UNIVERSITY OF CALIFORNIA

Santa Barbara

Realizing the Biotechnological Potential of Fungal Cellulosomes

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Chemical Engineering

by

Stephen Peter Lillington

Committee in charge:

Professor Michelle A. O’Malley, Co-Chair

Professor M. Scott Shell, Co-Chair

Professor Arnab Mukherjee

Professor Kevin W. Plaxco

June 2023

The dissertation of Stephen Peter Lillington is approved.

Arnab Mukherjee

Kevin W. Plaxco

Michelle A. O’Malley, Committee Co-Chair

M. Scott Shell, Committee Co-Chair

May 2023

Realizing the Biotechnological Potential of Fungal Cellulosomes

by

Stephen Peter Lillington

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all the people who supported me throughout my doctoral journey. First, I want to thank my undergraduate research advisors and mentors, Professor Keith Tyo and Dr. Will Bothfeld, who gave me my first research opportunity and who catalyzed my scientific career.

To my Ph.D. advisors, Professors Michelle O'Malley and Scott Shell, thank you for your unfaltering support and encouragement throughout my intellectual and professional journeys. My Ph.D. experience was incredibly enriching, in large part because of the unique access I had to your diverse perspectives on research problems. I knew from the moment I learned as a prospective, visiting student about the joint position in your labs that I wanted it, and I am so glad I took on the project. I would also like to thank my committee members, Professor Arnab Mukherjee and Professor Kevin Plaxco who provided constructive guidance and innovative ideas for elevating the impact and rigor of my research.

Thank you to my lab mates who made work fun, even when it meant late evenings at the lab bench, especially Susanna Seppala, Tom Lankiewicz, Tejas Navaratna, and Pat Leggieri. Thank you to my friends, especially my Northwestern squad, who gave me invaluable support and joy as I navigated these last five years.

Finally, to the most important people in my life, Katy, Mom, Dad, Dave – thank you for your unwavering support of my dreams and for serving as role models for me. I wouldn't have achieved this milestone without you all.

VITA OF STEPHEN PETER LILLINGTON
May 2023

EDUCATION

2018-2023 Doctor of Philosophy in Chemical Engineering
University of California, Santa Barbara

2013-2018 Bachelor of Science in Chemical Engineering
Northwestern University

PROFESSIONAL EMPLOYMENT

2018-2023 Graduate Student Researcher and Teaching Assistant
Department of Chemical Engineering, UC Santa Barbara
Santa Barbara, CA
Academic Advisors: Michelle A. O'Malley and M. Scott Shell

2016-2018 Undergraduate Research Assistant
Department of Chemical Engineering, Northwestern University
Evanston, IL
Advisor: Keith E.J. Tyo

2015-2016 Marketing and Continuous Improvement Co-op
Amcor Flexibles
Mundelein, IL

PUBLICATIONS

1. T.S. Lankiewicz, H. Choudhary, Y. Gao, B. Amer, **S.P. Lillington**, P.A. Leggieri, J.L. Brown, C.L. Swift, A. Lipzen, H. Na, M. Amirebrahimi, M.K. Theodorou, E.K. Baidoo, K. Barry, I.V. Grigoriev, V.I. Tomhokin, J. Gladden, S. Singh, J.C. Mortimer, J. Ralph, B.A. Simmons, S.W. Singer, M.A. O'Malley, "Lignin deconstruction by anaerobic fungi," *Nature Microbiology*, 2023.
2. T.S. Lankiewicz, **S.P. Lillington**, M.A. O'Malley, "Lignocellulosic enzyme discovery in anaerobic fungi (Neocallimastigomycetes) enables biorefinery innovation," *Microbiology and Molecular Biology Reviews*, 2022.
3. **S.P. Lillington**, W. Chrisler, C.H. Haitjema, S.P. Gilmore, C.R. Smallwood, V. Shutthanandan, J.E. Evans, M.A. O'Malley, "Cellulosome localization patterns vary across life stages of anaerobic fungi," *mBio*, 2021.
4. **S.P. Lillington**, P. Leggieri, K. Heom, M.A. O'Malley, "Nature's recyclers: anaerobic microbial communities drive crude biomass deconstruction," *Current Opinion in Biotechnology*, 2020.
5. S.P. Gilmore, **S.P. Lillington**, C.H. Haitjema, R. de Groot, M.A. O'Malley, "Designing chimeric enzymes for synthetic fungal cellulosomes," *Synthetic and Systems Biotechnology,* 2020.

PUBLICATIONS IN PREPARATION

1. **S.P. Lillington**, K. Deng, A. Parvate, H.M. Olson, S. Jin, J.E. Evans, M.A. O'Malley, "Composition-activity relationships governing lignocellulase activity of the *Neocallimastix californiae* cellulosome," *In preparation.*
2. **S.P. Lillington**, M.S. Shell, M.A. O'Malley, "Model-guided engineering of fungal cellulosome parts for pH-responsive protein assembly," *In preparation*.
3. A. Dementiev*, **S.P. Lillington***, R. Jedrzejczak, L. Welk, A. Joachimiak, M.A. O'Malley, "Structure and enzymatic characterization of CelD cellulase from the anaerobic fungus *Piromyces finnis*," *In preparation.* *Equal contribution

SELECTED PRESENTATIONS

Oral Presentations

1. **S.P. Lillington**, M.S. Shell, M.A. O'Malley, "Model-guided engineering of fungal cellulosome parts for pH-responsive protein assembly," *Annual Meeting of the American Institute of Chemical Engineers (AIChE)*, Phoenix, AZ. November 14, 2022.
2. **S.P. Lillington**, H.M. Olson, M.S. Shell, M.A. O'Malley, "Deciphering composition-activity relationships of fungal cellulosomes," *Symposium on Biomaterials, Fuels, and Chemicals*, New Orleans, LA. May 2, 2022.
3. **S.P. Lillington**, M.S. Shell, M.A. O'Malley, "Stimuli-responsive protein complexes inspired by anaerobic fungal cellulosomes," *American Chemical Society Spring National Meeting*, San Diego, CA. March 18, 2022.

Poster Presentations

1. **S.P Lillington**, W. Chrisler, J.E. Evans, M.A. O'Malley, "Cellulosome localization patterns vary across life stages of anaerobic fungi." *ICBE,* Santa Barbara, CA. January 6-9, 2021.
2. **S.P. Lillington**, M.S. Shell, M.A. O'Malley, "Comprehensive characterization of cellulosomes from anaerobic fungi." *Gordon Research Conference on CAZymes*, Andover, NH. July 20-25, 2019.
3. **S.P. Lillington**, W. Bothfeld, K.E.J. Tyo. "Understanding preferential consumption of aromatic compounds in *Acinetobacter baylyi* ADP1." *AIChE Annual Meeting*, Minneapolis, MN. October 30, 2017.
4. W. Bothfeld, **S.P. Lillington**, K.E.J. Tyo. "Optimization of dual alkylation/silylation derivatiziation method for quantitative metabolomics." *EBRC Spring Retreat*, Evanston, IL. March 24-25, 2017.

HONORS AND AWARDS

2020	Dow Discovery Fellowship Finalist, UCSB Dept. of Chemical Engineering
2018	Heslin Fellowship, UCSB Dept. of Chemical Engineering
2017	AIChE Undergraduate Poster Session, 1st place in Bioengineering
2017	Alumnae of Northwestern Undergraduate Research Grant

ABSTRACT

Realizing the Biotechnological Potential of Fungal Cellulosomes

by

Stephen Peter Lillington

Rising risks of climate change and supply chain insecurity highlight the need to develop alternative, greener synthesis routes to common materials currently sourced from petroleum. Biological systems excel at interconverting chemicals with exquisite specificity and speed, using networks of enzymes that perform catalysis at mild conditions. Protein complexes in nature colocalize complementary subunits to perform sophisticated biochemistry, and artificial, spatial organization of enzyme systems into synthetic complexes is an attractive strategy for improving biocatalytic process throughputs in industrial settings. While some sets of modular parts that enable designer protein complex construction exist, there is still a need to develop new components that are widely compatible with different enzymes and that are highly engineerable to impart desired self-assembly properties.

Fungal cellulosomes, modular protein machines produced by anaerobic fungi in the guts of herbivores to rapidly free sugars from plant matter, represent an unexplored framework for synthetic protein complex construction. Cellulosomes synergistically incorporate enzymes involved in biomass degradation into discrete complexes via modular protein-protein interactions between enzyme fused dockerin domains and cohesin domains repeated on a central scaffoldin protein. Over 80% of the degradative power anaerobic fungi possess is attributed to cellulosomes, but the mechanistic nature of their activity and their assembly mechanism remain unknown. These knowledge gaps have precluded the development of fungal cellulosomes or their parts as biocatalytic technologies with real world applications.

We apply a range of experimental techniques towards addressing how cellulosomes are produced in native anaerobic fungal cultures and characterizing the composition, nanostructure, and biochemical activity of purified, native cellulosomes. Immunofluorescence microscopy with cellulosome-labeling antibodies shows cellulosomes localize to the surfaces of cells, but that only cells at certain stages of the multi-staged life cycle produce cellulosomes under specific growth conditions. A robust cellulosome purification method we developed, in conjunction with mass spectrometry-based proteomics and biomass hydrolysis kinetic assays, provides high resolution details into the composition and lignocellulolytic activities of isolated cellulosomes produced by an anaerobic fungus, advancing our understanding of how cellulosomes can be engineered to enhance biomass hydrolysis rates.

Towards leveraging the modular cellulosome assembly framework for synthetic biology applications, we develop a suite of modular interacting parts for constructing protein complexes with fungal cellulosome proteins. Through a combination of molecular modeling and high-throughput screening, we engineer interacting domains with a range of pH dependent binding behaviors for building protein complexes whose composition and therefore function are modulated with and environmental trigger, pH. Together, these tools and insights shed light on how cellulosomes make anaerobic fungi prolific biomass degraders and provide a framework for engineering protein complexes inspired by fungal cellulosomes designed for a wide range of applications.

TABLE OF CONTENTS

I. Introduction

Parts of this chapter are adapted from S.P. Lillington, P.A. Leggieri, K.H. Heom, M.A. O'Malley, *Current Opinion in Biotechnology* **62** (2020). Copyright 2020, All rights reserved. Other parts are adapted from Stephen P. Lillington, et al, *mBio* 2021 [1]. Reprinted under a CC-BY Creative Commons license.

1.1 Abbreviations referenced in this thesis

CAZyme: Carbohydrate-active enzyme

GH: Glycoside hydrolase

CE: Carbohydrate esterase

GT: Glycosyl transferase

PL: Polysaccharide lyase

AA: Auxiliary activities

CBM: Carbohydrate binding module

MD: Molecular dynamics

REMD: Replica exchange molecular dynamics

1.2 Motivation

Recent global events and the looming threats of climate change emphasize the importance of diversifying chemical and material supply chains away from fossil fuels and towards natural resources that are domestically abundant, stable in price, and more environmentally sustainable. Biology is a master of chemical conversion, leveraging networks of fast, highly specific enzymes to convert sugars into useful products in aqueous environments at moderate temperatures. This is in stark contrast to traditional chemical processes that often require harsh conditions and hazardous materials, making them energy- and water-intensive. Biocatalytic processes that leverage engineered cells or enzyme systems for industrial scale production of chemicals typically sourced from petroleum thus present an attractive alternative strategy for manufacturing economically critical materials in a greener way.

While biomanufacturing as a field and industry has grown dramatically in recent years, few commercially successful products are currently on the market [2], and the primary barrier to success is higher cost of production relative to petroleum-based routes. In a typical bioprocess, the major costs are the cost of enzymes in cell-free bioprocessing and feedstock cost for fermentative bioprocessing [3]. Thus, any new innovations producing faster, more stable enzymes, increasing overall production per unit enzyme per unit time, or innovations that enable the use of cheaper, even negative cost feedstocks like waste materials, stand to accelerate the broader deployment of bioprocesses.

Lignocellulosic biomass, which makes up plant material, is an attractive low cost waste material for use in bioprocessing. Because it is a polymeric material, lignocellulosic bioprocessing requires a two-step process of saccharification and fermentation of sugars into product. Though conventional platform microbes cannot utilize lignocellulose, microbial biomass degradation occurs on a massive scale in the biosphere, and many non-model microbes have evolved highly efficient enzyme systems for hydrolyzing lignocellulose into sugars [4]. Mining these communities for superior enzymes that can be dropped into industrial bioprocesses to efficiently convert waste lignocellulose into sugar then a valuable product is a promising path towards realizing profitable bioprocesses.

