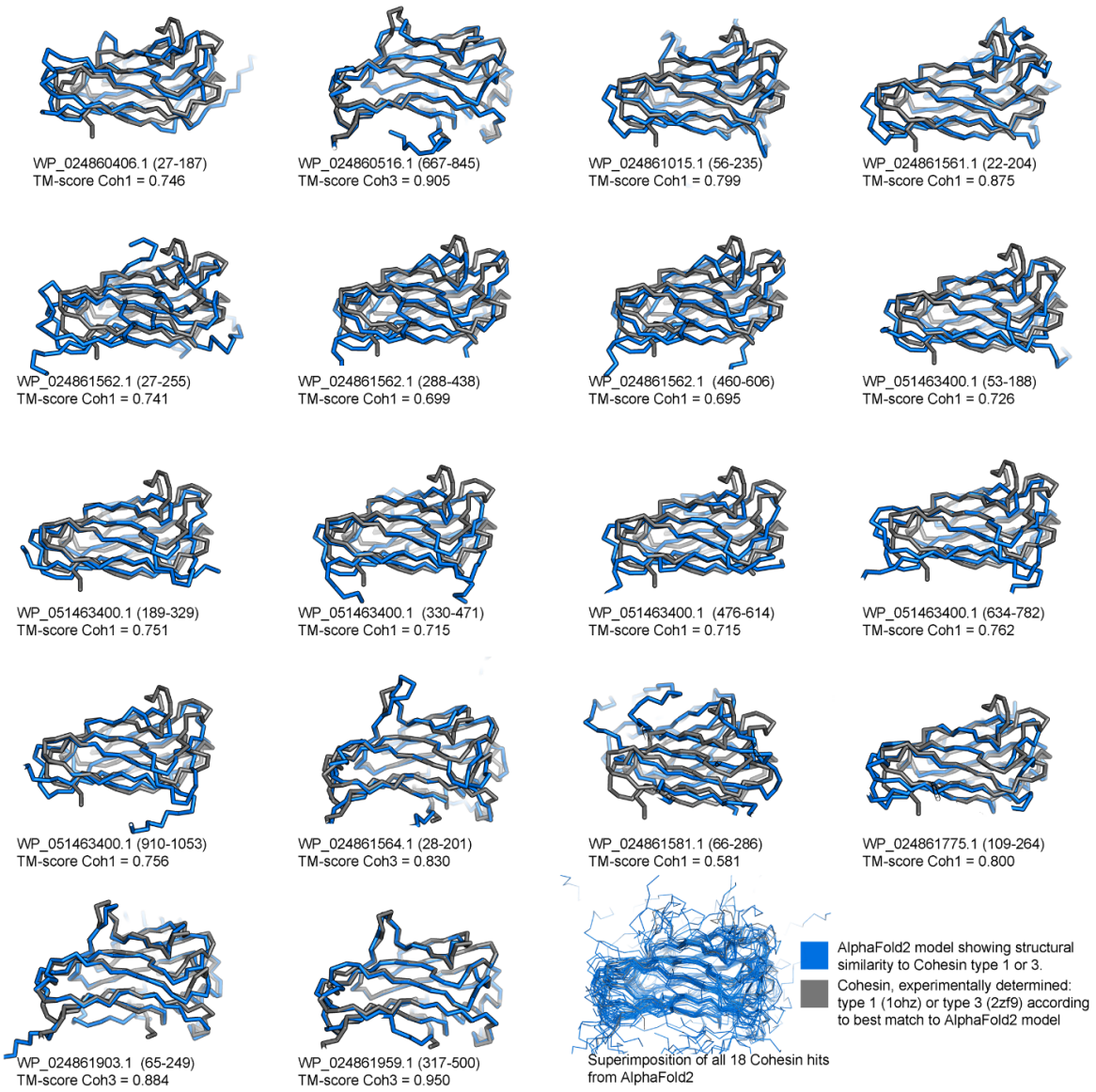


Figure S4.2 Superposition of AlphaFold2 identified cohesin domains within *Ruminococcus flavefaciens* (AE3010)

Eighteen cohesin domains were identified by AlphaFold2 in Ruminococcus flavefaciens (AE3010). The blue chain represents the predicted AlphaFold2 model, and the gray chain shows either an experimentally determined Coh1 (PDB: 1OHZ) or Coh3 (PDB: 2FZ9). Fourteen Coh1 domains and four Coh3 domains were captured with a TM-score above 0.5 to either

representative structures of Coh1 (PDB: 1OHZ) or Coh3 (PDB: 2FZ9), respectively. The last panel is a superposition of all 18 hits. Six domains had been identified using any of the following cohesin-type identifiers (coh1 (cd08548), coh2 (cd08547), coh3 (cd08759), coh (PF00963)). A total of 18 domains were identified by AlphaFold2 to be cohesin modules.

4.9 References

1. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 6506–6511 (2018).
2. Chundawat, S. P. S., Beckham, G. T., Himmel, M. E. & Dale, B. E. Deconstruction of lignocellulosic biomass to fuels and chemicals. *Annu. Rev. Chem. Biomol. Eng.* **2**, 121–145 (2011).
3. Liu, Y. *et al.* Challenges and Future Perspectives of Promising Biotechnologies for Lignocellulosic Biorefinery. *Mol. Basel Switz.* **26**, 5411 (2021).
4. Shukla, A. *et al.* Strategies of pretreatment of feedstocks for optimized bioethanol production: distinct and integrated approaches. *Biotechnol. Biofuels Bioprod.* **16**, 44 (2023).
5. Wang, Y., Zhang, Y., Cui, Q., Feng, Y. & Xuan, J. Composition of Lignocellulose Hydrolysate in Different Biorefinery Strategies: Nutrients and Inhibitors. *Mol. Basel Switz.* **29**, 2275 (2024).
6. Dassa, B. *et al.* Pan-Cellulosomics of Mesophilic Clostridia: Variations on a Theme. *Microorganisms* **5**, 74 (2017).
7. Artzi, L., Bayer, E. A. & Moraïs, S. Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. *Nat. Rev. Microbiol.* **15**, 83–95 (2017).
8. Gilbert, H. J. Cellulosomes: microbial nanomachines that display plasticity in quaternary structure. *Mol. Microbiol.* **63**, 1568–1576 (2007).
9. Bayer, E. A., Belaich, J.-P., Shoham, Y. & Lamed, R. The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annu. Rev. Microbiol.* **58**, 521–554 (2004).

10. Smith, S. P. & Bayer, E. A. Insights into cellulosome assembly and dynamics: from dissection to reconstruction of the supramolecular enzyme complex. *Curr. Opin. Struct. Biol.* **23**, 686–694 (2013).

11. Doi, R. H., Kosugi, A., Murashima, K., Tamaru, Y. & Han, S. O. Cellulosomes from mesophilic bacteria. *J. Bacteriol.* **185**, 5907–5914 (2003).

12. Lamed, R., Setter, E. & Bayer, E. A. Characterization of a cellulose-binding, cellulase-containing complex in Clostridium thermocellum. *J. Bacteriol.* **156**, 828–836 (1983).

13. Hong, W. *et al.* The contribution of cellulosomal scaffoldins to cellulose hydrolysis by Clostridium thermocellum analyzed by using thermotargetrons. *Biotechnol. Biofuels* **7**, 80 (2014).

14. Bhardwaj, N., Kumar, B., Agrawal, K. & Verma, P. Current perspective on production and applications of microbial cellulases: a review. *Bioresour. Bioprocess.* **8**, 95 (2021).

15. Sharma, A., Tewari, R., Rana, S. S., Soni, R. & Soni, S. K. Cellulases: Classification, Methods of Determination and Industrial Applications. *Appl. Biochem. Biotechnol.* **179**, 1346–1380 (2016).

16. Lu, Y., Zhang, Y.-H. P. & Lynd, L. R. Enzyme-microbe synergy during cellulose hydrolysis by Clostridium thermocellum. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 16165–16169 (2006).

17. Barba-Cedillo, V. & Montanier, C. Y. Effect of multimodularity and spatial organization of glycoside hydrolases on catalysis. *Essays Biochem.* **67**, 629–638 (2023).

18. Smith, S. P., Bayer, E. A. & Czjzek, M. Continually emerging mechanistic complexity of the multi-enzyme cellulosome complex. *Curr. Opin. Struct. Biol.* **44**, 151–160 (2017).

19. Bule, P., Pires, V. M., Fontes, C. M. & Alves, V. D. Cellulosome assembly: paradigms are meant to be broken! *Curr. Opin. Struct. Biol.* **49**, 154–161 (2018).

20. Bae, J., Morisaka, H., Kuroda, K. & Ueda, M. Cellulosome complexes: natural biocatalysts as arming microcompartments of enzymes. *J. Mol. Microbiol. Biotechnol.* **23**, 370–378 (2013).

21. Leibovitz, E. & Béguin, P. A new type of cohesin domain that specifically binds the dockerin domain of the Clostridium thermocellum cellulosome-integrating protein CipA. *J. Bacteriol.* **178**, 3077–3084 (1996).

22. Pagès, S. *et al.* Species-specificity of the cohesin-dockerin interaction between Clostridium thermocellum and Clostridium cellulolyticum: prediction of specificity determinants of the dockerin domain. *Proteins* **29**, 517–527 (1997).

23. Hamberg, Y. *et al.* Elaborate cellulosome architecture of Acetivibrio cellulolyticus revealed by selective screening of cohesin-dockerin interactions. *PeerJ* **2**, e636 (2014).

24. Phitsuwan, P., Moraïs, S., Dassa, B., Henrissat, B. & Bayer, E. A. The Cellulosome Paradigm in An Extreme Alkaline Environment. *Microorganisms* **7**, 347 (2019).

25. Artzi, L. *et al.* Cellulosomics of the cellulolytic thermophile Clostridium clariflavum. *Biotechnol. Biofuels* **7**, 100 (2014).

26. Carvalho, A. L. *et al.* Evidence for a dual binding mode of dockerin modules to cohesins. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 3089–3094 (2007).

27. Cameron, K. *et al.* Cell-surface Attachment of Bacterial Multienzyme Complexes Involves Highly Dynamic Protein-Protein Anchors. *J. Biol. Chem.* **290**, 13578–13590 (2015).

28. Brás, J. L. A. *et al.* Diverse specificity of cellulosome attachment to the bacterial cell surface. *Sci. Rep.* **6**, 38292 (2016).

29. Zhivin, O. *et al.* Unique organization and unprecedented diversity of the Bacteroides (Pseudobacteroides) cellulosolvens cellulosome system. *Biotechnol. Biofuels* **10**, 211 (2017).

30. Dassa, B. *et al.* Genome-wide analysis of acetivibrio cellulolyticus provides a blueprint of an elaborate cellulosome system. *BMC Genomics* **13**, 210 (2012).

31. Tao, X., Liu, J., Kempher, M. L., Xu, T. & Zhou, J. In vivo Functional Characterization of Hydrophilic X2 Modules in the Cellulosomal Scaffolding Protein. *Front. Microbiol.* **13**, 861549 (2022).