Anaerobic fungi found in the guts of herbivores represent one class of biomass degrading organisms from which better enzymes for lignocellulose bioprocessing may be sourced [5]. Anaerobic fungi hold the largest and one of the most diverse catalogs of carbohydrate-active enzymes (CAZymes) within their genomes (mycocosm.jgi.doe.gov). To efficiently hydrolyze lignocellulose into sugars, anaerobic fungi produce and secrete a variety of CAZymes that modularly self-assemble into complexes called cellulosomes. By bringing enzymes with complementary activities together, cellulosomes greatly enhance the rate and extent of lignocellulose degradation by CAZymes [6], making cellulosomes attractive technologies for bioprocessing. However, very few anaerobic fungal enzymes are experimentally characterized, and little is known about how fungal cellulosomes are produced, how they assemble, what their composition is, or how they biochemically function. Anaerobic fungi are difficult to culture in the lab, their proteins often do not express well in heterologous hosts, and no method for cellulosome purification at sufficient purity and yield for structural and kinetic characterization has been published. Understanding fungal cellulosome assembly and composition-structure relationships is key to engineering cellulosomes optimized for specific bioprocesses. These knowledge gaps have prevented most development of anaerobic fungal enzyme machinery, especially cellulosomes, into useful biotechnologies.

Fungal cellulosomes are interesting not only because of their potential in lignocellulose bioprocessing but also because of their self-assembly behavior. Modular, enzyme-fused dockerin domains mediate fungal cellulosome assembly, and these domains can be grafted onto many enzymes of interest to facilitate their incorporation into protein complexes [7]. Spatially organizing enzymes into complexes as fungal cellulosomes do is a strategy that has demonstrated improved enzyme system biocatalytic performance using a variety of assembling parts in many functional contexts besides biomass hydrolysis [8,9]. While existing parts for modular enzyme assembly have demonstrated many successes, there may not be a universal parts set for synthetic protein complex construction, as the mechanisms of rate enhancement are inconsistent and complex [10], and sometimes certain enzymes and assembly domains are incompatible [11,12]. Thus, new additions to the toolkit for constructing synthetic protein complexes remain highly desirable for designing optimized biocatalytic enzyme systems for wide-ranging applications. The protein domains mediating fungal cellulosome assembly represent one such parts set that has not been characterized.

While an encouraging platform, engineering enzyme complexes based on the fungal cellulosome framework, synthetic or natural, is a challenging problem that requires a molecular-level understanding of biophysics and biochemistry, as complex intra- and intermolecular interactions drive enzyme self-assembly and cooperation. Additionally, the sequence space for engineering proteins with desired functions is massive, meaning rational protein engineering is not always feasible. Thus, realizing the biotechnological potential of fungal cellulosomes by characterizing their structure, composition, and biochemistry to understand their native function, and by engineering fungal cellulosome parts to construct synthetic protein complexes with broad applications, requires a multi-pronged investigational approach. The development of molecular modeling frameworks as well as experimental, high-throughput screening techniques for fungal cellulosome characterization and engineering represent important steps towards this goal.

1.3 Organization of the dissertation

This dissertation contains six chapters. The first introduces the enzyme systems and organisms involved in natural lignocellulose hydrolysis, with a particular focus on anaerobic fungi and the fungal cellulosome. This chapter also explores the cellulosome framework as a platform with synthetic biology applications beyond biomass degradation. The second chapter describes insights into cellulosome production and localization in anaerobic fungal cultures enabled by cellulosome-labeling bioimaging approaches. The third presents a robust method for cellulosome purification from the anaerobic fungus *Neocallimastix californiae* and details novel insights into cellulosome composition, structure, and biochemistry gained from the characterization of purified cellulosomes. The fourth chapter presents a novel, highly engineerable parts set for constructing synthetic enzyme complexes inspired by fungal cellulosomes. This chapter further introduces a combined computational-experimental approach to engineer stimuli-responsive assembly behavior into these assembly parts, towards constructing synthetic enzyme complexes with post-translationally controlled composition and activity. Chapter five describes the structural and biochemical characterization of two heterologously produced anaerobic fungal enzymes predicted to be important components of fungal cellulosomes. The final, sixth chapter summarizes the impact of this thesis work and discusses future work and remaining challenges towards translating fungal cellulosomes into useful biotechnologies.

1.4 Discovery and characterization of enzyme systems involved in anaerobic biomass degradation

In nature, the degradation and recycling of organic carbon is largely mediated by anaerobic microorganisms that work together to divide-and-conquer the difficult biocatalytic steps of hydrolysis [13–16]. To depolymerize biomass, these organisms secrete an array of carbohydrate-active enzymes (CAZymes). Industry has sourced a number of CAZymes used in bioprocessing from anaerobic microbes [17,18], and even metabolically engineered the anaerobic microbes themselves [19,20] for biotechnological applications. For example, in terrestrial systems, microbial communities derived from the rumen microbiome are highly efficient at extracting sugars from plants [17,21–23], and have provided a number of industrially sourced enzymes to liberate carbohydrates from plant waste. Even anoxic marine zones and sediments have provided key strains and enzymatic machinery to aid in carbon and nitrogen recycling [24–26].

A suite of biochemical activities are required to hydrolyze lignocellulose

Lignocellulosic biomass is a composite material with three major biopolymers, cellulose, hemicellulose, and lignin. Cellulose, the major component, is a polymer of β-1,4 linked D-glucose molecules. Cellulose chains aggregate via hydrogen bonding networks to form crystalline and amorphous macrostructures, forming fibrils that make up the core of plant cell walls [27]. Crystalline cellulose is highly ordered and dense, making it recalcitrant to degradation, while amorphous cellulose is more easily degraded. Enzymatic degradation of cellulose to glucose proceeds primarily via the coordinated action of four Glycoside Hydrolase (GH) family enzymes – endoglucanases, which cleave internal chain β-O-1,4 bonds, reducing end-acting exoglucanases, which processively cleave cellobiose units from the reducing end of a cellulose chain, cellobiohydrolases, which processively cleave cellobiose units from the non-reducing end of a cellulose chain, and β-glucosidase, which cleave cellobiose into glucose.

Lignin is typically the second most abundant polymer in lignocellulose behind cellulose and is a highly heterogeneous, aromatic biopolymer that primarily makes up the outer plant cell wall, providing structural rigidity and forming a shell around core cellulose fibrils. Aromatic subunits in lignin are connected via C-C and C-O bonds, making lignin highly recalcitrant to microbial degradation. The heterogeneous lignin backbone often contains branching hydroxycinnamic ester side chains, which can be terminal or form covalent crosslinks with hemicellulose [28]. Lignin is also proposed to form random covalent linkages to cellulose as well, further complicating the degradation of sugar polymers in lignocellulose. All characterized lignin-active enzymes employ free-radical generation to break lignin C-C and C-O bonds and are found in the Auxiliary Activity (AA) families of CAZymes [29].

Hemicellulose is the third most abundant component of lignocellulose. Like cellulose, hemicelluloses are made of sugars, but are much more heterogeneous than cellulose, containing polymers with monomeric units including xylose, glucose, mannose, galactose, arabinose, fucose, glucuronic acid, and galacturonic acid in various proportions, linked via different chemical bonds [30]. Xylans are the major component of hemicellulose, and most xylans possess a backbone of β-1,4 linked xylose residues that is decorated with other sugars or O-acetyl groups, though heteroxylans may contain other sugars in the linear backbone. An important, common xylan derivative abundant in many grasses are arabinoxylans, in which the xylan backbone is decorated with arabinose residues via α-1,2 or α-1,3 linkages. Arabinoxylans are frequently esterified with aromatic hydroxycinnamic acids present in lignin, cross-linking different arabinoxylan chains to each other and to lignin [31]. These crosslinks are thought to contribute significantly to the indigestibility of grasses by microorganisms. Other hemicellulose polymers include mannans, xyloglucans, and galactans, which form linear β-1,4 linked chains of mannose, xylose and glucose, and galactose sugars respectively, that are each functionalized with different sugar side chains [30]. Overall, because hemicelluloses contain a diversity of sugars of both α- and β- anomeric forms linked via 1,2-, 1,3-, 1,4-, and 1,6- ether bonds, in addition to acetyl ester and hydroxycinnamic ester modifications, a wide diversity of enzymes are required to completely hydrolyze hemicellulose. Indeed, at least 12 different CAZyme families perform distinct chemistries involved in hemicellulose degradation, and these are largely non-overlapping with enzyme families that participate in cellulose and lignin degradation (www.cazy.org).

Meta-omics of anaerobic communities identifies a wealth of CAZymes

Numerous meta-omics studies suggest the presence of a wide distribution of CAZymes in anaerobic biomass-degrading communities (Table 1) [32–34]. CAZymes, defined as enzymes that synthesize, degrade, or bind saccharides, are classified into six distinct classes - Glycoside Hydrolases (GH), Glycosyltransferases (GT), Polysaccharide lyases (PL), Carbohydrate Esterases (CE), Carbohydrate-Binding Modules (CBM), and Auxiliary Activities (AA), which includes ligninolytic enzymes and lytic polysaccharide monooxygenases. These classes contain multitudes of families and subfamilies, reflecting a huge diversity of activity and specificity among these enzymes. The authoritative database of annotated CAZyme sequences, CAZy (http://www.cazy.org) [35,36], has ballooned in size from ~340,000 protein entries in 2013 to over 1.4 million today, greatly surpassing growth in the number of experimentally characterized CAZymes.

Metagenomics and metatranscriptomics efforts have contributed to this immense growth in sequence information, but it should be noted that these analyses reflect the genomic potential and expressed genes in a sequenced sample, and do not necessarily reflect the production of active proteins. Metaproteomics analyses of biomass-degrading microbial communities to date only report on the order of tens of positively identified CAZymes (Table 1.1) [32,37], reflecting the challenges in confidently identifying proteins in complex samples using mass spectrometry proteomics. Furthermore, homology-based annotation of gene, transcript, or protein function using omics methods remains putative, and large-scale experimental characterization is needed to confidently link sequence and function. The success of this approach was recently demonstrated in the biochemical characterization of over 500 CAZymes in the CAZy database, which led to the discovery of new CAZyme families and assigned functions to 25 previously classified subfamilies [38].

Table 1.1. Numbers and classes of Carbohydrate Active Enzymes (CAZymes) present in different anaerobic microbial communities as surveyed from meta-omics data.

Environmental sample	**Omics techniques employed**	**Total CAZymes**	**# GH**	**# CE**	**# PL**	**# GT**	**# AA**	**# CBM**	**Ref**
Cow rumen	metatranscriptomics	12,237	10,209	1,404	624	-	-	-	34
Poplar adapted consortium	metagenomics	14,274	12,548	94	-	742	890	-	33
Corn stover-adapted consortium	metagenomics	1,798	691	270	41	348	56	392	32
Wheat straw compost liquid culture	metatranscriptomics and metaproteomics	88	26	16	1	2	7	36	39
Anaerobic digester consortia on filter paper	metaproteomics	16	13	-	-	-	-	3	37
Corn stover-adapted consortium	metaproteomics	56	30	4	-	-	1	21	32

The distribution of CAZymes among specific members of anaerobic, biomass-degrading consortia is far from uniform. Consortia that break down lignocellulose in the moose rumen, for example, form a specialized metabolic network of polymer degraders and sugar fermenters that cooperate to efficiently break down plant carbon [40]. Prolific CAZyme producers in these communities employ different strategies that empower their biomass-degrading ability. While most aerobic cellulolytic fungi and bacteria secrete free enzymes, many anaerobic bacteria and fungi produce large complexes called cellulosomes [41,42]. It has been hypothesized that the low energy nature of anaerobic fermentation drove the evolution of cellulosomes as a strategy to enhance the efficiency of cellulose hydrolysis and product uptake, but the nature of their evolution and their exception from aerobic environments remains unclear [6]. To date, only a handful of organisms are known to produce CAZymes with multiple catalytic domains [43,44], but it is extremely likely that many other interesting cellulases are yet to be found in anaerobic lignocellulolytic communities.

Cellulosomes enable synergy between CAZymes to accelerate biomass degradation

Cellulosomes were first discovered by Edward Bayer and colleagues in *Clostridium thermocellum*, a thermophilic bacterium found to produce cellulolytic enzyme complexes

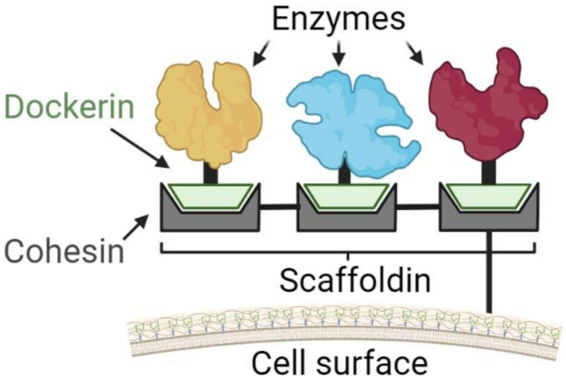


Figure 1.1. Diagram of cellulosome structure. Cellulosomes modularly tether enzymes via noncovalent interactions between enzyme-fused dockerin domains and cohesin domains repeated on a central, usually cell-bound scaffoldin protein. Created with Biorender.

~2.1 MDa in size containing 15 distinct subunits that could be isolated from cell culture [45]. Once the amino acid sequences of the bacterial cellulosome assembly domains (dockerin and cohesin) were uncovered [46], cellulosome assemblies were discovered in a number of other anaerobic bacteria [47] . Generally, a canonical bacterial cellulosome complex forms from the assembly of enzyme-fused dockerin domains on a central scaffoldin protein containing several (as many as nine) cohesin repeats which non-covalently bind the dockerins [48] (Figure 1.1). In *Clostridium thermocellum,* cellulosomes are primarily anchored to the cell surface via interaction between a cohesin-rich primary scaffoldin containing a carbohydrate binding module and a single dockerin domain and an anchoring scaffoldin bound to the cell surface either non-covalently via an S-layer homology motif or covalently via a sortase domain [49]. However, recent work found that *C. thermocellum* and other cellulosome-producing bacteria also secrete a cell-free cellulosome system [50,51]. Consistent with an emerging paradigm of extracellular metabolism among biomass-degrading bacteria, evidence also exists showing cellulosomes localized to the surfaces of secreted membrane vesicles [52].

Bacterial cellulosomes possess many enzyme activities, including cellulase, xylanase, pectinase, mannanase, and xyloglucanase activities. All known cellulosome-producing bacteria produce a GH48 exoglucanase as the major cellulosome component which is complemented by a repertoire of GH9 endoglucanases. Additional major enzyme components include GH5, GH10, GH11, and GH43, an array of enzymes capable of hydrolyzing cellulose and hemicellulose. These different enzyme families may synergize when part of a cellulosome through several mechanisms: 1) endoglucanases increase the number of chain ends for attack by exoglucanases, 2) hemicellulases improve access to the cellulose core for attack by cellulases, or 3) relief of product inhibition (Figure 1.2).

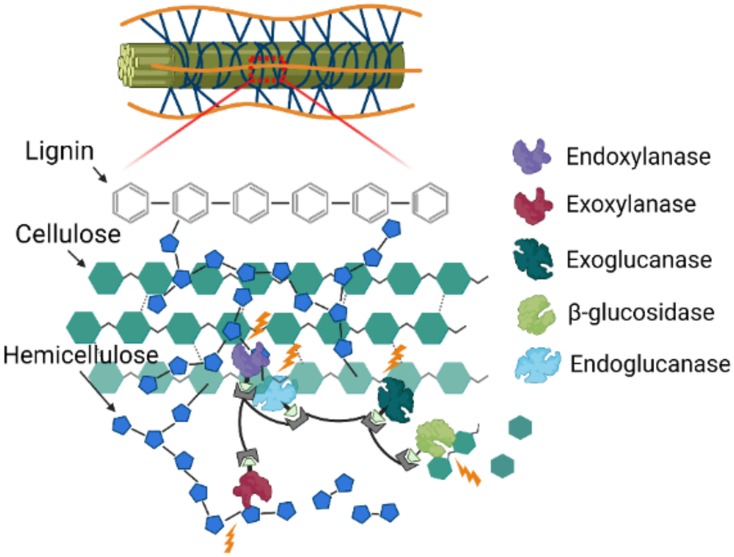


Figure 1.2. Enzyme synergy during lignocellulose hydrolysis by cellulosomes. Cellulosomes spatially organize enzymes that break different chemical bonds within biomass (indicated by lighting bolts). Endoglucanases and endoxylanases internally cleave hemicellulose and cellulose chains, creating more end sites for cleavage by exocellulases and exoxylanases. Β-glucosidases convert disaccharide, which inhibits cellulase activity, into glucose. Created with Biorender.

Designer cellulosomes using bacterial dockerins and cohesins have served as an informative platform for investigating how various lignocellulolytic activities synergize. The core synergism in bacterial cellulosomes appears to occur between GH48 exocellulases and GH9 endoglucanases [53]. While the trifunctional complexes of recombinant *Clostridium cellulovorans* enzymes examined in [53] were generally two to six-fold less active than native cellulosomes from *C. cellulovorans*, the GH48 and GH9 subunits contributed ~75% of the overall enzymatic activity regardless of the third subunit when tested against crystalline cellulose. On the lignocellulosic substrate wheat straw, the authors show that complexation of GH48 and GH9 with a multifunctional GH10, feruloyl esterase enzyme enhances hydrolysis four-fold over a mixture of the free enzymes. This result highlights the benefit of colocalizing enzymes that hydrolyze hemicellulose and hemicellulose-lignin linkages in addition to those that act on cellulose.

Generally, it is assumed that colocalization of hemicellulases is less beneficial for hemicellulose degradation compared to cellulose given hemicellulose's amorphous structure. However, evidence shows incorporating enzymes with endoxylanase and xylobiase activities into synthetic cellulosomes increases both the rate and extent of wheat straw hydrolysis compared to free enzymes [54].

The work highlighted here suggests that the core enzyme families that should be represented in cellulosomes for optimal lignocellulose hydrolysis are exocellulase, endocellulase, β-glucosidase, β-xylosidase, endoxylanase, and feruloyl/coumaroyl esterase. Indeed, the major components of the native *C. clariflavum* cellulosome are GH48 and GH9 proteins, regardless of growth substrate; GH5, GH10/11, and CE families are also represented among the 20 most abundant proteins measured by quantitative proteomics [51]. In *C. thermocellum*, GH48, GH11, GH9, GH5, and GH8 enzymes, representing the key cellulase and xylanase activites, were also among the top 10 cellulosome proteins when the organism was grown on a suite of substrates ranging in complexity from cellobiose to switchgrass [55].

Cellulosome complexes from anaerobic fungi were first described in 1992 [56] in *Neocallimastix frontalis*, in which a culture supernatant fraction containing multi-subunit, highly cellulolytic enzyme complexes ~750-1000 kDa in size was characterized. Cellulosome complexes were also observed in the supernatants of a *Piromyces* species [57], suggesting cellulosome production was a general strategy evolved by anaerobic fungi to degrade lignocellulose. This hypothesis was largely confirmed with the annotation and demonstration of fungal dockerin domains as required for cellulosome binding [58,59]; many fungal dockerin-containing genes are annotated in all published anaerobic fungal genomes sequenced to date. Fungal dockerin has neither sequence nor structural similarity to bacterial dockerin, suggesting cellulosome formation evolved separately in the two kingdoms [58]. Fungal dockerins are fused to catalytic domains on either the N- or C- terminus and are most frequently encoded as tandem repeats of two domains (termed a double dockerin), though single and triple dockerin-fused enzymes are also commonly observed [7]. Given the shared feature of modular interacting domains mediating enzyme assembly, fungal cellulosomes are thought to mirror bacterial ones in overall structure (Figure 1.1b). However, the fact that a true, dockerin-binding cohesin has yet to be discovered in anaerobic fungi, complicates investigations into fungal cellulosome enzyme synergy and additionally, has led some to question whether the bacterial model for fungal cellulosomes is accurate.

Enzyme activities biochemically verified to be present in fungal cellulosomes include endoglucanase, exoglucanase, cellobiohydrolase, β-glucosidase, feruloyl/coumaroyl esterase, xylanase, mannanase, and acetyl xylan esterase activities [57,60,61]. Direct comparison of the cellulosome and free enzymes isolated from the culture supernatant of *Piromyces* sp. E2 showed that, despite both possessing all three cellulolytic activities required to degrade crystalline cellulose, cellulosomes achieved complete conversion of 2% (w/v) Avicel to glucose while the free enzymes accomplished only 25% conversion over 12 days [62]. A similar study of the cellulosome and free enzyme systems of *Neocallimastix frontalis* observed the same behavior for cotton solubilization [56]. While the cellulosome complex accounted for the majority of lignocellulolytic activity in gut fungal supernatants, in both studies the presence of both cellulosomes and free enzymes synergized substrate degradation via mechanisms that are not understood.

Equivalent investigations into the hemicellulolytic activity of fungal cellulosomes vs free enzymes in gut fungal supernatants are lacking, and whether enzyme colocalization benefits cellulosomes to the same degree remains an interesting question since hemicellulose lacks the highly ordered crystalline structure of cellulose in plants. While most work on hemicellulolytic synergy has focused on bacterial or aerobic fungal enzymes, a notable exception studying recombinant dockerin-containing feruloyl esterase, EstA, from *Piromyces equi* found that supplementing *P. equi* EstA with xylanase from *Trichoderma viride* boosted the release of ferulic acid from destarched wheat bran over 100-fold, illustrating how esterase and xylanase activities synergize in degrading a lignocellulosic substrate [60].

1.5 An overview of anaerobic fungal biology

Anaerobic fungi (*Neocallimastigomycota*) found in the guts of herbivores are some of the most prolific biomass degraders found in nature. Despite only accounting for an estimated 8% of the gut microflora by mass, anaerobic fungi account for over 49% of crude biomass saccharification that occurs in the animal gut [63–65]. Anaerobic fungi were first described by Colin Orpin in 1975 [66], isolating the species *Neocallimastix frontalis* from the rumen of sheep.

Neocallimastigomycetes are close relatives in fungal phylogeny to *Chytridiomycota* and process through a similar life cycle that starts with a motile, flagellated single-celled zoospore [67] (Figure 1.3). A maturing zoospore sheds its flagella and begins to grow rhizoids, filamentous structures similar to hyphae produced by filamentous fungi such as *Aspergillus* or *Trichoderma*, transitioning to a new life stage termed a thallus. Progeny zoospores are then formed within the head-like structure (also termed a thallus), forming a multi-celled body called a zoosporangium which eventually bursts to release nascent zoospores and restart the cycle.

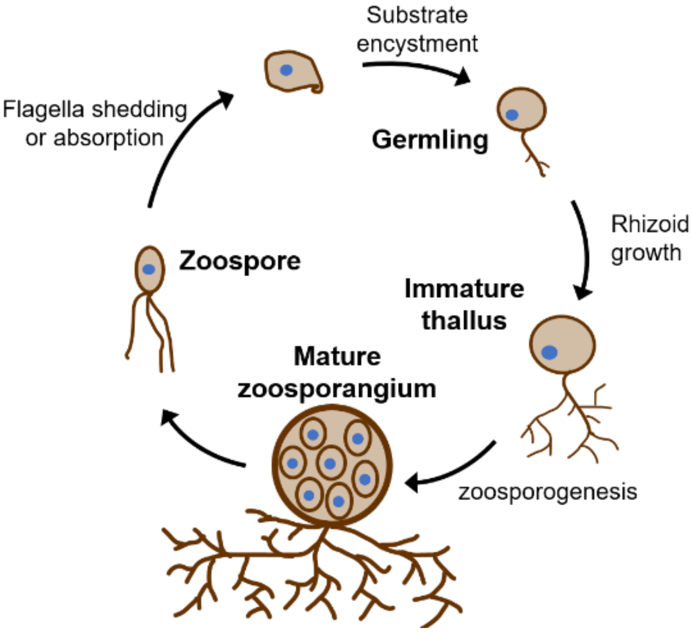


Figure 1.3. Anaerobic fungi replicate through a multi-step life cycle accompanied by dramatic morphological change. Motile, flagellated zoospores follow chemiosmotic gradients to a carbon source and, on insoluble substrates, encyst themselves through the extension of root-like rhizoids into the substrate. The main zoospore body simultaneously matures into a thallus that, in the reproductive stage of growth, hosts newly formed zoospores that eventually break out of the zoosporangium to repeat the growth cycle. This schematic represents reproduction of a monocentric anaerobic fungus with multiflagellar zoospores.

While this life cycle is believed to describe all genera within *Neocallimastigomycota*, different genera do display morphological differences. In contrast to all other genera, *Caecomyces* cells, isolated from the hindgut of a horse, do not produce root-like rhizoids [68]. Whether zoospores are monoflagellated or polyflagellated is also an observed difference among genera, as is whether thalli are monocentric (one nucleus per thallus) or polycentric [69].