32. Raman, B. *et al.* Impact of pretreated Switchgrass and biomass carbohydrates on Clostridium thermocellum ATCC 27405 cellulosome composition: a quantitative proteomic analysis. *PloS One* **4**, e5271 (2009).

33. Nölling, J. *et al.* Genome sequence and comparative analysis of the solvent-producing bacterium Clostridium acetobutylicum. *J. Bacteriol.* **183**, 4823–4838 (2001).

34. Li, W. *et al.* RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res.* **49**, D1020–D1028 (2021).

35. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform. Int. Conf. Genome Inform.* **23**, 205–211 (2009).

36. Marchler-Bauer, A. *et al.* CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* **30**, 281–283 (2002).

37. Sonnhammer, E. L., Eddy, S. R. & Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405–420 (1997).

38. Hulo, N. *et al.* The PROSITE database. *Nucleic Acids Res.* **34**, D227-230 (2006).

39. Mende, D. R. *et al.* proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Res.* **45**, D529–D534 (2017).

40. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116-120 (2005).

41. Käll, L., Krogh, A. & Sonnhammer, E. L. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036 (2004).

42. Pandurangan, A. P., Stahlhacke, J., Oates, M. E., Smithers, B. & Gough, J. The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. *Nucleic Acids Res.* **47**, D490–D494 (2019).

43. Sigrist, C. J. A. *et al.* New and continuing developments at PROSITE. *Nucleic Acids Res.* **41**, D344-347 (2013).

44. Wang, J. *et al.* The conserved domain database in 2023. *Nucleic Acids Res.* **51**, D384–D388 (2023).

45. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

46. Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025 (2022).

47. Hallgren, J. *et al.* DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. Preprint at https://doi.org/10.1101/2022.04.08.487609 (2022).

48. Yu, N. Y. *et al.* PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinforma. Oxf. Engl.* **26**, 1608–1615 (2010).

49. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinforma. Oxf. Engl.* **30**, 2068–2069 (2014).

50. Sabathé, F., Bélaïch, A. & Soucaille, P. Characterization of the cellulolytic complex (cellulosome) of Clostridium acetobutylicum. *FEMS Microbiol. Lett.* **217**, 15–22 (2002).

51. Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **37**, D233-238 (2009).

52. Ben David, Y. *et al.* Ruminococcal cellulosome systems from rumen to human. *Environ. Microbiol.* **17**, 3407–3426 (2015).

53. Dassa, B. *et al.* Rumen cellulosomics: divergent fiber-degrading strategies revealed by comparative genome-wide analysis of six ruminococcal strains. *PloS One* **9**, e99221 (2014).

54. Ding, S. Y. *et al.* Cellulosomal scaffoldin-like proteins from Ruminococcus flavefaciens. *J. Bacteriol.* **183**, 1945–1953 (2001).

55. Rincon, M. T. *et al.* Novel organization and divergent dockerin specificities in the cellulosome system of Ruminococcus flavefaciens. *J. Bacteriol.* **185**, 703–713 (2003).

56. Rincón, M. T. *et al.* ScaC, an adaptor protein carrying a novel cohesin that expands the dockerin-binding repertoire of the Ruminococcus flavefaciens 17 cellulosome. *J. Bacteriol.* **186**, 2576–2585 (2004).

57. Dawson, N. L. *et al.* CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* **45**, D289–D295 (2017).

58. Dong, R., Pan, S., Peng, Z., Zhang, Y. & Yang, J. mTM-align: a server for fast protein structure database search and multiple protein structure alignment. *Nucleic Acids Res.* **46**, W380–W386 (2018).

59. Ze, X. *et al.* Unique Organization of Extracellular Amylases into Amylosomes in the Resistant Starch-Utilizing Human Colonic Firmicutes Bacterium Ruminococcus bromii. *mBio* **6**, e01058-01015 (2015).

60. Wang, Y., Li, L., Xia, Y. & Zhang, T. Reliable and Scalable Identification and Prioritization of Putative Cellulolytic Anaerobes With Large Genome Data. *Front. Bioinforma.* **2**, 813771 (2022).

61. Artzi, L., Morag, E., Barak, Y., Lamed, R. & Bayer, E. A. Clostridium clariflavum: Key Cellulosome Players Are Revealed by Proteomic Analysis. *mBio* **6**, e00411-00415 (2015).

62. Rettenmaier, R., Gerbaulet, M., Liebl, W. & Zverlov, V. V. Hungateiclostridium mesophilum sp. nov., a mesophilic, cellulolytic and spore-forming bacterium isolated from a biogas fermenter fed with maize silage. *Int. J. Syst. Evol. Microbiol.* **69**, 3567–3573 (2019).

63. Aikawa, S. *et al.* Characterization and high-quality draft genome sequence of Herbivorax saccincola A7, an anaerobic, alkaliphilic, thermophilic, cellulolytic, and xylanolytic bacterium. *Syst. Appl. Microbiol.* **41**, 261–269 (2018).

64. Kato, S. *et al.* Clostridium straminisolvens sp. nov., a moderately thermophilic, aerotolerant and cellulolytic bacterium isolated from a cellulose-degrading bacterial community. *Int. J. Syst. Evol. Microbiol.* **54**, 2043–2047 (2004).

65. Xu, Q. *et al.* A novel Acetivibrio cellulolyticus anchoring scaffoldin that bears divergent cohesins. *J. Bacteriol.* **186**, 5782–5789 (2004).

66. Xu, Q. *et al.* Architecture of the Bacteroides cellulosolvens cellulosome: description of a cell surface-anchoring scaffoldin and a family 48 cellulase. *J. Bacteriol.* **186**, 968–977 (2004).

67. Chow, N. & Wu, J. H. D. The Cellulosome. in *Biotechnology of Microbial Enzymes* 243–266 (Elsevier, 2017). doi:10.1016/B978-0-12-803725-6.00010-8.

68. Blouzard, J.-C. *et al.* Modulation of cellulosome composition in Clostridium cellulolyticum: adaptation to the polysaccharide environment revealed by proteomic and carbohydrate-active enzyme analyses. *Proteomics* **10**, 541–554 (2010).

69. Celik, H. *et al.* A two-component system (XydS/R) controls the expression of genes encoding CBM6-containing proteins in response to straw in Clostridium cellulolyticum. *PloS One* **8**, e56063 (2013).

70. Kampik, C. *et al.* A Novel Two-Component System, XygS/XygR, Positively Regulates Xyloglucan Degradation, Import, and Catabolism in Ruminiclostridium cellulolyticum. *Appl. Environ. Microbiol.* **86**, e01357-20 (2020).

71. Bhaskar, Y., Su, X., Xu, C. & Xu, J. Predicting Selective RNA Processing and Stabilization Operons in Clostridium spp. *Front. Microbiol.* **12**, 673349 (2021).

72. Xu, C. *et al.* Cellulosome stoichiometry in Clostridium cellulolyticum is regulated by selective RNA processing and stabilization. *Nat. Commun.* **6**, 6900 (2015).

73. Nataf, Y. *et al.* Clostridium thermocellum cellulosomal genes are regulated by extracytoplasmic polysaccharides via alternative sigma factors. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 18646–18651 (2010).

74. Muñoz-Gutiérrez, I. *et al.* Decoding Biomass-Sensing Regulons of Clostridium thermocellum Alternative Sigma-I Factors in a Heterologous Bacillus subtilis Host System. *PloS One* **11**, e0146316 (2016).

75. Kahel-Raifer, H. *et al.* The unique set of putative membrane-associated anti-sigma factors in Clostridium thermocellum suggests a novel extracellular carbohydrate-sensing mechanism involved in gene regulation. *FEMS Microbiol. Lett.* **308**, 84–93 (2010).

76. Rincon, M. T. *et al.* A novel cell surface-anchored cellulose-binding protein encoded by the sca gene cluster of Ruminococcus flavefaciens. *J. Bacteriol.* **189**, 4774–4783 (2007).

77. Jindou, S. *et al.* Cellulosome gene cluster analysis for gauging the diversity of the ruminal cellulolytic bacterium Ruminococcus flavefaciens. *FEMS Microbiol. Lett.* **285**, 188–194 (2008).

78. Salama-Alber, O. *et al.* Crystallization and preliminary X-ray characterization of a type III cohesin-dockerin complex from the cellulosome system of Ruminococcus

flavefaciens. *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.* **68**, 1116–1119 (2012).

79. Tindall, B. J. Replacement of the illegitimate genus name Hungateiclostridium Zhang et al. 2018 in Hungateiclostridium mesophilum Rettenmaier et al. 2019 by Acetivibrio Patel et al. 1980, creating Acetivibrio mesophilus (Rettenmaier et al. 2019). *Int. J. Syst. Evol. Microbiol.* **69**, 3967–3968 (2019).

80. Liu, Y. *et al.* Characterization of Two α-l-Arabinofuranosidases from Acetivibrio mesophilus and Their Synergistic Effect in Degradation of Arabinose-Containing Substrates. *Microorganisms* **9**, 1467 (2021).

81. Monserrate, E., Leschine, S. B. & Canale-Parola, E. Clostridium hungatei sp. nov., a mesophilic, N2-fixing cellulolytic bacterium isolated from soil. *Int. J. Syst. Evol. Microbiol.* **51**, 123–132 (2001).

82. Ren, Z., You, W., Wu, S., Poetsch, A. & Xu, C. Secretomic analyses of Ruminiclostridium papyrosolvens reveal its enzymatic basis for lignocellulose degradation. *Biotechnol. Biofuels* **12**, 183 (2019).

83. Kakiuchi, M. *et al.* Cloning and DNA sequencing of the genes encoding Clostridium josui scaffolding protein CipA and cellulase CelD and identification of their gene products as major components of the cellulosome. *J. Bacteriol.* **180**, 4303–4308 (1998).

84. Rettenmaier, R., Kowollik, M.-L., Klingl, A., Liebl, W. & Zverlov, V. Ruminiclostridium herbifermentans sp. nov., a mesophilic and moderately thermophilic cellulolytic and xylanolytic bacterium isolated from a lab-scale biogas fermenter fed with maize silage. *Int. J. Syst. Evol. Microbiol.* **71**, (2019).

85. Desvaux, M. Clostridium cellulolyticum: model organism of mesophilic cellulolytic clostridia. *FEMS Microbiol. Rev.* **29**, 741–764 (2005).

86. Nishiyama, T., Ueki, A., Kaku, N. & Ueki, K. Clostridium sufflavum sp. nov., isolated from a methanogenic reactor treating cattle waste. *Int. J. Syst. Evol. Microbiol.* **59**, 981–986 (2009).

87. Collins, M. D. *et al.* The phylogeny of the genus Clostridium: proposal of five new genera and eleven new species combinations. *Int. J. Syst. Bacteriol.* **44**, 812–826 (1994).