Regardless of genera, all published *Neocallimastigomycota* genomes contain a large and diverse array of CAZymes, many of which contain dockerin domains, suggesting all known species of *Neocallimastigomycota* produce cellulosomes.

1.6 Cellulosomes as a biotechnology platform

Enzymes drive the complex network of chemical transformations that mediates nearly all critical life processes. To enhance and direct the activity of enzyme systems, nature evolved strategies to spatiotemporally organize enzymes into multi-functional, ordered complexes (reviewed in [8,9,70]). Exemplifying its importance, enzyme complexation as a strategy for efficient multi-step chemical transformation has evolved in unique biological contexts, including natural product biosynthesis [71], biomass saccharification [72], and signal transduction [73,74]. As it does for enzyme systems that naturally incorporate into complexes, enzyme colocalization can dramatically enhance product fluxes when heterologous enzyme sets are artificially colocalized, making it an attractive strategy for improving industrial biocatalyst performance [9,75].

Cellulosome's modular assembly architecture is a very attractive framework for engineering enzyme systems. Indeed, bacterial dockerins and chimeric scaffoldins sourced from different cellulosome-producing bacteria have been exploited to enhance reaction rates in many different biocatalytic cascades, simply by fusing a dockerin domain to each pathway enzyme and co-producing a scaffoldin protein onto which subunits assemble [76,77]. Endowing the platform industrial yeast *S. cerevisiae* with bacterial cellulosome machinery also produced an efficient consolidated bioprocess for producing ethanol from crystalline cellulose [78]. However, not all heterologous enzyme systems are compatible with the bacterial cellulosome framework. Naturally, bacterial dockerins mainly exist as C-terminal fusions to enzymes, and producing synthetic chimeras with dockerin at the N-terminus has demonstrated limited success [11].

A mechanistic understanding of bacterial cellulosome assembly has enabled translation and optimization of this framework into technologies with diverse applications. Surprisingly, enzyme arrangement within natural bacterial cellulosomes is somewhat random, as dockerin-cohesin interactions are non-specific within the same species [72]; some degree of arrangement specificity is attributed to interactions between subunits [79]. Local enzyme arrangement is, however, critical for proper function of heterologous enzyme pathways in a synthetic complex [75,80], and effects of enzyme arrangement can be interrogated explicitly in complexes based on bacterial cellulosome parts.

The unknown identity of fungal cohesin has severely limited biotechnological development of fungal cellulosomes. A number of fungal cohesin candidates have been proposed from dockerin affinity chromatography experiments [81], and the most-comprehensive evidence suggests fungal cohesin is within a class of large, repeat-rich proteins (annotated as scaffoldins) conserved across anaerobic fungal genomes; a fragment of one of these proteins demonstrated dockerin binding with a $K_d = 1\ \mu M$ [82]. A specific sequence and structure for the cohesin domain within this fragment remains elusive.

In contrast, fungal dockerins are well characterized with a known structure and sequence [59] and are ideal domains for synthetic protein assembly provided a suitable binding partner is discovered or engineered. Across anaerobic fungal genomes, dockerin sequences are fused to the N- or C-termini of 99 different Pfam protein families [83–85], and numerous fusions of fungal dockerins to proteins from other organisms facilitate incorporation of the functional protein into the cellulosome [86]. Furthermore, the dockerin fold accommodates high sequence diversity, implying high functional engineerability without disrupting binding activity [83–85].

II. Anaerobic fungi localize cellulosomes to the cell-biomass interface to drive plant matter hydrolysis

Parts of this chapter are adapted from Stephen P. Lillington, et al, *mBio* 2021 [1]. Reprinted under a CC-BY Creative Commons license.

2.1 Introduction

Anaerobic fungi (phylum *Neocallimastigomycota*) are commonly found in the digestive tracks of large, herbivorous animals where they play an important role in colonizing and degrading ingested plant biomass [69,87]. Despite only accounting for ~8% by mass of the gut microflora [63], reducing the ruminal anaerobic fungi population has been shown to cause a 49-70% decrease in unpretreated biomass consumption by sheep and cattle compared to those with a natural abundance of fungi in the rumen [64,65]. Furthermore, anaerobic fungi have been shown to preferentially degrade the more recalcitrant, lignin-rich plant matter that cellulolytic rumen bacteria cannot catabolize, making them attractive potential hosts for converting unpretreated waste biomass into high-value bioproducts [87,88]. Since the first description of these organisms in 1975 by Colin Orpin [66], more than 28 species of anaerobic fungi have been isolated and characterized to better understand both their ecological role in the rumen microbiome and their biomass degrading machinery [89,90].

Genomic and transcriptomic sequencing have shown that anaerobic fungi harbor a wealth of carbohydrate active enzymes (CAZymes) that enable their deconstruction of crude lignocellulose. Whole genome sequencing of 6 species of anaerobic fungi [82,91–93] revealed that they encode on average over four-fold more CAZymes compared to *Trichoderma reesei* and *Aspergillus niger*, the sources of the most popular cellulolytic cocktails in industry (www.mycocosm.jgi.gov, [94]). The functional diversity of the encoded gut fungal CAZymes is similarly astounding; represented among the published genomes of *Neocallimastigomycota* are 43 Glycoside Hydrolase (GH) families, 28 Glycosyl Transferase (GT) families, 9 Carbohydrate Esterase (CE) families, 5 Polysaccharide Lyase (PL) families, and 19 Carbohydrate Binding Module (CBM) families (www.mycocosm.jgi.gov).While their genomic potential is impressive, more biochemical data and functional knowledge of how this diverse group of hydrolytic enzymes function *in vivo* are necessary before the degradative machinery of anaerobic fungi or the organisms themselves can be developed into useful platforms for the conversion of waste biomass.

Like several species of anaerobic bacteria, anaerobic fungi incorporate many of these CAZymes into multi-enzyme complexes called cellulosomes, which colocalize lignocellulolytic enzymes of complementary function to greatly enhance degradative activity [47]. Fungal cellulosomes are thought to mimic bacterial ones in their general structure, in which modular dockerin domains attached to catalytic proteins non-covalently bind repeated cohesin domains on a central, membrane-anchored scaffoldin. These interacting parts have been well characterized in bacterial cellulosomes; the molecular details of their interaction are well described, as is their localization on the cell surface *in vivo* [46,95]. As a result of its characterization, the bacterial cellulosome has served as a template for designing synthetic protein complexes with wide-ranging applications in nanobiotechnology, recently reviewed in [47].

In contrast, while the fungal dockerin domain has a known sequence and structure divergent from the bacterial dockerin [59], the identity of a conserved cohesin domain in anaerobic fungi remains elusive, though a conserved group of large (>500kDa), repeat-rich scaffoldins with no sequence homology to any bacterial cellulosome component was recently shown to bind recombinant fungal dockerin with a K_D of 994 nM [82]. No *in vivo* colocalization studies have cemented these scaffoldin's role in native fungal cellulosomes yet, but comparative genomic analyses of fungal cellulosome composition and domain architecture suggest fungal cellulosome parts may serve as better templates than bacterial ones for certain synthetic systems. An observed roadblock to constructing designer cellulosomes using bacterial components is failure to construct stable, functional enzyme-dockerin chimeras, particularly for enzymes not of bacterial origin or for N-terminal dockerin fusions [11]. In contrast to bacterial dockerins, fungal dockerin domains are observed as N- and C-terminal fusions to enzymes and the diversity of dockerin-containing proteins in anaerobic fungi significantly exceeds that of bacteria [82], suggesting more general compatibility of fungal cellulosome components in the construction of synthetic complexes. Where cellulosomes localize in native systems and how these structures attach to cells and biomass are key unresolved details to inform the successful design of synthetic fungal cellulosomes. No previous work has investigated the *in vivo* localization of the newly discovered fungal scaffoldins, and only one study observed dockerin localization on the cell surface of *Orpinomyces sp.* PC-2 using electron microscopy and an immunogold-labeled anti-dockerin antibody [96].

Deriving insight into the cellulosome's functional role by observing its spatial localization patterns is complicated by the complex life cycle of anaerobic gut fungi, since the composition and role of cellulosome-associated proteins [82] likely change during fungal life cycle progression. As members of *Chytridiomycota*, anaerobic fungi proliferate through a life cycle in which a motile zoospore encysts on plant biomass, growing root-like rhizoids and maturing into a zoospore-filled sporangium that releases more zoospores (Figure 1.3). The drastic morphological changes seen during life cycle progression raises the question of which life stage is predominantly responsible for rapid lignocellulose hydrolysis. Furthermore, as potential ingredients in lignocellulolytic enzyme cocktails, it is important to understand how production of cellulosomes is regulated during life cycle progression. It has generally been assumed the rhizoid-bearing, maturing cells are responsible for both robust cellulosome production and rapid lignocellulose hydrolysis, since rhizoids bear resemblance to hyphae, the primary sites of enzyme secretion and biomass hydrolysis in filamentous fungi like *Aspergillus niger* [97]. Experimental evidence confirming the functional importance of mature cells or their rhizoids to biomass hydrolysis is still lacking.

In this work, we developed imaging probes unique to fungal cellulosomes that label key cellulosome protein domains in the anaerobic fungus *Piromyces finnis*. Antibodies raised against the fungal scaffoldin, dockerin, and GH48 domain, the most abundant dockerin-fused enzyme in the *P. finnis* cellulosome, enabled localization of these domains in samples from two different genera within *Neocallimastigomycota.* Using these tools, we show that rhizoids are the dominant location of lignocellulose hydrolysis in mature cells, displaying high coverage of both GH48 and dockerin domains localized to the surface, but only when anaerobic fungi are challenged with insoluble, complex carbon sources such as Whatman cellulose paper. In contrast, zoospores display both GH48 and dockerin-containing proteins on their main body regardless of growth substrate complexity, insinuating differential regulation of cellulosomal CAZyme production throughout the fungal life cycle. Our results suggest zoospores accompany rhizoid-bearing cells as important biomass degraders in the rumen environment. They also imply that comparative analysis of zoospore and mature cell gene expression data may yield genetic targets to engineer anaerobic fungi for overproduction of CAZymes and cellulosomes throughout the fungal life cycle under all growth conditions [98]. These first insights into cellulosome spatial localization and life cycle-dependent regulation of cellulosome production will benefit both future laboratory study seeking to engineer fungal cellulosomes and future development of anaerobic fungi into platforms for waste biomass upcycling or hydrolytic enzyme production.

2.2 Results

Anaerobic fungi use rhizoids to penetrate and disrupt plant biomass

A key characteristic of rhizoid-forming anaerobic fungi is the highly branched morphology of mature fungal cells, which is unique among members of the rumen microbiota. In the rumen, root-like rhizoids are known to facilitate biomass colonization by anaerobic fungi and are hypothesized to play a major role in lignocellulose hydrolysis, both by mechanically disrupting ingested biomass [5,88,99] and by attaching fungal cellulosomes, the primary source of cellulolytic power [100–103]. To our knowledge, neither of these assumptions has been explicitly verified for any anaerobic fungal isolate. Prior works report size estimates for fungal cellulosomes based on size exclusion chromatography and electron microscopy ranging from 700 kDa to tens of MDa [100,101], suggesting these structures might be visualized by Helium Ion Microscopy (HIM). We sought to determine whether fungal rhizoids host fungal cellulosomes using HIM analysis of anaerobic fungi growing on lignocellulosic substrates.