88. López-Contreras, A. M. *et al.* Production by Clostridium acetobutylicum ATCC 824 of CelG, a cellulosomal glycoside hydrolase belonging to family 9. *Appl. Environ. Microbiol.* **69**, 869–877 (2003).

89. López-Contreras, A. M. *et al.* Substrate-induced production and secretion of cellulases by Clostridium acetobutylicum. *Appl. Environ. Microbiol.* **70**, 5238–5243 (2004).

90. Tomazetto, G. *et al.* Complete genome analysis of Clostridium bornimense strain M2/40(T): A new acidogenic Clostridium species isolated from a mesophilic two-phase laboratory-scale biogas reactor. *J. Biotechnol.* **232**, 38–49 (2016).

91. Gilroy, R. *et al.* Extensive microbial diversity within the chicken gut microbiome revealed by metagenomics and culture. *PeerJ* **9**, e10941 (2021).

92. Levi Hevroni, B., Moraïs, S., Ben-David, Y., Morag, E. & Bayer, E. A. Minimalistic Cellulosome of the Butanologenic Bacterium Clostridium saccharoperbutylacetonicum. *mBio* **11**, e00443-20 (2020).

93. You, M. *et al.* Insights into lignocellulose degradation: comparative genomics of anaerobic and cellulolytic Ruminiclostridium-type species. *Front. Microbiol.* **14**, 1288286 (2023).

94. Giallo, J., Gaudin, C. & Belaich, J. P. Metabolism and Solubilization of Cellulose by Clostridium cellulolyticum H10. *Appl. Environ. Microbiol.* **49**, 1216–1221 (1985).

95. Usai, G. *et al.* Clostridium cellulovorans metabolism of cellulose as studied by comparative proteomic approach. *J. Proteomics* **216**, 103667 (2020).

96. Li, L.-L., Taghavi, S., Izquierdo, J. A. & van der Lelie, D. Complete genome sequence of Clostridium sp. strain BNL1100, a cellulolytic mesophile isolated from corn stover. *J. Bacteriol.* **194**, 6982–6983 (2012).

97. Koeck, D. E., Hahnke, S. & Zverlov, V. V. Herbinix luporum sp. nov., a thermophilic cellulose-degrading bacterium isolated from a thermophilic biogas reactor. *Int. J. Syst. Evol. Microbiol.* **66**, 4132–4137 (2016).

98. Wylensek, D. *et al.* A collection of bacterial isolates from the pig intestine reveals functional and taxonomic diversity. *Nat. Commun.* **11**, 6389 (2020).

99. Kosugi, A., Amano, Y., Murashima, K. & Doi, R. H. Hydrophilic domains of scaffolding protein CbpA promote glycosyl hydrolase activity and localization of cellulosomes to the cell surface of Clostridium cellulovorans. *J. Bacteriol.* **186**, 6351–6359 (2004).

100. Tarraran, L., Gandini, C., Luganini, A. & Mazzoli, R. Cell-surface binding domains from Clostridium cellulovorans can be used for surface display of cellulosomal scaffoldins in Lactococcus lactis. *Biotechnol. J.* **16**, e2100064 (2021).

101. Kataeva, I. A. *et al.* The fibronectin type 3-like repeat from the Clostridium thermocellum cellobiohydrolase CbhA promotes hydrolysis of cellulose by modifying its surface. *Appl. Environ. Microbiol.* **68**, 4292–4300 (2002).

102. Alahuhta, M. *et al.* Structure of a fibronectin type III-like module from Clostridium thermocellum. *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.* **66**, 878–880 (2010).

103. Zhang, J. *et al.* Iocasia fonsfrigidae NS-1 gen. nov., sp. nov., a Novel Deep-Sea Bacterium Possessing Diverse Carbohydrate Metabolic Pathways. *Front. Microbiol.* **12**, 725159 (2021).

104. Li, J. *et al.* Paenibacillus guangzhouensis sp. nov., an Fe(III)- and humus-reducing bacterium from a forest soil. *Int. J. Syst. Evol. Microbiol.* **64**, 3891–3896 (2014).

105. Pradel, N., Fardeau, M.-L., Tindall, B. J. & Spring, S. Anaerohalosphaera lusitana gen. nov., sp. nov., and Limihaloglobus sulfuriphilus gen. nov., sp. nov., isolated from solar saltern sediments, and proposal of Anaerohalosphaeraceae fam. nov. within the order Sedimentisphaerales. *Int. J. Syst. Evol. Microbiol.* **70**, 1321–1330 (2020).

106. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).

107. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

108. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinforma. Oxf. Engl.* **30**, 1236–1240 (2014).

109. Gough, J., Karplus, K., Hughey, R. & Chothia, C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903–919 (2001).

110. Letunic, I., Khedkar, S. & Bork, P. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res.* **49**, D458–D460 (2021).

111. Attwood, T. K. *et al.* The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012. *Database J. Biol. Databases Curation* **2012**, bas019 (2012).

112. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

113. Zhu, K., Su, H., Peng, Z. & Yang, J. A unified approach to protein domain parsing with inter-residue distance matrix. *Bioinforma. Oxf. Engl.* **39**, btad070 (2023).

Chapter 5

AlphaFold Structural Proteomics Reveals Complex Cellulosome Machinery for Carbohydrate Degradation in Human Gut Bacterium *Ruminococcus callidus* and Related Species

The work described in this chapter is a version of the in-progress manuscript to be submitted for publication

Minor C, T**akayesu A**, Ha SM , Salwinski L, Pellegrini M, Gunsalus RP, Arbing MA, Sawaya MR, and Clubb RT (2025)

I contributed to this work by determining the crystal structure of one of the newly identified cohesin types, making figures, writing, and analysis of the data.

5.1 Overview

One health benefit attributed to the human gut microbiome arises from its capacity to break down indigestible plant fibers. Among these, cellulose is the most abundant carbohydrate polymer in nature, remains difficult to break down by natural human enzymes and primarily relies on specialized gut bacteria. These gut bacteria have dedicated supramolecular surface-anchored multi-protein complexes that facilitate the efficient breakdown of complex polysaccharides from our diets. To date, only three species from the *Ruminococcus* genus have been identified, *Ruminococcus champanellensis*, *Ruminococcus hominiciens*, and *Ruminococcus primaciens*, within the human gut microbiome to contain cellulosome machinery. These cellulosome producing species were identified using sequence-based homology searches targeting cohesin- and dockerin- containing proteins, which are key structural components of the cellulosome. With the goal of expanding the depth of cellulosome producing species, here we survey a dataset of published non-redundant microbial genomes from the human gut and take a structure-based homology approach to identify a divergent set of cohesin containing proteins. To establish a robust novel cohesin profile, our search for divergent cohesin domains was expanded to over 300k genomes from the NCBI Reference Sequence Database (RefSeq). AlphaFold was used to predict the structures of 26 species' proteomes to reveal previously 170 previously undetectable cohesin domains. We establish that *Ruminococcus callidus*, a prominent gut microbe in healthy individuals, has a previously unidentified cellulosome and potential cellulolytic activity.

5.2 Introduction

The human gut microbiota plays a crucial role in the digestion of complex carbohydrates, including plant-derived polysaccharides such as hemicellulose, pectin, resistant starch, and, to a limited extent, cellulose [1]. Cellulose is the most abundant carbohydrate polymer in nature, but its crystalline, insoluble structure makes it resistant to degradation by human enzymes [2]. Instead, gut bacteria ferment cellulose into short-chain fatty acids that support host energy metabolism, immune function, and overall gut health [3,4]. While the exact diversity of the human gut microbiome relies heavily on an individual's diet and geographical location, in general the human microbiome is dominated by four main bacterial phyla; *Bacteroidota* (formerly *Bacteroidetes*) (17%), *Bacillota* (formerly *Firmicutes*) (70.6%), *Pseudomonadota* (formerly *Proteobacteria*) (5%), and *Actinomycetota* (formerly *Actinobacteria*) (3.6%) [5]. In addition, roughly 1.2% of the rumen microbiome is composed of methanogenic archaea [6]. Interestingly, many environmental and herbivore-associated bacterial species can efficiently degrade cellulose using large, surface-anchored, multi-enzyme complexes known as cellulosomes that concentrate cellulolytic activity on the microbial surface [7,8]. However, despite the presence of over 5,000 microbial species in the human-gut, only three species have the genetic capacity to produce cellulosomes [9]. We hypothesized that reliance on sequence-based homology searches has hindered cellulosome detection, as the primary sequences of many cellulosome components may be too divergent to identify structural homologs. Here, we demonstrate the power of proteome-scale, structure-based analysis using AlphaFold to uncover novel, previously undetectable cellulosome architectures. Using this approach, we reveal for the first time that the human commensal bacterium *Ruminococcus callidus,* which is part of the core microbiome found in healthy individuals, encodes a previously undetectable cellulosome, consisting of divergent cohesin modules.

Cellulosomes are present in a range of anaerobic environmental bacteria, and were first discovered in *Acetivibrio thermocellus* (formerly known as *Clostridium thermocellum* and *Hungateiclostridium thermocellum*) [10,11]. These extracellular multi-enzyme assemblies consist of scaffoldin proteins that organize numerous glycoside hydrolases (GHs) through high-affinity, non-covalent interactions between cohesin domains on the scaffoldins and dockerin domains fused to the GHs (referred to as DocGH enzymes). Scaffoldins often include additional functional domains, such as carbohydrate-binding modules (CBMs), cell wall-anchoring motifs like S-layer homology (SLH) domains, and dockerins that enable interaction with secondary scaffoldins anchored at the cell surface. In *A. thermocellus*, for example, the primary scaffoldin ScaA includes several cohesin domains for GH recruitment, a CBM3 domain for cellulose binding, and a C-terminal dockerin that attaches to other cell wall–associated scaffoldins including ScaB, ScaC, ScaD, and ScaF [12]. Cellulose degradation by the cellulosome relies on the concerted activity of three enzyme types: endoglucanases that cleave internal β-1,4 linkages (e.g., GH7, GH12), exoglucanases that act on chain ends (e.g., GH5, GH9), and β-glucosidases that convert oligosaccharides into glucose [13,14]. Though the substrate specificities vary, GH families such as GH5, GH10, GH11, GH43, and GH48 are commonly represented [7]. Additionally, other types of carbohydrate-active enzymes (CAZymes), including polysaccharide lyases and carbohydrate esterases, are also incorporated via dockerin fusions. The spatial organization and proximity of enzymes within the cellulosome enhance lignocellulose degradation by facilitating enzyme synergy, improving substrate targeting, and promoting close interactions between the microbial cell surface and the polysaccharide substrate.

Our recent comparative genome analysis identified 33 cellulosome-producing bacterial species, primarily from the *Acetivibrio*, *Clostridium*, *Ruminiclostridium*, and *Ruminococcus* genera [15]. Further classification based on the number of scaffoldin and DocGHs encoding genes distinguished them into ‘complex’ (*Acetivibrio* and *Ruminococcus*) and ‘simple’

(*Clostridium* and *Ruminiclostridium*) cellulosome producing bacteria. Complex cellulosomes typically have distinct mechanisms of attachment to the cell wall and unique primary scaffoldin proteins, while simple cellulosomes characteristically have fewer scaffoldin proteins and DocGH enzymes, suggesting they exhibit lower cellulolytic activity. Based on their primary sequences, three types of cohesins and dockerins have been identified, referred to here as Coh1 to Coh3 and Doc1 to Doc3, respectively. In general, cohesin-dockerin interactions are both species- and type-specific (e.g., only Coh1 and Doc1 proteins produced by the same species interact with each other) [16,17]. In non-*Ruminococcus* species, GH enzymes are typically fused to Doc1 domains, enabling them to bind Coh1 domains within the scaffoldin. Meanwhile, Coh2-Doc2 interactions mediate scaffoldin-scaffoldin assembly within the cellulosome. In contrast, Coh3 and Doc3 domains predominate in Ruminococcus species, which in some instances employ sortase transpeptidase enzymes to covalently attach their cellulosomes to the cell wall.

In the human gut, members of the genus *Ruminococcus* contribute significantly to the breakdown of dietary polysaccharides and their altered abundance is associated with metabolic disorders, inflammatory bowel disease, and other gut-related conditions. In herbivores, *Ruminococcus* bacteria producing cellulosomes are present in the rumen where they efficiently breakdown plant cell wall polysaccharides (*R. flavefaciens* and *R. albus*) [18]. However, of the ten Ruminococcus species that inhabit the human gut (*R. bromii, R. gnavus, R. torques, R. champanellensis, R. lactaris, R. callidus, R. hansenii, R. obeum* (now known as *Blautia obeum*)), only *R. champanellensis* has an identifiable cellulosome based on its genome sequence [19]. Identifiable cellulosomes within metagenomic assembled genomes of human gut bacteria are also rare, with the notable exception of the newly discovered *R. primaciens* and *R. hominiciens* [9]. Indeed, an inspection of sequenced *Ruminococcus* genomes reveal that many species encode a large number of DocGH enzymes that are presumably capable of cellulosome binding, but only a limited number of identifiable scaffoldin genes (**Figure 5.1**). Thus, these

Ruminococcus microbes either possess divergent carbohydrate degrading systems and/or cellulosomes with unconventional architectures that can not be detected using traditional sequence-homology based methods.

It has long been known that proteins with highly divergent primary sequences (<20% sequence identity) can nevertheless adopt related tertiary structures that have similar biological functions. Thus, primary sequence alone is not always sufficient to infer protein function. To overcome this limitation, we performed a high throughput application of AlphaFold to identify cohesin domains within 24 *Ruminococcus* genomes. This analysis revealed a significant number of novel cohesin domain encoding genes that were otherwise undetectable based on their primary sequence using existing Hidden Markov Model profiles within the InterPro database. Experimental atomic structures of several of these domains confirm predictions, and provide evidence of newly discovered cohesin families in a range of *Ruminococcus* species. Notably, we identify for the first time that *Ruminococcus callidus*, a human gut symbiont, encodes a structurally complex cellulosome, underscoring its likely importance in human dietary fiber degradation and microbial community structure.

5.3 Results

5.3.1 AlphaFold structural proteomics identifies novel cohesin containing scaffoldins.

We used AlphaFold structure prediction methods to identify cellulosome components in Ruminococcus species and other bacteria that are known to inhabit the human gut (**Figure 5.2**). The following genomes were analyzed: (1) 305,693 prokaryotic genomes from the NCBI Reference Sequence Database (RefSeq) (July 10, 2023 release) and (2) metagenome-assembled genomes (MAGs) from recently identified gut bacteria (*R. primaciens*, *R. hominiciens*, and *R. ruminiciens*) [9, 20]. Initially, a HMMR search was performed on each of the RefSeq annotated genomes to identify DocGH and cohesin domain encoding genes as previously described. A total of 23 Ruminococcus genomes were found to contain genes suggesting that the microbe might produce a cellulosome; at least five genes encoding DocGH proteins and one gene encoding a multi-cohesin containing protein that might function as a scaffoldin. Of the 23 that were identified, eight genomes are complete and 15 are metagenome-assembled genomes (MAGs). Because the proteome scale application of AlphaFold is computationally intensive, only the full proteomes of the eight complete Ruminococcus genomes from RefSeq and the three established human gut MAGs were structurally predicted. For the remaining 15 MAGs, only those proteins annotated to have a signal peptide by SignalP (SignalP v. 6.0) were structurally predicted, this greatly reduced the computational time given that on average only 16% of the genome had identified signal peptides.

The atomic structures of all proteins were predicted using a locally implemented version of the AlphaFold2 program (29,657 atomic structures were predicted corresponding to ~2,800 structures per complete proteome and 470 structures per MAG). Each protein structure was then partitioned into its component domains using the program Unidoc and their similarity to experimentally determined cohesin structures determined using the TM-align algorithm. **Figure 5.3** shows representative plots of the template matching (TM) score between predicted domain

structures and the structure of an experimentally determined Coh1 domain (PDB:1OHZ) domain for *R. flavefaciens*, and *R. champanellensis*, and *R. callidus*. Only a few domains in each microbe exhibit TM score ≥ 0.7 when compared to Coh1 indicating that they are structurally related, consistent with the small magnitude Root Mean Square Deviation (RMSD) differences between the domains and Coh1.

We identified predicted domains that were structurally similar to Coh1 (PDB:1OHZ), Coh2 (PDB:1TYJ) and Coh3 (PDB:2ZF9) using a TM score threshold of ≥0.7. As a negative control, their similarity to the structure of a carbohydrate binding module family 3 (CBM3) protein (PDB:6UFW) was also calculated, as CMB3 and cohesins have distinct functions but are known to adopt related anti-parallel β-sandwich folds [21]. This AlphaFold-based screen identified a significant number of scaffoldins in the Ruminococcus proteomes that were otherwise undetectable using primary sequence homology-based methods (**Fig. 4A**). In particular, a total of 404 cohesin domains were detected across the surveyed Ruminococcus genomes based on their AlphaFold predicted structures, whereas only 155 cohesins could be identified based on a primary sequence analysis using InterProScan. AlphaFold also identified more cohesins as compared to sequence analyses of the proteomes using BLAST and representative sequence profiles for the different types of cohesins. Importantly, with only 2 exceptions, all cohesin domains that can be detected based on their primary sequence using BLAST or InterProScan were identified using AlphaFold, while 170 domains were only detectable after their structures were predicted by AlphaFold.

5.3.2 Discovery of cellulosomes in *R. callidus* and other *Ruminococcus* spp.

The AlphaFold structural proteomics reveals the presence of an array of new scaffoldins in 21 of the 26 Ruminococcus species that were investigated. The largest number of novel scaffoldins were detected in the proteome of the human gut bacterium *R. callidus*, which based on primary sequence analysis encodes 26 DocGH enzymes across 76 dockerin-fused proteins

but only two single-cohesin scaffoldins. Based on the AlphaFold predictions this microbe contains an additional 6 scaffoldins which collectively harbor 14 putative cohesin domains. These newly identified scaffoldins demonstrate hallmarks of a *Ruminococcus* complex cellulosome as it contains both sortase-attached cell-wall anchoring and adapter scaffoldins. AlphaFold-identified scaffoldins in *R. callidus* also include two large multi-cohesin proteins with C-terminally fused dockerin domains that resemble ScaA- and ScaB-like scaffoldins typically seen in other established cellulosome producing *Ruminococcus* [18]. Novel AlphaFold discovered cohesins are also present in previously characterized *Ruminococcus* species that inhabit either the rumen (*R. flavefaciens* and *R. albus)* or human gut (*R. champanellensis, and R. bromii)* revealing the produce cellulosomes of greater complexity than previously appreciated. Notably, an AlphaFold analysis of environmental bacteria from other genera known to produce cellulosomes (*A. thermocellus* and *Ruminiclostridium cellulolyticum*) did not uncover new cohesin modules, suggesting that unlike Ruminococcus Spp, cohesin sequence profiles are generally adequate to detect cohesins in bacteria from these genera (**Figure 5.4B**).

5.3.3 Structure of two novel *Ruminococcus* cohesin groups

Having identified new cohesins within the *R. callidus* genome, we sought to determine their atomic structures using X-ray crystallography in order to both verify our AlphaFold predictions and potentially identify novel features not present in the *in silico* model (**Figure 5.5A**). For this, we selected one representative cohesin domain each from two distinct groups which appeared to cluster with other cohesin domains unique to *R. callidus*-type cellulosomes based on sequence identity. The selected cohesin domains were Rcal_0153 (WP_370809552.1), domain 5 (Rcal_0153_5) and Rcal_2942 (WP_370809561.1), domain 4 (Rcal_2942_4). Crystallization screens yielded high-quality crystals for data collection, allowing for structural determination of both cohesin domains via a molecular replacement strategy

utilizing the generated AlphaFold models for each respective domain. Rcal_0153_5 crystals diffracted to a maximum resolution of 1.8Å and belonged to the orthorhombic space group $C222_1$ with unit cell dimensions of a=42.5Å, b=84.6Å, and c=118.4Å. Crystals for Rcal_02942_4 diffracted to a maximum resolution of 1.2Å and belonged to space group $P2_12_12_1$, with unit cell dimensions of a=34.0Å, b=37.2Å, and c=110.4Å. The final data refinement and structure-quality statistics are listed in Table 5.1.

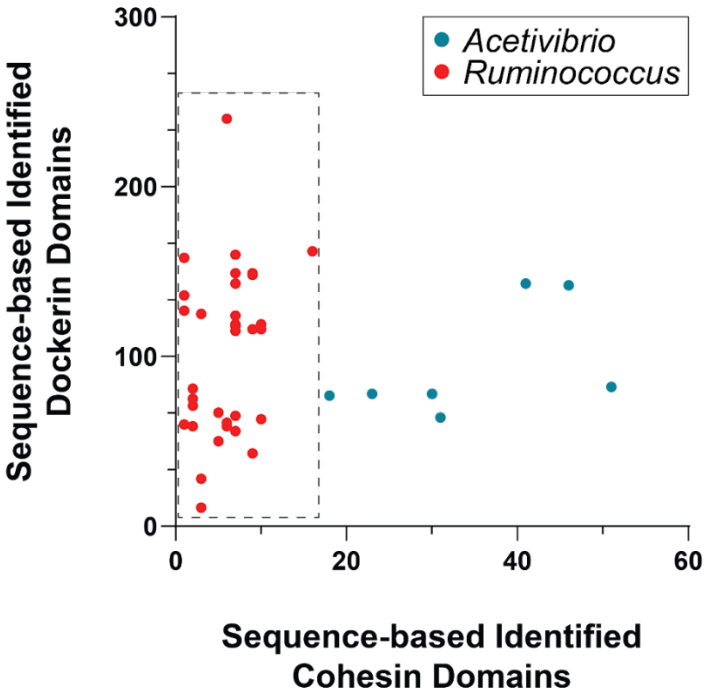
The overall structure of Rcal_2942_4 presents a classical two-sheeted β-sandwich jellyroll topology consisting of 9 antiparallel β-strands as is typical for a cohesin. The solved atomic structure and predicted AlphaFold model of Rcal_2942_4 are closely related, with a backbone atom RMSD of 0.380 Å. Despite not sharing significant sequence homology to any previously reported cohesins, Rcal_2942_4 displays several other features commonly seen in cohesin structures. Similar to the 3_{10}-helix observed in the *R. flavefaciens*' ScaB Coh5, Rcal_2942_4 contains a more well defined α-helix spanning from His521 to Lys525 [22]. Further, Rcal_2942_4 also contains an elongated β-flap interrupting the β8 strand of the cohesin, lying at the Coh:Doc binding interface. This β-flap feature can also be observed in the *R. flavefaciens* ScaC cohesin (PDB 5LXV) as well as other non-*Ruminococcus* type cohesins such as those from ScaC, Coh3 of *Acetivibrio cellulolyticus* (PDB: 4UYP) and ScaB of *Pseudobacteroides cellulosolvens* (PDB: 1TYJ) [23-25]. Unique to the Rcal_2942_4 cohesin is the presence of an additional accessory α-helix not predicted in the original AlphaFold model formed by residues Phe574 to Ser579 between β-strands 5 and 6.