Lower magnification HIM micrographs of *Piromyces finnis* grown on dried switchgrass biomass demonstrate the rhizoids of mature cells are the major interface with grass particles

(Fig. 2.2A-B). Rhizoidal morphology greatly increases the interfacial surface area between cells and their carbon source, likely an important factor in enhancing lignocellulose hydrolysis. Though HIM is a surface microscopy technique and cannot resolve internal structures, the apparent growth of rhizoids into the grass particles in Fig. 2.2A-B suggests these structures penetrate the colonized substrate to access trapped carbon. This phenomenon is well documented in samples taken from a live animal rumen, showing heavy internal rhizoid colonization of damaged, living plant tissues [88,104], but is not confirmed to

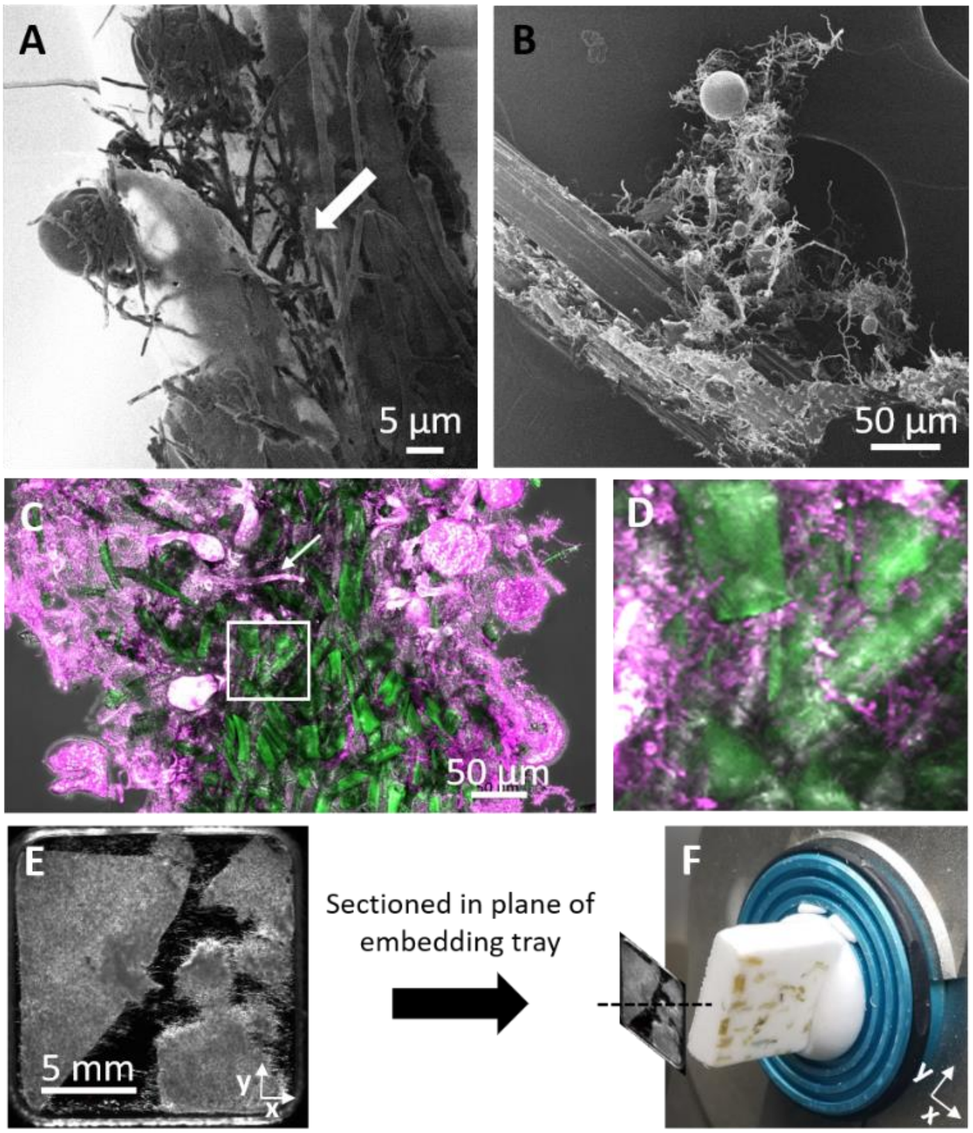


Figure 2.2. Rhizoids mediate substrate attachment and penetration by mature *Piromyces* cells. Helium ion micrographs show that rhizoids of mature *Piromyces finnis* cells mediate contact with the lignocellulosic substrate and suggest a penetrative growth phenotype, indicated by the white arrow that aids in fungal biomass degradation (A-B). Substrate penetration by rhizoids was confirmed by cryo-sectioning a formaldehyde-fixed culture of *Piromyces finnis* grown on cellulose paper, showing clear growth into the substrate interior, including one rhizoid, identified by the arrow, penetrating over 100 μm (C). A magnified portion of the maximum intensity projection, the white box in (C), highlights the extent to which the rhizoid network colonizes the substrate interior (D). Sections were prepared by cutting in the plane of the embedded sample as shown in (E-F). Confocal micrographs from 33 2 μm Z-steps were obtained with a LD C-Apochromat 40x/1.1 W Korr M27 objective and collapsed to a maximum intensity projection using Zen software (Zeiss). Stains: FM 1-43 membrane stain (Pink) and SYBR Gold (Green).

occur during anaerobic fungal growth on dried (ligno)cellulosic substrates, which are the recommended biomass source for biorefinery concepts [105]. To determine if biomass colonization patterns are similar on dried substrates, we cryo-sectioned and imaged by confocal microscopy formaldehyde-fixed samples of an anaerobic fungus growing on Whatman cellulose paper (Figure 2.2E-F). Though sacrificing in resolution using confocal instead of helium ion microscopy, the sectioned samples enabled unequivocable observation of rhizoid penetration into the substrate interior, with one rhizoid shown growing over 100 μm into the substrate (Figure 2.2C-D).

Without coincident immunolabeling, it was difficult to identify cellulosome structures in the HIM micrographs directly, though we frequently observed bumpy regions in what appeared to be highly branched rhizoid systems with globular features 10-100 nm in diameter (Figure 2.3). It is important to note that biomass-only negative controls possessed somewhat similar fibrillar networks of a similar length scale, and a generally wispy microstructure that obfuscated which structural features arose from cells versus the substrate. As such, it is possible Figure 2.3 depicts enzyme-degraded reed canary grass, with

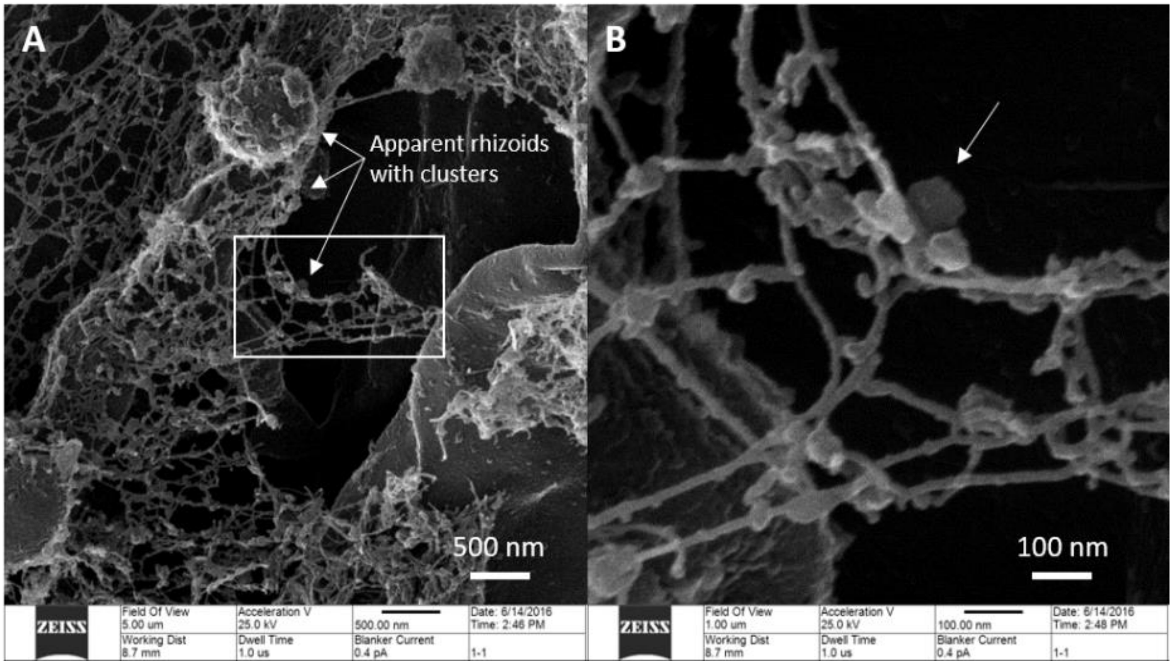


Figure 2.3. Potential cellulosome structures from *Piromyces finnis* visualized by Helium Ion Microscopy (HeIM). Images were obtained from fixed, dehydrated samples of *P. finnis* grown on reed canary grass as described in the Materials and Methods. Globular structures with diameters in the 10 nm to 100 nm range, consistent with MDa-sized protein complexes, are apparent on the surface of filamentous structures ~10 nm in diameter. Filaments this small could possibly originate from the reed canary grass substrate, in which case the globular structures may be bound, cell-free cellulosome complexes.

the highly branched fibrillar network comprising reed canary grass filaments and the globular structures possibly representing bound proteins or cellulosomes. However, HIM micrographs in which rhizoids were clearly emerging from sporangia definitively highlighted rhizoids' "rough" surfaces, possibly characteristic of surface-displayed cellulolytic enzymes and/or cellulosomes that may localize to these structures (Figure 2.4). Such structural features could have been artifacts of chemical fixation during sample preparation, and as such, these data were used solely to generate hypotheses about the localization of fungal cellulosomes *in vivo*.

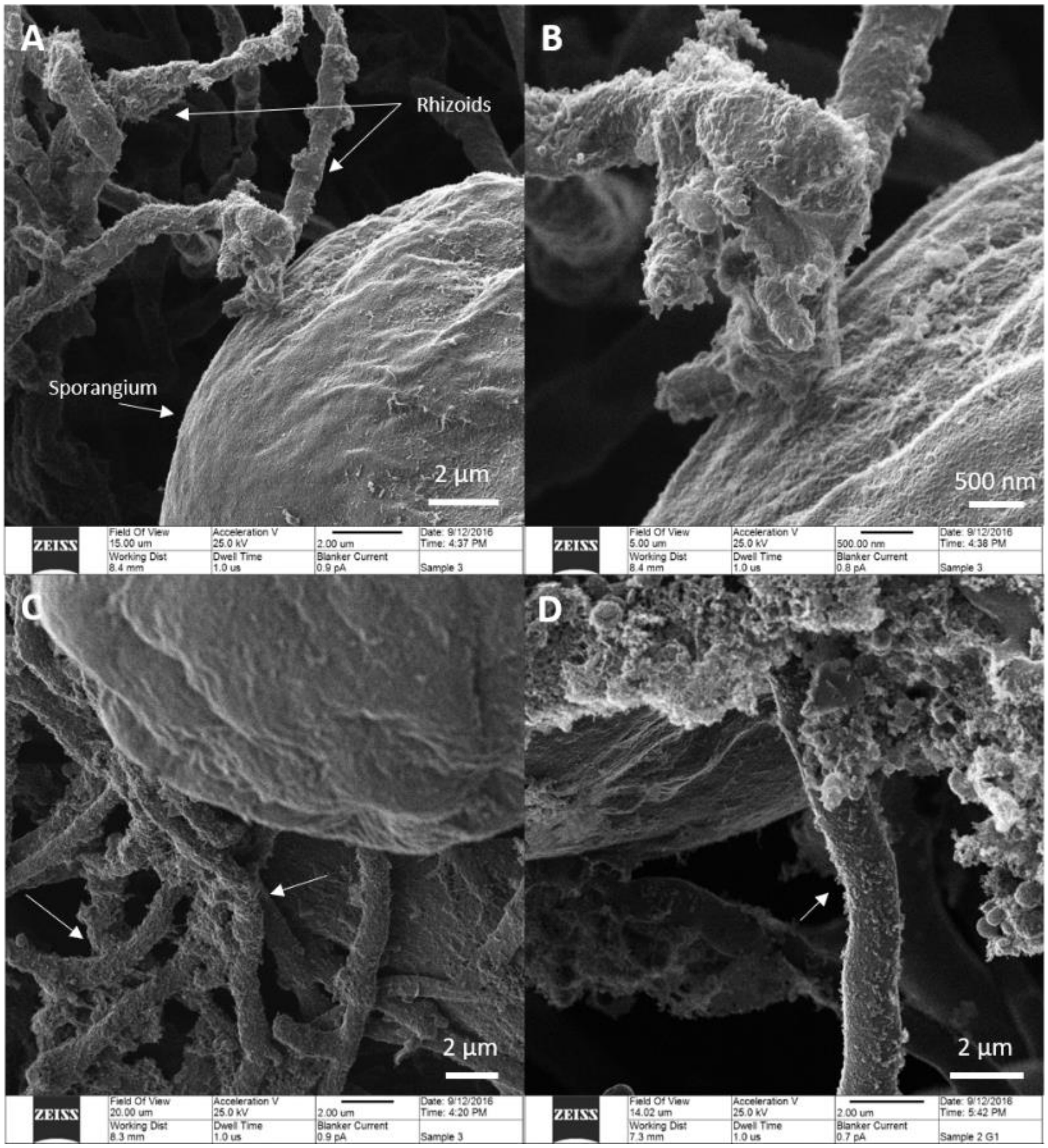


Figure 2.4. Larger rhizoids of *Piromyces finnis* and *Neocallimastix californiae* cells have rough surfaces covered with globular structures that may be proteins or protein complexes. Panels A-C show micrographs from *P. finnis* cultured on reed canary grass, in which B is a magnified close-up of part of the rhizoid annotated in A).