Although Rcal_0153_5 contains the same core 9 β-stranded sandwich topology of classical cohesins, there are several exceptions where the cohesin deviates structurally (**Fig 5.5B**). Most notable of these deviations is the elongated two-β-stranded hairpin that forms between residues Thr1080 and Asp1098. Similar to the β-flap seen interrupting the β8 strand in Rcal_2942_4 and previously reported cohesins, the flap of Rcal_0153_5 forms a break between

strands β7 and β10 and appears to form a structured β8/9-strand hairpin. Additionally, this β-hairpin also lies adjacent to the cohesin:dockerin interface, suggesting it may play a role in forming important contacts for specific dockerin binding. The AlphaFold model for the domain also predicts this β-hairpin, with the full-length structures exhibiting a strong structural similarity (backbone RMSD = 0.748Å). 500 ns molecular dynamics (MD) simulations performed with the crystal structure of Rcal_0153_5 using a CHARMM36 all-atom force field in water reveal that the extended β-hairpin arm does show a high degree of flexibility, demonstrating relatively large conformational fluctuations (up to 7.5Å RMSF). In addition to the extended β-hairpin, there are several additional small α-helices formed in the linker regions between the β-strands that make up the cohesin core. Two of these α-helices formed between residues Glu962 to Leu967 and Val979 to Tyr986 appear in an uncharacteristically long linker region between β3 and β4. Lastly, there are two additional helices formed between residues Leu1004 to Lys1009 and Glu1048 to Tyr1058 which lie β4/5 and β6/7 strands respectively.

5.7 Figures

A



B

Species	Cohesins	Dockerins
Ruminoclostridium cellulolyticum	10	62
Ruminoclostridium herbifermentans	37	80
Acetovibrio thermocellus	29	73
Clostridium acetobutylicum	6	10
Ruminococcus flavefaciens 007c	1	65
Ruminococcus champanellensis	7	154
Ruminococcus callidus	2	76
Ruminococcus albus	1	60

Figure 5.1 High DocGH *Ruminococcus* species show a disparity in cohesin domain counts.

(A) Plot showing the relationship between identifiable dockerin and cohesin domains within known cellulosome producing bacteria. The domains were identified based on their primary sequences using InterProScan and correspond to select bacteria within the *Acetivibrio* (blue)

and *Ruminococcus* (red) genera, both thought to have complex cellulosomes. Enclosed within the dashed box are the *Ruminococcus* bacteria, which are atypical because based on their primary sequences they produce a large number of dockerin containing proteins, but nevertheless are predicted to produce only a few cohesin domains. (B) Table indicating select cellulosome producing microbes and their cohesin and dockerin domain counts. The pie charts represent the number of cohesin (orange) and dockerin (blue) domains present in each species.

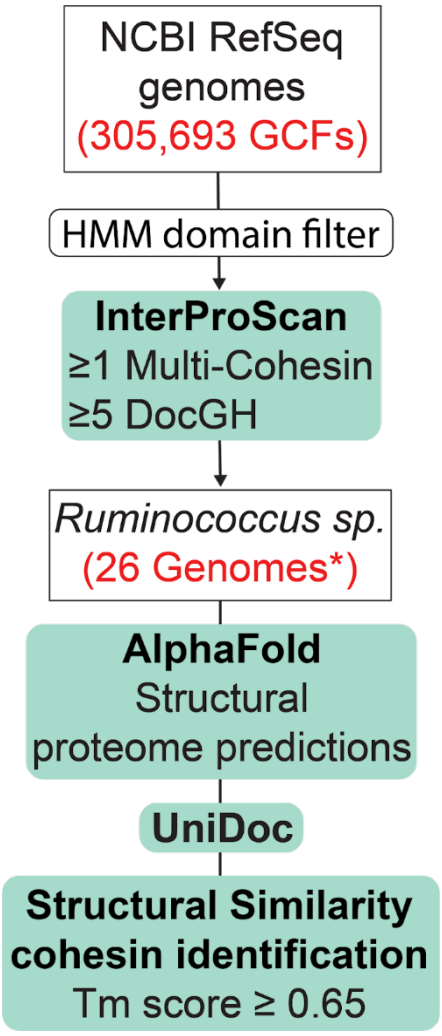


Figure 5.2 AlphaFold-based strategy to identify novel cohesin domains within the proteomes of *Ruminococcus* bacteria.

Strategy used to identify cohesin domains in *Ruminococcus* genomes using AlphaFold. Over 300,000 complete genomes from the NCBI Reference Sequence database (RefSeq) were searched using Hidden Markov Model (HMM) profiles to identify genomes that contained cohesin proteins and DocGH enzymes, hallmarks of cellulosome producing bacteria. A subset of those genomes was then submitted for *de novo* annotation via Prokka (v. 1.14.16) followed by a more stringent analysis for domain identification by InterProScan (v. 5.59-9.10) where 23 *Ruminococcus* genomes had more than one cohesin containing protein and at least five

dockerin proteins. An additional three *Ruminococcus* genomes (*R. hominiciens*, *R. primaciens*, and *R. ruminiciens*) were considered in the analysis as they were recently identified as cellulosome containing *Ruminococcus* species [9]. Protein structures from the 26 genomes were then predicted by AlphaFold, and these structures were parsed into domains by UniDoc. Each domain was structurally compared to representative Coh1, 2, and 3 structures and the similarity evaluated by Tm score.

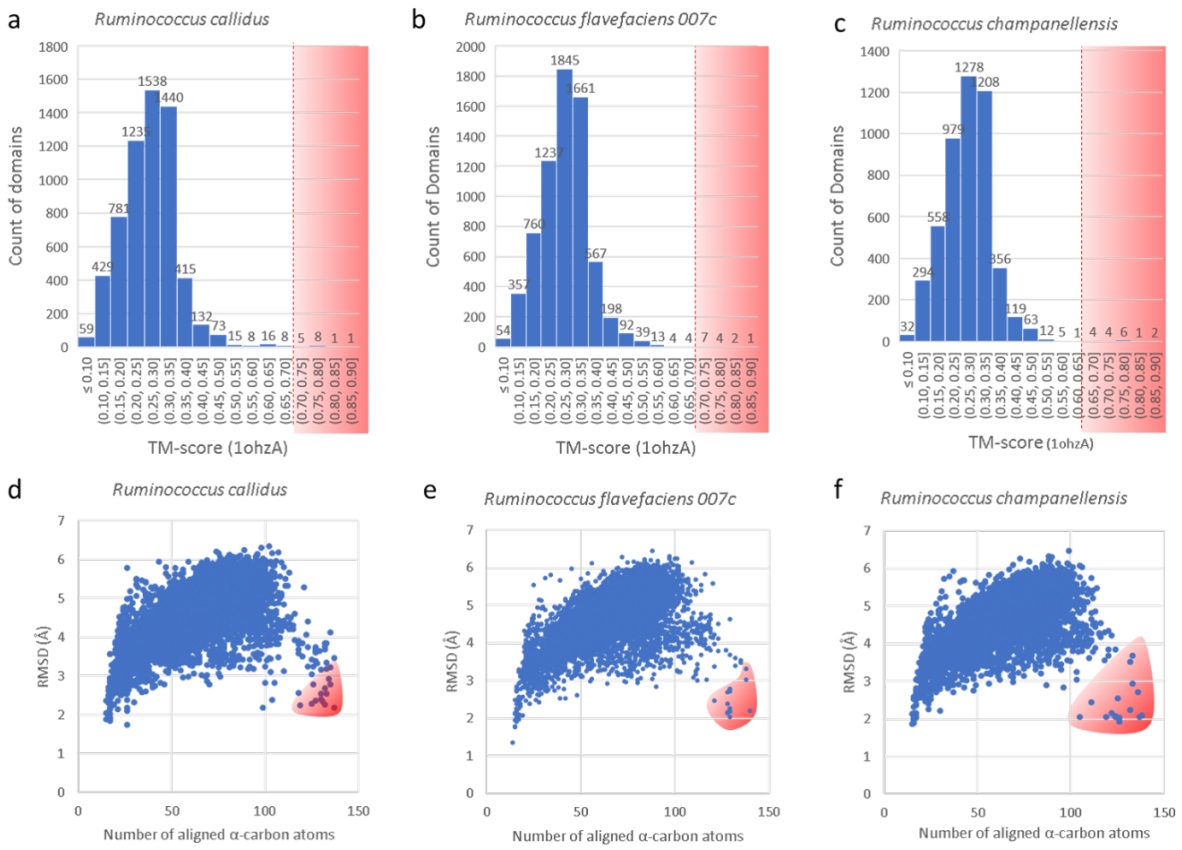


Figure 5.3 Implementation of a structural similarity threshold score to discriminate whether an AlphaFold model qualifies as a putative cohesin domain.

(A) Illustrates the degree to which the predicted structures of the 6164 domains in the *Ruminococcus callidus* proteome resemble the structure of Coh1 (PDB ID 1ohz). The resemblance is plotted as a histogram of TM-scores (structural similarity) scores. The threshold TM-score (dashed red line) was chosen by estimating the dividing line between the low-scoring population (left) and the high-scoring population (right). (B) and (C) illustrate analogous histograms for *Ruminococcus flavefaciens 007c*, and *Ruminococcus champanellensis* proteomes, respectively. The threshold TM-score was between 0.65 and 0.70 for all the organisms examined. (D), (E), and (F) illustrate the use of a different metric to evaluate the resemblance of predicted domain structures to Coh1: a scatter plot of RMSD versus number of aligned alpha-carbon atoms (Nalign). We observed that the sets of putative cohesin domains

identified by TM-score threshold in panels (A)-(C) (points inside red rectangle), were identical to those identified in plots of RMSD versus number of aligned alpha-carbon atoms (Nalign) (points inside red triangles). This correlation between panels (A)-(C) and (D)-(F) illustrate the robustness of our TM-score threshold as a discriminator.