Rhizoids of mature fungal cells display cellulosomes and other lignocellulolytic proteins

To determine whether the putative rhizoid-localized proteins were cellulosomes, antibodies were raised against three recombinantly-produced proteins from *Piromyces finnis*

– a double dockerin domain which is the prevailing form in cellulosome-associated proteins [82], a fragment of the dockerin's putative binding partner ScaA, and a CAZyme of the GH48 family, the dominant enzyme in the *Piromyces* cellulosome [42,106]. Western blots show the anti-dockerin and anti-GH48 antibodies bind to multiple proteins from the *Piromyces finnis* cellulosome, several of which were verified to contain the antibodies' target domain by mass spectrometry (MS) proteomics [82] (Figure 2.5A-B). The anti-ScaA antibody bound specifically to the recombinant fragment used for immunization (Figure 2.5C) but did not appear to bind any protein from the *P. finnis* cellulosome in a Western blot. One possible explanation is that, though MS proteomics confirmed the presence of ScaA in the *P. finnis* cellulosome [82], the antigen's abundance may have been below the detection limit of standard Western blot (data not shown). As a secreted protein, native ScaA is also expected to be post-translationally modified in ways the recombinant protein was not, which may provide an alternative explanation for this observation if the anti-ScaA epitope is post-translationally modified. An additional possibility is that the antibody binding site may not be accessible in an *in vivo* context, either buried in the core of the full length ScaA or blocked by the binding of other peptides.

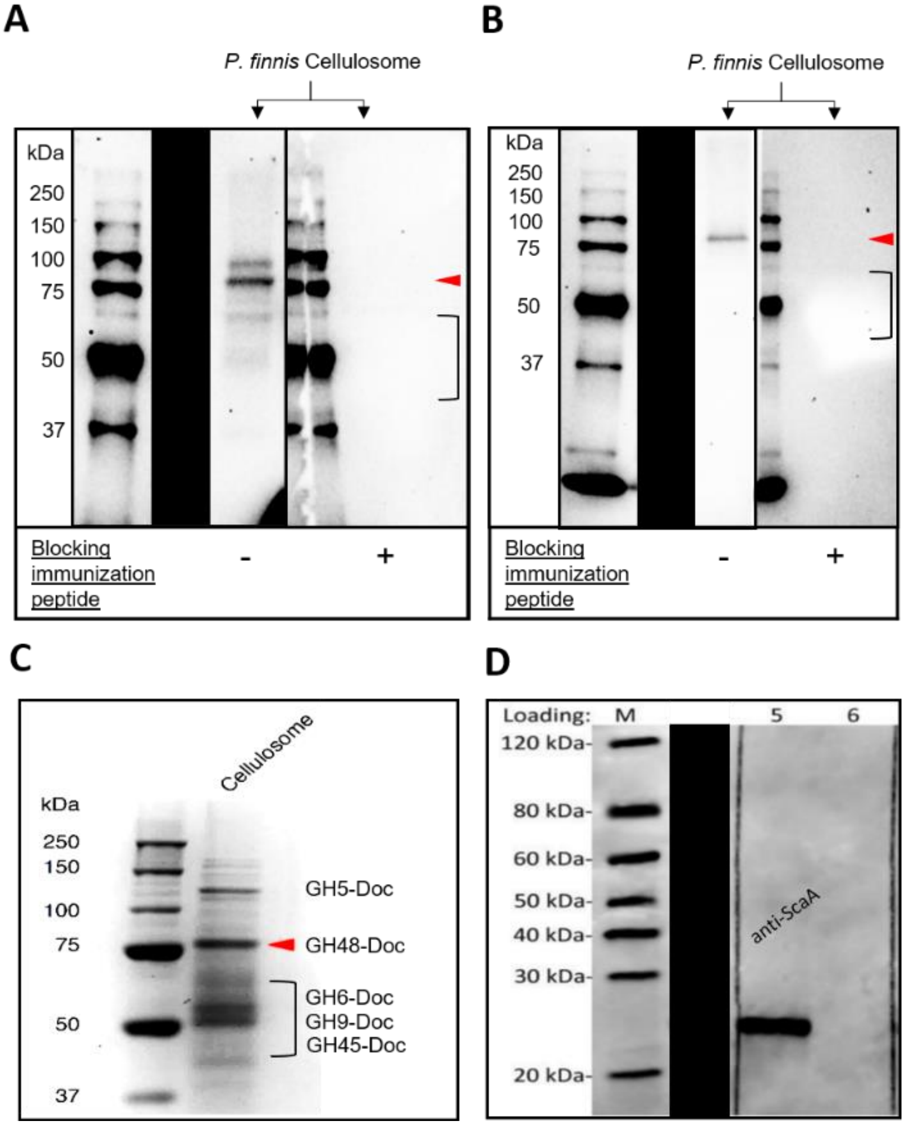


Figure 2.5. Antibodies generated against recombinant cellulosome protein fragments bind their expected targets in the native *P. finnis* cellulosome. (A) Our anti-dockerin antibody binds multiple proteins from the *P. finnis* cellulosome. Preincubating the antibody with the immunizing peptide ablates signal, demonstrating high antibody-epitope specificity. **(**B) Our anti-GH48 antibody binds a single target in *P. finnis* cellulosome. Preincubating the antibody with the immunizing peptide ablates signal, demonstrating high antibody-epitope specificity. (C) A Coomassie Blue-stained SDS-PAGE gel of the *P. finnis* cellulosome. Putative band identities annotated by arrows or brackets in (A) and (B) are assigned from qualitative MS proteomics data provided in Haitjema et al, 2017 (10). "Doc" indicates the protein contains at least one fused dockerin domain. (D) A Western blot of mouse monoclonal antibody against a recombinant fragment of the *P. finnis* ScaA scaffoldin protein shows specific binding to the scaffoldin fragment. A black bar separates non-adjacent lanes from the same SDS-PAGE gel for clarity. No binding to any protein in the *P. finnis* cellulosome in (C) was observed.

We characterized dockerin and GH48 localization patterns by analyzing formaldehyde-fixed cultures of *Piromyces finnis* after 72-96 hours of growth on Whatman cellulose paper. Whatman paper was chosen as a substrate both because it is known to induce robust production of CAZymes and dockerin-containing proteins in anaerobic fungi [42] and because, in addition to fungal cell walls, it is also stained by Calcofluor white, facilitating visualization. Samples were stained with Calcofluor white and one or multiple of the three antibodies before visualization under a Zeiss LSM 710 confocal microscope equipped with a LD C-Apochromat 40x/1.1 W objective. As shown in Figure 2.6, both anti-dockerin and anti-GH48 signals localize intensely to the surface of mature fungal cell rhizoids during growth on cellulose paper, consistent with the HIM images depicting protein-like structures on rhizoid surfaces (Figures 2.2-2.3). Surface-localized dockerin and GH48 appear on all parts of a cell's rhizoid system, both at the oldest growth near the thallus or sporangial head and at the newer, more highly branched growth (Figure 2.6C-F). The absence of signal in samples stained only with secondary antibodies provides strong evidence these observations are not an artifact of our visualization method (Figure 2.6A-B). Importantly, no anti-dockerin or anti-

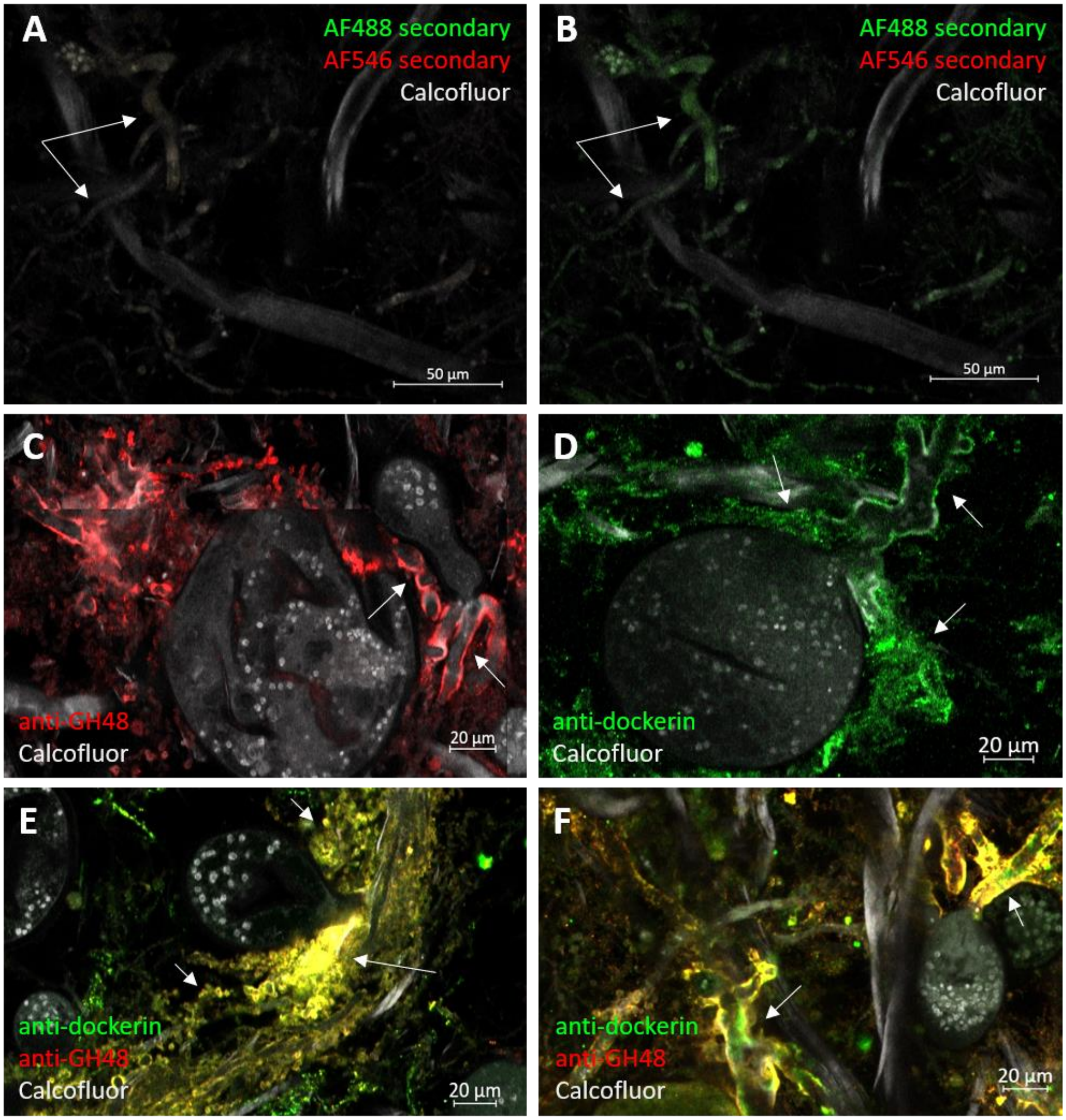


Figure 2.6. Mature *Piromyces finnis* cells localize cellulose-degrading machinery to rhizoids when growing on complex substrates. Immunofluorescence staining of fixed *P. finnis* cells grown on Whatman filter paper shows localization of anti-GH48 (red) and anti-dockerin (green) signal to rhizoids indicated with white arrows (C-F). A negative control stained only with Calcofluor white and secondary antibodies (donkey anti-rabbit IgG AlexaFluor 488, donkey anti-rabbit IgG AlexaFluor 546, and goat anti-mouse IgG AlexaFluor 647) shows low background signal (A-B). Anti-dockerin labeling was lower than anti-GH48 across samples, and the different image display settings used to produce panel C) and D) are reproduced in A) and B) respectively. E) and F) show maximum intensity projections from five 1 μm confocal image stacks that illustrate the abundant colocalization of anti-dockerin and anti-GH48 to cell rhizoids, evidenced by apparent yellow staining caused by co-emission of red and green light. Each panel, except A) and B) which are the same micrograph with different image processing, is representative of three technical replicate samples from a single culture tube. The negative controls are representative of five technical replicates. The same display settings and gamma parameters in Zen (Zeiss Microscopy) were used to generate panels A,C,E,F. Different but consistent display settings and gamma parameters were used to generate panels B and D. Confocal micrographs were obtained with a LD C-Apochromat 40x/1.1 W Korr M27 objective. Antibodies and stains used: Calcofluor white (gray); GH48 – rabbit anti-GH48 (primary), donkey anti-rabbit AlexaFluor 546 (secondary, red); Dockerin – rabbit anti-dockerin (primary), donkey anti-rabbit AlexaFluor 488 (secondary, green).