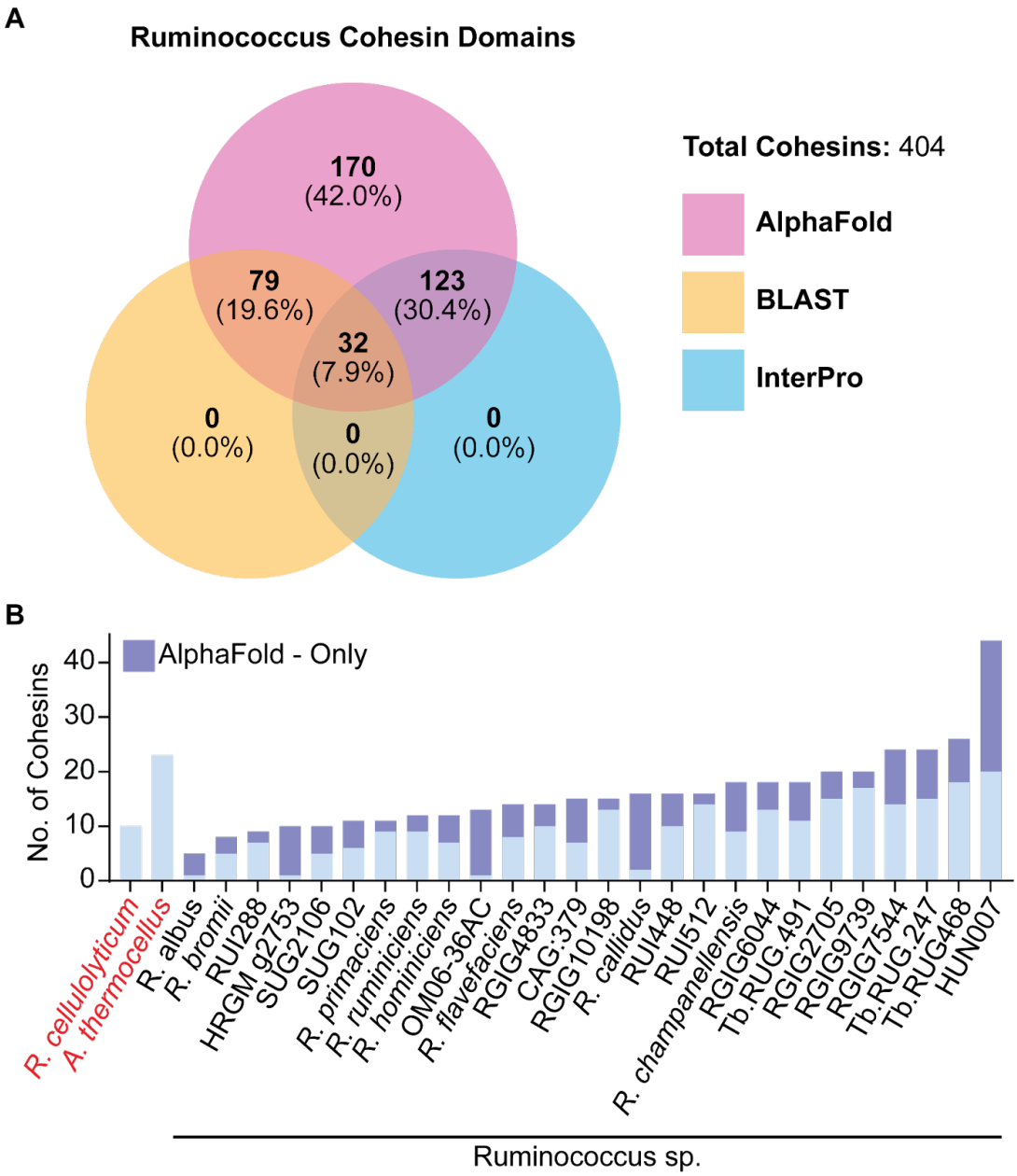


Figure 5.4 AlphaFold identifies a large number of novel cohesin domains.

(A) Venn diagram of the total number of cohesin domains identified by AlphaFold (pink), InterProScan (cyan), and BLAST (yellow) in analyzed *Ruminococcus* species. Two cohesin domains exclusively detected by BLAST were detected in genomes for *R. hominiciens* and *Ruminococcus* sp. RGIG7544 but were excluded as the annotated sequences were severely truncated compared to a full-length cohesin **(B)** Bar graph showing total number of cohesin

domains identified per *Ruminococcus* genome. The cohesin domain counts for *Ruminiclostridium cellulolyticum* and *Acetivibrio thermocellus* are shown as representative non-Ruminococcus species that produce simple and complex cellulosomes, respectively. Total bar height represents the total number of cohesin domains identified across all annotation methods used (AlphaFold, InterProScan, BLAST). The total subset of domains that were exclusively identified as cohesins by AlphaFold (purple) are indicated.

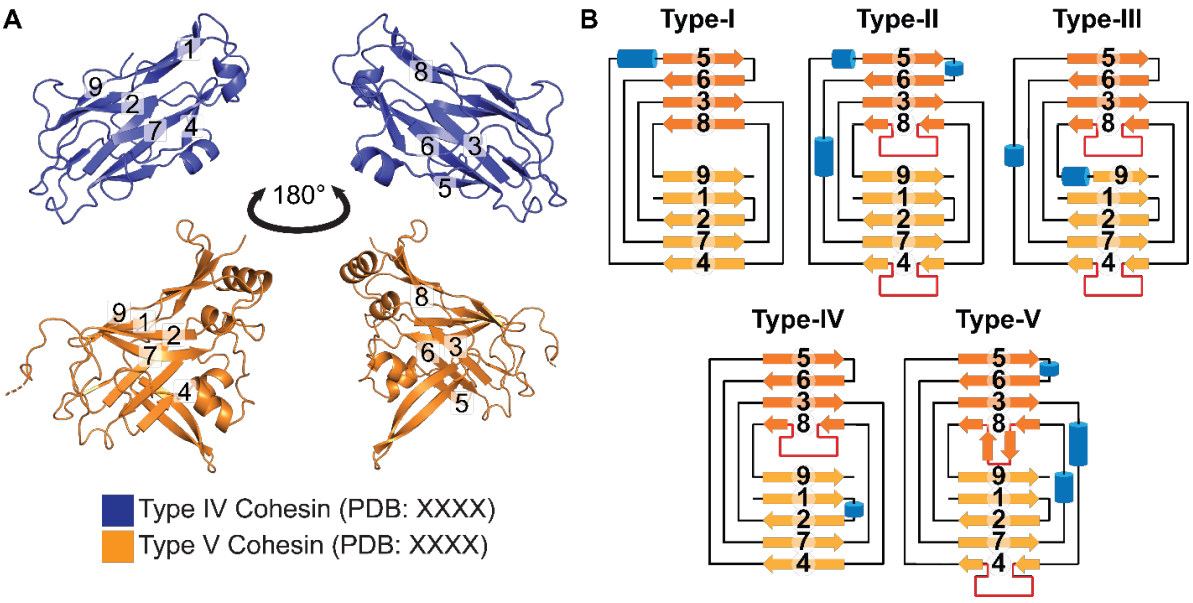


Figure 5.5. **Two novel cohesin families found in *R. callidus***

(A) Representative crystal structures of Type IV and Type V cohesins from *R. callidus*. Rotated cartoon ribbon representation of the type IV cohesin, Rcal_2942_4 (blue), and of the type V cohesin, Rcal_0153_5 (orange). Strand numbers are indicated on each β-sheet face. (B) Topology diagram of novel cohesin families. The flattened 2D-projection of the novel Type-IV and Type V cohesins identified in *R. callidus* are shown in comparison to previously established Type I-III cohesins. All five cohesin families exhibit a core nine-stranded β-sandwich jelly-roll topology with strands 5, 6, 3, and 8 (orange) forming the dockerin-cohesin interface and strands 9, 1, 2, 7, and 4 forming the other sheet face. The presence and length of various accessory helices (blue) highlight the unique structures of each cohesin type. Additional features that deviate from a Type-I cohesin are shown in red for each cohesin type.

Table 5.1 X-ray diffraction data collection and refinement statistics for *R. callidus* cohesins

Crystal	**Rcal0153_5**	**Rcal02942_4**
Data Collection		
Beamline	APS 24-ID-E	APS-24-ID-E
Space group	C2221	P212121
Resolution (Å)	1.80 (1.85-1.80)*	1.20 (1.27-1.20)
Unit cell dimensions: a,b,c (Å)	42.5, 84.6, 118.4	34.0, 37.2, 110.4
Unit cell angles: α,β,γ (°)	90, 90, 90	90, 90, 90
Measured reflections	157327(5854)	564095 (85641)
Unique reflections	19021(976)	84166 (13514)
Overall completeness (%)	94.1 (65.5)	99.4 (98.5)
Overall redundancy	8.3 (6.0)	6.7 (6.3)
Overall R_{merge}	0.128 (0.591)	0.095 (1.03)
$CC_{1/2}$	99.1 (76.8)	99.8 (55.3)
Overall I/δ	11.6 (2.6)	10.3 (1.6)
Structure Refinement		

R_{work} / R_{free}	0.188/ 0.234	0.148 / 0.174
RMSD bond length (Å)	0.006	0.005
RMSD angle (°)	1.2	0.8
Number of protein atoms	1754	2577
Number of water atoms	184	145
Number of other solvent atoms	4	0
Average B-factor of protein ($Å^2$)	18.6	17.1
Average B-factor of water ($Å^2$)	23.6	25.1
Average B-factor other solvent ($Å^2$)	33.8	N/A
PDB ID code	Not yet submitted	Not yet submitted

5.9 References

1. M. Parizadeh and M.-C. Arrieta, “The global human gut microbiome: genes, lifestyles, and diet,” *Trends Mol. Med.*, vol. 29, no. 10, pp. 789–801, Oct. 2023, doi: 10.1016/j.molmed.2023.07.002.

2. S. Fujimori, “Humans have intestinal bacteria that degrade the plant cell walls in herbivores,” *World J. Gastroenterol.*, vol. 27, no. 45, pp. 7784–7791, Dec. 2021, doi: 10.3748/wjg.v27.i45.7784.

3. S. E. Schwartz and G. D. Levine, “Effects of dietary fiber on intestinal glucose absorption and glucose tolerance in rats,” *Gastroenterology*, vol. 79, no. 5 Pt 1, pp. 833–836, Nov. 1980.

4. K. Makki, E. C. Deehan, J. Walter, and F. Bäckhed, “The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease,” *Cell Host Microbe*, vol. 23, no. 6, pp. 705–715, Jun. 2018, doi: 10.1016/j.chom.2018.05.012.

5. S. Andreu-Sánchez *et al.*, “Global genetic diversity of human gut microbiome species is related to geographic location and host health,” *Cell*, vol. 188, no. 15, pp. 3942-3959.e9, Jul. 2025, doi: 10.1016/j.cell.2025.04.014.

6. C. M. Chibani *et al.*, “A catalogue of 1,167 genomes from the human gut archaeome,” *Nat. Microbiol.*, vol. 7, no. 1, pp. 48–61, Jan. 2022, doi: 10.1038/s41564-021-01020-9.

7. L. Artzi, E. A. Bayer, and S. Moraïs, “Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides,” *Nat. Rev. Microbiol.*, vol. 15, no. 2, pp. 83–95, Feb. 2017, doi: 10.1038/nrmicro.2016.164.

8. V. D. Alves, C. M. G. A. Fontes, and P. Bule, “Cellulosomes: Highly Efficient Cellulolytic Complexes,” *Subcell. Biochem.*, vol. 96, pp. 323–354, 2021, doi: 10.1007/978-3-030-58971-4_9.

9. S. Moraïs *et al.*, “Cryptic diversity of cellulose-degrading gut bacteria in industrialized humans,” *Science*, vol. 383, no. 6688, p. eadj9223, Mar. 2024, doi: 10.1126/science.adj9223.

10. R. Lamed, J. Naimark, E. Morgenstern, and E. A. Bayer, “Specialized cell surface structures in cellulolytic bacteria,” *J. Bacteriol.*, vol. 169, no. 8, pp. 3792–3800, Aug. 1987, doi: 10.1128/jb.169.8.3792-3800.1987.

11. R. Lamed, E. Setter, and E. A. Bayer, “Characterization of a cellulose-binding, cellulase-containing complex in Clostridium thermocellum,” *J. Bacteriol.*, vol. 156, no. 2, pp. 828–836, Nov. 1983, doi: 10.1128/jb.156.2.828-836.1983.

12. S. Yoav *et al.*, “How does cellulosome composition influence deconstruction of lignocellulosic substrates in Clostridium (Ruminiclostridium) thermocellum DSM 1313?,” *Biotechnol. Biofuels*, vol. 10, p. 222, 2017, doi: 10.1186/s13068-017-0909-7.

13. N. Bhardwaj, B. Kumar, K. Agrawal, and P. Verma, “Current perspective on production and applications of microbial cellulases: a review,” *Bioresour. Bioprocess.*, vol. 8, no. 1, p. 95, Oct. 2021, doi: 10.1186/s40643-021-00447-6.

14. A. Sharma, R. Tewari, S. S. Rana, R. Soni, and S. K. Soni, “Cellulases: Classification, Methods of Determination and Industrial Applications,” *Appl. Biochem. Biotechnol.*, vol. 179, no. 8, pp. 1346–1380, Aug. 2016, doi: 10.1007/s12010-016-2070-3.

15. C. M. Minor *et al.*, “A genomic analysis reveals the diversity of cellulosome displaying bacteria,” *Front. Microbiol.*, vol. 15, p. 1473396, 2024, doi: 10.3389/fmicb.2024.1473396.

16. S. Pagès *et al.*, “Species-specificity of the cohesin-dockerin interaction between Clostridium thermocellum and Clostridium cellulolyticum: prediction of specificity determinants of the dockerin domain,” *Proteins*, vol. 29, no. 4, pp. 517–527, Dec. 1997.

17. R. Haimovitz *et al.*, “Cohesin-dockerin microarray: Diverse specificities between two complementary families of interacting protein modules,” *Proteomics*, vol. 8, no. 5, pp. 968–979, Mar. 2008, doi: 10.1002/pmic.200700486.

18. B. Dassa *et al.*, “Rumen cellulosomics: divergent fiber-degrading strategies revealed by comparative genome-wide analysis of six ruminococcal strains,” *PloS One*, vol. 9, no. 7, p. e99221, 2014, doi: 10.1371/journal.pone.0099221.

19. S. Moraïs *et al.*, “Lysozyme activity of the Ruminococcus champanellensis cellulosome,” *Environ. Microbiol.*, vol. 18, no. 12, pp. 5112–5122, Dec. 2016, doi: 10.1111/1462-2920.13501.

20. A. Almeida *et al.*, “A new genomic blueprint of the human gut microbiota,” *Nature*, vol. 568, no. 7753, pp. 499–504, Apr. 2019, doi: 10.1038/s41586-019-0965-1.

21. O. Yaniv, Y. Halfon, L. J. W. Shimon, E. A. Bayer, R. Lamed, and F. Frolow, “Structure of CBM3b of the major cellulosomal scaffoldin subunit ScaA from Acetivibrio cellulolyticus,” *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.*, vol. 68, no. Pt 1, pp. 8–13, Jan. 2012, doi: 10.1107/S174430911104807X.

22. P. Bule *et al.*, “Higher order scaffoldin assembly in Ruminococcus flavefaciens cellulosome is coordinated by a discrete cohesin-dockerin interaction,” *Sci. Rep.*, vol. 8, no. 1, p. 6987, May 2018, doi: 10.1038/s41598-018-25171-8.

23. I. Noach *et al.*, “Crystal structure of a type-II cohesin module from the Bacteroides cellulosolvens cellulosome reveals novel and distinctive secondary structural elements,” *J. Mol. Biol.*, vol. 348, no. 1, pp. 1–12, Apr. 2005, doi: 10.1016/j.jmb.2005.02.024.

24. P. Bule *et al.*, “Single Binding Mode Integration of Hemicellulose-degrading Enzymes via Adaptor Scaffoldins in Ruminococcus flavefaciens Cellulosome,” *J. Biol. Chem.*, vol. 291, no. 52, pp. 26658–26669, Dec. 2016, doi: 10.1074/jbc.M116.761643.

25. K. Cameron *et al.*, “Cell-surface Attachment of Bacterial Multienzyme Complexes Involves Highly Dynamic Protein-Protein Anchors,” *J. Biol. Chem.*, vol. 290, no. 21, pp. 13578–13590, May 2015, doi: 10.1074/jbc.M114.633339.

APPENDIX

The work described in the Appendix features on-going, incomplete, and/or otherwise unpublished work that I have carried out during the time of my graduate studies. Additional details, experimental procedures and data analysis for material covered in the appendix can be accessed from my digital scientific notebook:

https://ucla-site.signalsresearch2.revvitycloud.com/elements/entity/journal:42df48ee-bcc4-468a-a6b6-c700b7c44ab3?focus=experiment%3A2b46fb6a-abb4-4d60-9022-c78f99a09353

A.1 Search for RsgI9's cognate SigI partner

A.1.2 Overview

RsgI9 is unique from the other eight RsgI present in *C. thermocellum* in several ways: (1) RsgI9 is an "orphan" anti-σ factor, having no known SigI encoded within the same operon or gene neighborhood. (2) The cytoplasmic N-terminal portion of RsgI9 contains an additional 95 residue, low-complexity insert sequence between the N-terminal anti-σ factor domain and the transmembrane helix. (3) The ectodomain of RsgI9 does not contain any domains predicted to bind to or have activity against polysaccharides like six of the remaining RsgI [1]. In order to find a potential "SigI9", I expressed and purified a 6xHis-tagged N-terminal RsgI9 anti-σ factor domain ($RsgI9^{1-54}$) as bait for a set of pull-down experiments using cell lysate collected from *C. thermocellum* grown in cellobiose. From these experiments, I was unable to find a corresponding SigI9 protein, with the most abundant proteins seen from the LC-MS/MS corresponding to a chaperonin and an enolase.

A.1.2 Results

To identify potential binding partners to the N-terminal anti-σ factor domain of RsgI9, I constructed a C-terminal 6xHis-tagged RsgI9$^{1-54}$ protein to serve as a bait for affinity pull-down assays. After the bait RsgI9$^{1-54}$ was loaded onto a Co-NTA affinity column, lysate harvested from *C. thermocellum* during mid-log phase (OD ~ 0.750) was flowed over the resin, washed, and subsequently eluted for visualization by SDS-PAGE (Figure 7.1). In all replicates, RsgI9$^{1-54}$ co-eluted with two other species at molecular weights of approximately 50 kDa and 60 kDa, respectively. These gel-bands were excised and submitted for LC-MS/MS identification. In the 60 kDa protein band, the highest abundance proteins seen corresponded to a 60 kDa chaperonin (*groL*) protein and an Arginine-tRNA ligase (*argS*). The presence of the chaperonin protein would suggest that the RsgI9$^{1-54}$ construct is not properly folded, despite the removal of the large cytoplasmic insert. In the 50 kDa band sample, the most abundant protein observed was an enolase (*eno*). Further studies by either CD-spectroscopy or NMR of the RsgI9$^{1-54}$ construct are needed to confirm if it is unfolded before these affinity experiments can be repeated. qPCR experiments carried out in *C. thermocellum* suggest that SigI levels remain relatively low when the activating signal substrate for the RsgI is not present. As such, additional experiments where lysates are grown in different polysaccharide-containing growth media are required as any potential SigI binding partners to RsgI9 may be in too low abundance to be seen in cellobiose growth conditions.

A.2 Investigation of RsgI/SigI crosstalk in *C. thermocellum*

A.2.1. Overview

The regulomes controlled by ECF/anti-σ factors are tightly regulated, with very few cases of non-cognate crosstalk between them [2]. However, previous work from Chrstine Minor of the Clubb lab (unpublished) has suggested that the RsgI/SigI binding pairs within *C. thermocellum* may be more promiscuous than previously thought. She utilized a split tripartite-GFP reporter using N-terminal RsgI anti-σ factor ($RsgI^N$) and C-terminal SigI ($SigI^C$) domains taken from *C. thermocellum* fused to the split GFP fragments. Using Fluorescence-Activated Cell Sorting (FACS) as a readout, she was able to assess *in vitro* RsgI:SigI binding partners in a high-throughput manner for many of the possible 72 RsgI/SigI pairs. Interestingly, her data suggested that some noncognate pairs could interact, particularly RsgI5 and SigI2. In *C. thermocellum*, RsgI5 is predicted to bind to and control genes responsible for the degradation of arabinoxylan, due to the presence of a CBM42 in RsgI5's ectodomain [1]. Conversely, RsgI2 has a predicted CBM3 module as its ectodomain, suggesting it is responsible for regulating cellulose-degrading DocGHs [3]. The presence of crosstalk in the RsgI/SigI system of *C. thermocellum*, could suggest a more complex, orchestrated regulatory response to determine the simultaneous expression of varied cellulosomal genes. To test this crosstalk, I used 2D-NMR to investigate and validate the positive interaction seen between a ^{15}N-labeled RsgI5N and $SigI2^C$. The results of these experiments showed a low affinity between the tested pair but difficulties in SigI2 expression and purification may suggest the protein was misfolded during titration experiments.