GH48 signal localized to thallus or sporangial heads, suggesting rhizoids are the primary site for lignocellulose hydrolysis. The anti-ScaA antibody did not yield specific positive staining comparable in appearance with that of the dockerin or GH48 to any part of the cells. Much of the anti-ScaA signal appeared non-specifically bound to cellulose paper fibrils, suggesting the anti-ScaA antibody likely does not bind its intended target.

Growth conditions control production of key cellulosome proteins across the life stages of anaerobic fungi

The HIM and initial immunofluorescence micrographs suggested that fungal rhizoids primarily serve as the sites for cellulosome localization. However, further examination of immunofluorescence images from the *Piromyces* filter paper cultures showed an abundance of small spherical bodies consistent in size with zoospores (~5 µm in diameter) that also displayed high dockerin and GH48 detection signal well beyond that seen in the negative control (Figure 2.7A-C).

The apparent presence of surface-displayed cellulosomes by both zoospores and mature fungal thalli is surprising given the stark difference in cell morphology and previously described roles of the two life stages, in which rhizoid-bearing thalli are the chief biomass degraders, while motile zoospores search for new carbon sources to colonize [107]. These results also imply an interesting connection between cellulosome localization and

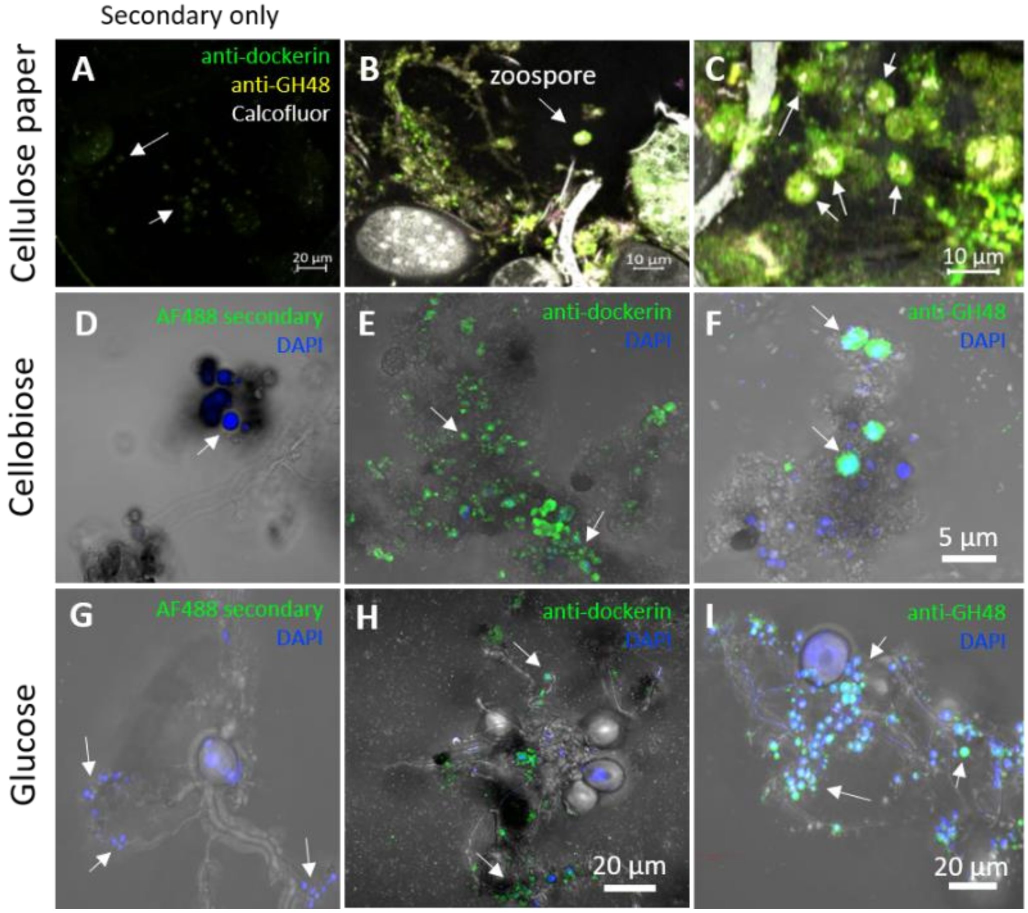


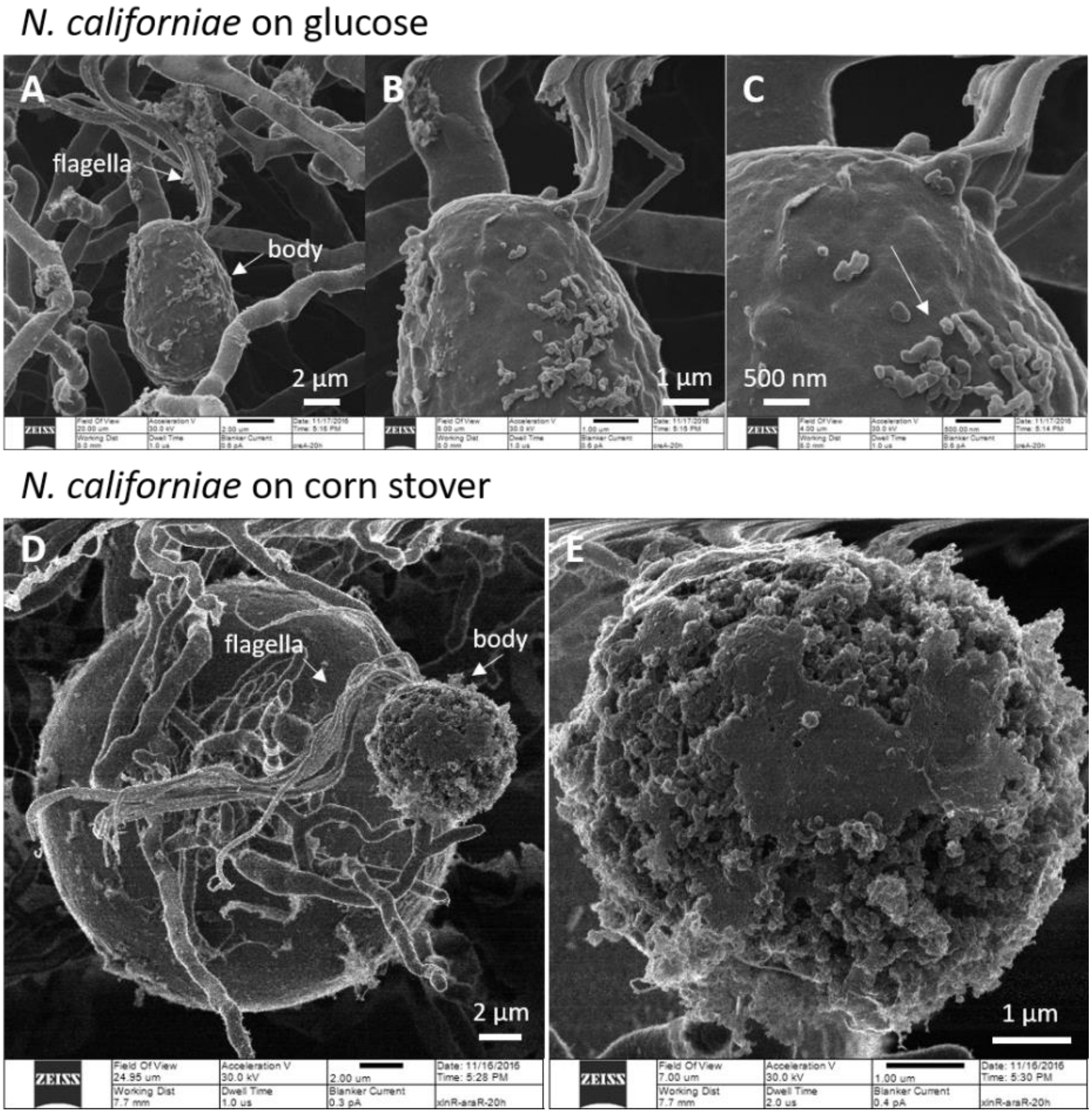
Figure 2.7. Zoospores actively express and display cellulosome proteins during growth on simple and complex substrates. Micrographs of *P. finnis* zoospores grown on Whatman paper show intense anti-dockerin and anti-GH48 signal (B-C) relative to a secondary antibody-only control (A). Representative zoospores are labeled by white arrows. Panels A-C were all processed by Zen using the same display settings and gamma parameters. During growth on cellobiose, anti-dockerin and anti-GH48 is restricted to *P. finnis* zoospores with no staining of mature cells (E-F). *Neocallimastix californiae* demonstrates the same staining pattern when grown on glucose, suggesting that substrate-dependent cellulosome rhizoid display and substrate-independent zoospore cellulosome display are general phenotypes of monocentric anaerobic fungi (H-I). Panels D and G represent negative controls stained only with donkey anti-rabbit AlexaFluor 488. The images are representative of at least two technical replicates per carbon source, with one biological replicate for cellulose paper cultures and two biological replicates for cellobiose and glucose cultures. All confocal micrographs were obtained with a LD C-Apochromat 40x/1.1 W Korr M27 objective. Antibodies and stains used: (A-C) Calcofluor white (gray), (D-I) DAPI (blue); GH48 – rabbit anti-GH48 (primary), (A-C) anti-rabbit AlexaFluor 546 (secondary, yellow), (F,I) anti-rabbit AlexaFluor 488 (secondary, green); Dockerin – rabbit anti-dockerin (primary), (A-I) anti-rabbit AlexaFluor 488 (secondary, green).

progression through the fungal life cycle, in which localization shifts from the zoospore body during the substrate search to the growing rhizoids after substrate encystment.

Since cellulosome and CAZyme expression is highly dependent on the growth substrate [42], we hypothesized dockerin and GH48 localization patterns would change as a function of substrate complexity. To test this hypothesis, we stained formaldehyde-fixed samples of *Piromyces finnis* grown on the soluble disaccharide cellobiose for 72-96 hours with anti-dockerin and anti-GH48 antibodies. Intriguingly, zoospores remained hot spots for dockerin and GH48 signal under these conditions, but the rhizoids of thalli appeared devoid of both proteins (Figure 2.7E-F). The same patterns were observed for glucose-grown samples of a different anaerobic fungus, *Neocallimastix californiae* (Figure 2.7H-I), indicating that the unique expression and display of cellulosomal CAZymes by zoospores during growth on soluble sugars may be general to monocentric anaerobic fungi. Consistent with the immunofluorescence observations, HIM micrographs of *N. californiae* zoospores from cultures grown on both glucose and corn stover also indicated the presence of surface-displayed proteins, as evidenced by the clustered globular structures apparent on zoospore bodies under both growth conditions (Figure 2.8).

2.3 Discussion

Immunofluorescence microscopy with antibodies generated against the fungal dockerin domain and GH48 domain provided a useful tool to visualize cellulosome and CAZyme localization in anaerobic fungi grown on an array of carbon substrates. Our observation that cellulosome-associated dockerin domains and GH48 catalytic domains are highly localized to the rhizoid networks of mature *Piromyces finnis* cells during growth on a cellulosic substrate is strong evidence this rhizoidal cell morphology is an important driver of plant


N. californiae on glucose

N. californiae on corn stover

Figure 2.8. Possible cellulosomes observed on the surface of *N. californiae* zoospores. (A-C) When grown on glucose, *N. californiae* zoospores imaged by HeIM appear to show a relatively smooth surface with clusters of globular structures as annotated by the white arrow in panel C). In contrast, a zoospore imaged by HeIM from a culture grown on corn stover has a much rougher surface, characteristic of the presence of many more surface proteins, which may be cellulosomes (D-E). It remains possible that this change in surface roughness is characteristic of different stages of zoospore development, but the increased deployment of surface-displayed cellulosomes would be consistent with the upregulation of degradative machinery observed for anaerobic fungi cultured on substrates more complex than glucose (K. Solomon et al, *Science*, **2016**). The presence of surface-displayed globular structures under both glucose and corn stover growth conditions is also consistent with the immunofluorescence results presented in Figure 5, suggesting these structures may indeed be cellulosomes.

matter deconstruction by anaerobic fungi in the rumen. Additionally, our localization data for both the GH48 and dockerin domains is consistent with a prior study of other anaerobic fungal species, which found unlysed cell pellets possessed cellulolytic activity comparable to the culture supernatant, suggesting surface display of cellulolytic enzymes [108]. Given these results and the known genotypic and phenotypic similarities among anaerobic fungi, we speculate that cellulosome proteins localize to the rhizoids of rhizoid-forming anaerobic fungi from other genera besides *Piromyces* as well. While HIM micrographs captured cellulosome-like structures on the rhizoid networks of mature fungal cells, we were unable to address how fungal cellulosomes are attached to cell structures with our failure to observe colocalized ScaA and dockerin or GH48 signal in fixed cell samples. An alternative strategy to address this gap is likely needed, since the anti-ScaA antibody target shares homology with many classified scaffoldins, a large group of proteins encompassing a wide size range that also includes polypeptides with and without transmembrane anchors.