A.2.2 Results & Conclusions

In order to orthogonally confirm the results of the FACS-based assay, I used 2D-NMR chemical shift perturbation (CSP) experiments utilizing purified ^{15}N-labeled RsgI^{N} and SigI^{C} proteins to demonstrate *in vitro* binding, specifically of RsgI5 and SigI2 (Figure 7.2). The apo ^{15}N-RsgI5^{N} sample was stably expressed and showed a well-dispersed 2D-TROSY-HSQC spectra, suggesting the N-terminal anti-σ factor fragment is folded. Further, the number of amide resonances observed within the spectra (~50 peaks) corresponds well with the number of residues within the RsgI5^{N} fragment, indicating that the sample is also of sufficient purity with minimal degradation products present. Upon titration of unlabeled SigI2^{C} in a 1:2 RsgI:SigI ratio, there were a minimal number of CSPs and broadened peaks observed, indicating relatively weak and/or non-specific binding. While the results of these experiments contradict the previous FACS-assay data, the instability and purity of the SigI2^{C} sample may have led to the negative results seen in the 2D-TROSY-HSQC. Previous work with truncated SigI constructs from *C. thermocellum* have noted issues in stability [4]. Indeed, the SigI2^{C} construct showed signs of sample degradation on SDS-PAGE and visible aggregate within 12 hours. Further, purification of the SigI2^{C} protein proved challenging due to the presence of an additional ~20 kDa co-contaminant that could not be removed via size exclusion chromatography or stringent affinity-column washes. An unfolded purification method using 8 M urea was required in order to remove the contaminant from the unlabeled SigI2^{C} sample, which significantly reduced yield and potentially led to improper re-folding of the SigI. Future experiments with this system would require further validation that the purified SigI2^{C} protein is folded using NMR or CD-spectroscopy. Improved yields and purity could also be possible using reverse-phase HPLC purification of the SigI2^{C} sample but it would require more intensive refolding procedures. Lastly, the remaining RsgI/SigI pairs that were identified as potential cross-talk partners need to be validated using NMR or other biophysical approaches such as ITC or SPR.

A.3 Figures

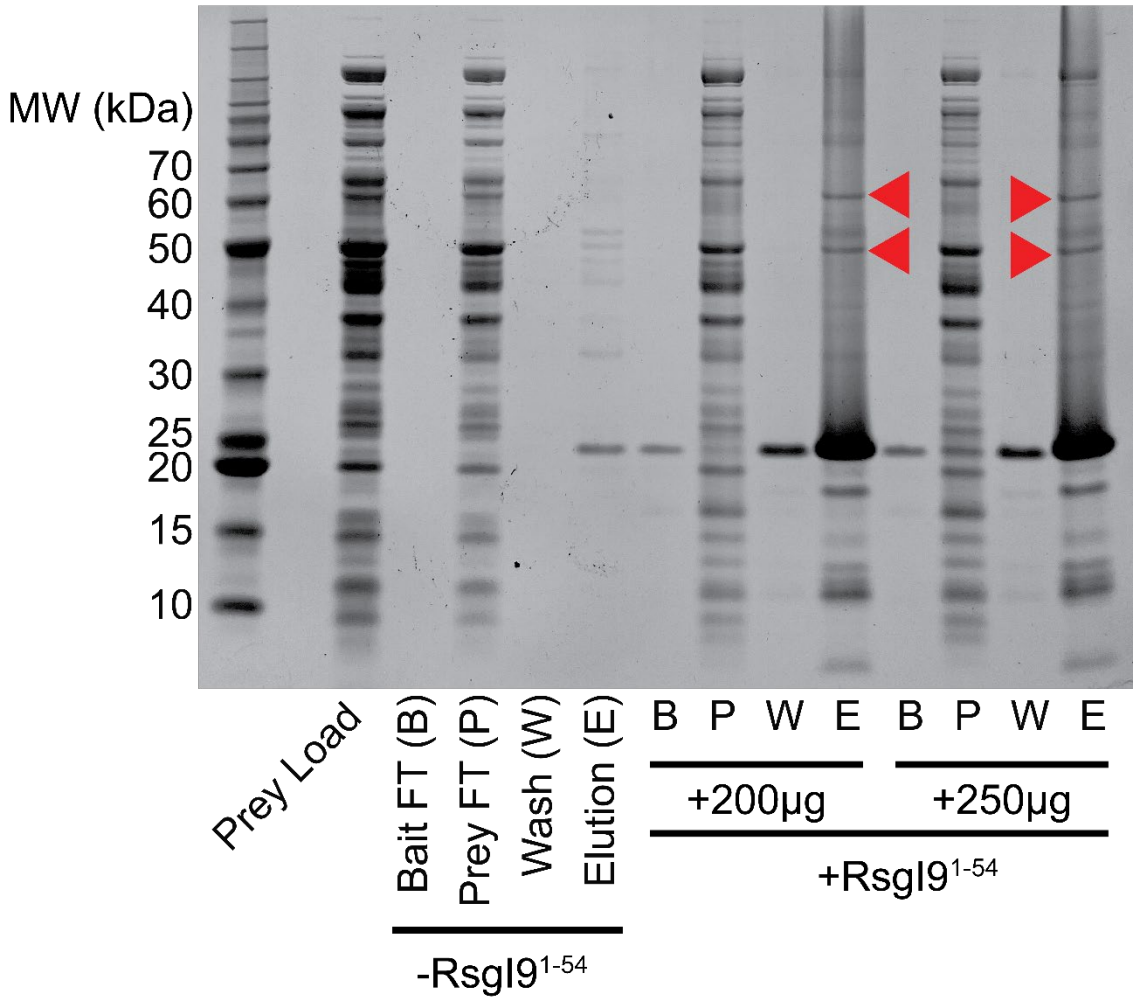


Figure A.1 Affinity Pull-Down Search for "SigI9"

SDS-PAGE gel for affinity pull-down experiments done with an immobilized RsgI9$^{1-54}$ construct and lysate harvested from *C. thermocellum* at mid-log phase (OD: 0.750). The flow-through samples for the bait (B), prey (P), washes (W), and final elution (E) are shown for pull-down experiments with no bait (-RsgI9$^{1-54}$) and increasing concentrations of bait. The co-elution bands observed at 50 kDa and 60 kDa molecular weights are indicated with red arrows.

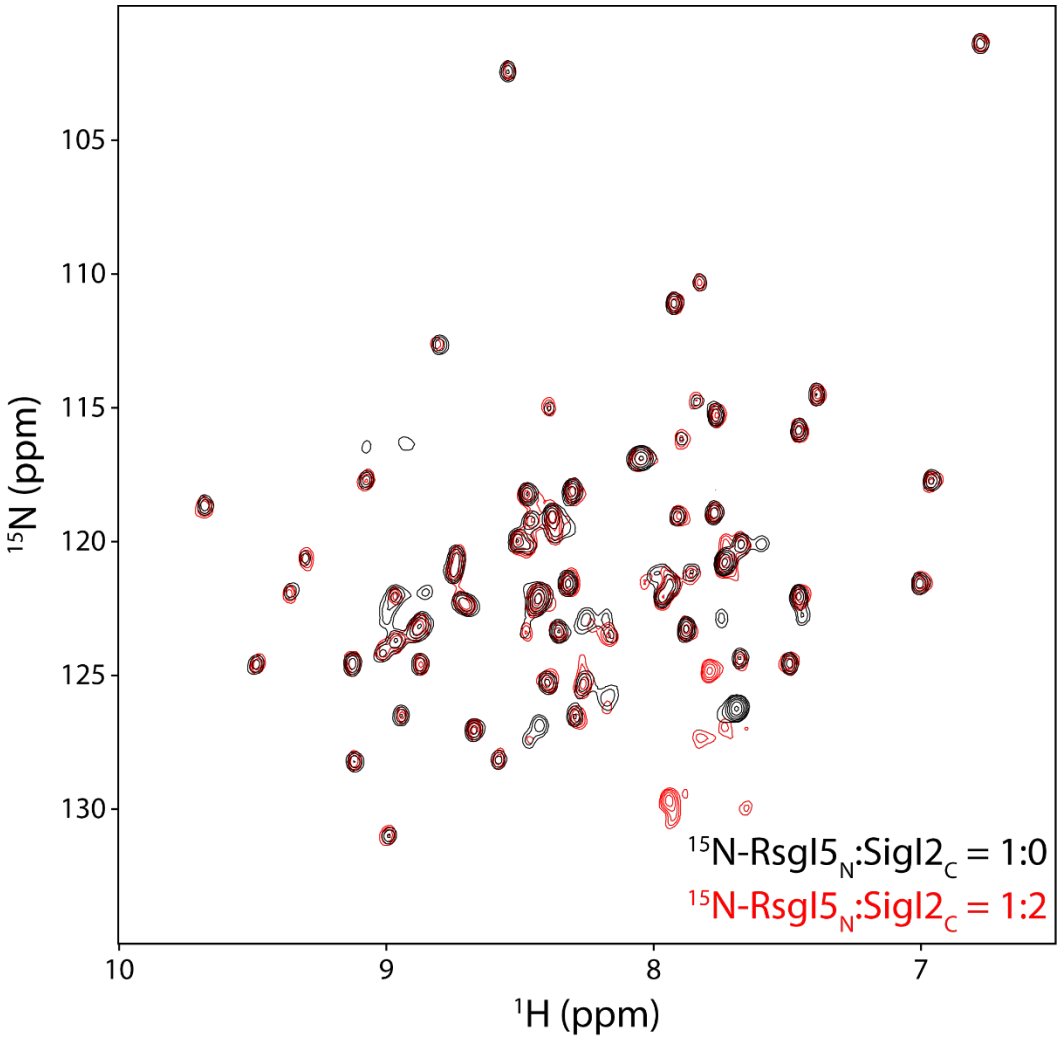


Figure A.2 Chemical Shift Perturbation Studies with RsgI5^N and SigI2^C

Two overlayed 2-D TROSY-HSQC NMR spectra of ^{15}N-labeled RsgI5^N titrated with (black) and without purified SigI2^C in a 1:2 ratio (red).

A.3 References

1. Y. Nataf *et al.*, “Clostridium thermocellum cellulosomal genes are regulated by extracytoplasmic polysaccharides via alternative sigma factors,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 43, pp. 18646–18651, Oct. 2010, doi: 10.1073/pnas.1012175107.
2. V. A. Rhodius *et al.*, “Design of orthogonal genetic switches based on a crosstalk map of σs, anti-σs, and promoters,” *Mol. Syst. Biol.*, vol. 9, p. 702, Oct. 2013, doi: 10.1038/msb.2013.58.
3. H. Wei *et al.*, “Comparison of transcriptional profiles of Clostridium thermocellum grown on cellobiose and pretreated yellow poplar using RNA-Seq,” *Front. Microbiol.*, vol. 5, p. 142, 2014, doi: 10.3389/fmicb.2014.00142.
4. Z. Wei *et al.*, “Alternative σI/anti-σI factors represent a unique form of bacterial σ/anti-σ complex,” *Nucleic Acids Res.*, vol. 47, no. 11, pp. 5988–5997, Jun. 2019, doi: 10.1093/nar/gkz355.