In addition to mature fungal rhizoids, zoospore surfaces also showed intense dockerin and GH48 localization. Some of the earliest investigations of anaerobic fungi following their initial discovery found that cell lysates from captured zoospores of three different genera all contained many of the enzymatic activities present in the fungal supernatant after growth, including cellulase and hemicellulase activity, highlighting the potential importance of this life stage in the biomass degradation process [109]. Our data, consistent with this previous evidence that motile zoospores play a degradative role, provides the first evidence that zoospores also display cellulosomes on their surface, suggesting a broader physiological role for cellulosomes that likely assist with both substrate attachment and degradation, as is the case with cellulolytic bacteria [110–112].

A particularly interesting finding of our work is that fungal zoospores, in contrast to thalli and zoosporangia, display cellulosomes during growth on simple soluble substrates such as glucose or cellobiose. Previous work demonstrated that culture supernatants from glucose or cellobiose-grown cultures contained significantly less CAZyme activity and abundance compared to those from cells grown on lignocellulosic substrates [42]. It was subsequently shown by bulk RNA sequencing that CAZyme expression in three strains of anaerobic fungi is strictly catabolite repressed by these simple carbohydrates [113]. However, RNA from cells at all stages of the anaerobic fungal life cycle was used in these analyses, confounding any differences in gene expression response by cells at different life stages. This immunofluorescence microscopy approach provides the first evidence that production and display of major cellulosome protein domains is regulated differently by zoospores and thalli or sporangia. While we could not quantitatively measure changes or lack thereof in abundance of labeled zoospore surface proteins between growth conditions, the decrease in rhizoid-localized dockerin and GH48 to sub-detectable levels during growth on glucose or cellobiose relative to lignocellulose supports the hypothesis that production of major cellulosome proteins may be uniquely constitutive in zoospores, regardless of growth conditions. This hypothesis would partially explain why a small subset of CAZyme- and dockerin-encoding genes showed no expression level change by RNA-seq analysis of *P. finnis* cells grown on glucose vs. lignocellulose [42]. Indeed, separate transcriptomic analysis of zoospores and sporangia of a parasitic chytrid, *Batrachochytrium dendrobatidis*, under the same environmental conditions found that more than half of the genes in the genome exhibited differential expression between the two life stages, including several peptidase gene families involved in pathogenicity, emphasizing the importance of cellular life stage in chytrid expression patterns and phenotype [114].

The potential importance of the anaerobic fungal life cycle in CAZyme production by these organisms is also reflected in the shifting cellulosomal CAZyme localization from zoospore body to mature cell rhizoid that was concomitant with cell maturation. As part of their development, monocentric anaerobic fungal zoospores shed or absorb their flagella, encyst, then germinate to form rhizoids that grow into plant material, while the nucleus-containing zoospore body develops into the sporangium [69,88]. It is logical that fungal thalli direct their cellulolytic machinery to the rhizoids after attachment to a carbon source, but the cell biology of protein trafficking and secretion in anaerobic fungi has not been studied, though it is clearly of great importance given these organisms' prolific enzyme production when cultured on lignocellulosic substrates [42]. Recent work has demonstrated that rhizoids of chytrid fungi closely resemble hyphae of filamentous fungi (Ascomycota), which are better characterized [67,115]. Enzyme secretion in filamentous fungi such as *Aspergillus niger* is known to occur predominantly at the hyphal tips in a growth-coupled process [97]; inhibition of processes that enable hyphae growth such as actin polymerization significantly reduced enzyme secretion and localization to the extracellular cell wall [116]. Addition of the actin polymerization inhibitor cytochalasin B or a cell wall synthesis inhibitor, caspofungin, similarly stunted rhizoid growth in the chytrid *Rhizoclosmatium globosum*, but changes in protein secretion were not investigated [67]. Given these established similarities, we hypothesize that CAZyme secretion is coupled to rhizoid growth during culturing on insoluble substrates, making the growing thallus life stage of interest for optimizing protein production by anaerobic fungi.

2.4 Conclusions

In summary, immunofluorescence microscopy of native anaerobic fungal cultures using antibodies raised against major fungal cellulosome protein domains provided strong evidence for the cell-associated *in vivo* localization of these molecular machines under different growth conditions. We confirmed the importance of rhizoids as centers for fungal cellulosome localization and biomass hydrolysis in mature fungal cells and established foundational data that zoospores display cellulosome proteins, paving the way for future study of how cellulosomes impact zoospore biology. Our approach also avoided the difficulties inherent in traditional omics-type analysis of organisms with complex life cycles to uncover uniquely constitutive production of cellulosome components by zoospores regardless of substrate complexity. These findings highlight how life cycle-dependent cell morphology and cellulosome localization contribute to biomass degradation by anaerobic fungi and importantly, provide new support for the significance of zoospores in that process. Elucidating the uncharacterized regulatory relationships between life cycle progression and cellulosome production and function will benefit both laboratory efforts to engineer fungal cellulosomes for nanobiotechnology and industrial efforts to realize anaerobic fungi as platforms for bioprocessing. Higher resolution structural biology studies detailing how cellulosomes orient enzyme active sites with different chemistries are still needed and will complement these efforts nicely. Such insights into the natural system will ultimately be key in the successful deployment of anaerobic fungi or their cellulosomes in industrial biotechnology.

2.5 Materials and Methods

Cell culture and fixation

Anaerobic fungal isolates were grown anaerobically under a headspace of 100% CO_2 at 39°C in Hungate tubes containing CO_2-flushed Medium C [117] supplemented with various carbon sources at 1% (w/v) for insoluble substrates and 0.5% (w/v) for soluble substrates. Fungal cultures were passaged every 3-7 days to maintain viable cell populations by diluting 1mL of growing culture into 9mL of fresh media containing a carbon source. Cultures harvested for microscopy were incubated for 3-4 days after inoculation. For all samples, fungal colonies were scraped off the Hungate tube walls and the entire tube contents transferred to a 15mL Falcon tube. Fungal cells (and insoluble substrate if present) were pelleted in a fixed angle rotor at 3000x g for 3 minutes and resuspended in cold PBS + 4% (w/v) formaldehyde. After fixation for at least an hour at 4°C, samples were washed once with an equal volume of PBS to remove excess formaldehyde and stored in PBS at 4°C prior to analysis.

Antibody production

Genes encoding each of the three key protein domains (ScaA fragment, dockerin, and GH48) were cloned into the pET-28a vector (Addgene) for expression of the target with N- and C-terminal 6x His tags in *Escherichia coli* BL21 (DE3). For the ScaA fragment, the full gene product was used for animal immunization and antibody generation. Anti-dockerin and anti-GH48 antibodies were generated by animal immunization with synthetic 15-mer peptides taken from representative full dockerin and GH48 sequences. The three amino acid sequences used for antibody generation are included in Supplementary Table S1. Purified protein product for immunization was prepared by protein expression and cell lysis followed by immobilized metal affinity chromatography (IMAC). Briefly, *E. coli* strains were grown at 37°C in Luria-Bertoni (LB) media supplemented with 50 μg/mL kanamycin. Protein synthesis was induced when the cells reached an absorbance at 600 nm of ∼0.6 by adding 0.1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) to the medium. Cultures were incubated at 30°C overnight (16-24 hrs) following induction. To harvest protein product, cells were pelleted by centrifugation at >3,000 x g for >10 min and resuspended in 1x PBS + 10mM imidazole (pH 7.4) at 1% of the original culture volume. Cells were lysed by vortexing rigorously with 0.5 mm silica beads and soluble supernatant recovered by centrifugation at 10,000 x g for 10 minutes. 6x His-containing protein was purified from the soluble supernatant using HisPur™ Ni-NTA resin following the manufacturer's instructions (Thermo Fisher Scientific).

Purified protein products were sent to Genscript as needed for antibody generation. All antibodies were verified with enzyme-linked immunosorbent assay (ELISA) titers $\geq 64{,}000$.

Western blot

Cellulosome proteins isolated from *P. finnis* cultures as described in [82] were separated by SDS-PAGE and subsequently blotted onto a PVDF membrane using a Bio-Rad TransBlot Turbo Transfer System (Bio-Rad Laboratories, Hercules, CA). The membrane was then blocked with Tris-buffered saline + 0.1% Tween-20 (TBS-T) supplemented with 5% milk powder for one hour at room temperature. After being washed with TBS-T three times, the membrane was incubated with primary antibody for one at 4°C. To perform the antigen blocking experiment, the membrane, containing identical protein samples, was split in two.

1 μg/mL primary antibody in TBS-T was used to label the control half, while 1 μg/mL primary antibody pre-incubated with 5 μg/mL immunizing peptide (sequence in Table S1) for 1 hour prior to labeling. Both halves were subsequently labeled with goat anti-rabbit HRP-conjugated secondary antibody (ThermoFisher Scientific #31460). Blots were then developed with ECL blotting substrate (ThermoFisher Scientific #32209) and imaged using a ChemiDoc imaging system (Bio-Rad Laboratories, Hercules, CA).

Immunofluorescence microscopy

Formaldehyde-fixed samples were washed three times in PBS for 10 min, blocked with 1% BSA for 1 hour, and incubated with primary antibody overnight at 4°C. Antibody specifications were as follows: anti-ScaA, mouse monoclonal, 10 mg/ml in PBS, anti-dockerin and anti-GH48, rabbit polyclonal each at 10 mg/ml in PBS + 1% BSA. After primary antibody incubation, samples were washed with PBS and incubated at room temperature for 1 hour with a corresponding secondary antibody - goat anti-mouse AlexaFluor 647 (ThermoFisher), donkey anti-rabbit IgG AlexaFluor 488 (ThermoFisher), or donkey anti-rabbit IgG AlexaFluor 594 (ThermoFisher) at 1 mg/ml concentration in PBS. Unless otherwise specified, samples were labeled with both anti-GH48 and anti-dockerin antibodies, necessitating a multi-step labeling process. After the first labeling step, dual-labeled samples were washed 3 times in PBS followed by fixation with 4% paraformaldehyde (Electron Microscopy Sciences, Inc.) for 10 minutes at room temperature. The above-described labeling process was repeated with appropriate primary and secondary antibodies. After immunolabeling, the samples were washed 3 times with PBS and finally counterstained with Calcofluor White stain (Sigma Aldrich). Confocal microscope images were acquired at 1 mm z-steps on a Zeiss LSM 710 scanning head confocal microscope with a Zeiss plan apo 40X/1.1 objective. Excitation lasers were 405, 488, 561 and 633 nm for the blue, green, orange and red emission channels, respectively. Laser dwell times were 0.79 ms each channel. Image processing was completed with Zen (Zeiss), ImageJ (NIH) and Volocity (Quorum Technologies Inc.).

Cryosectioning

Paraformaldehyde fixed samples were placed into cryomolds, embedded in Tissue-Plus O.C.T Compound (ThermoFisher) and then frozen at -20°C. Thin sections were generated on a CryoStar NX70 Cryostat (ThermoFisher) and placed onto a #1 coverslip. The thin sections were then stained with 5 µg/mL FM 1-43 membrane stain (ThermoFisher) and 1X SYB Gold Nucleic Acid Stain (ThermoFisher). Tiled confocal microscope images were acquired at 2 mm z-steps on a Zeiss LSM 710 scanning head confocal microscope with a Zeiss plan apo 40X/1.1 objective. Excitation lasers were 405 and 488 nm for the blue and green emission channels, respectively. Laser dwell times were 2.55 ms each channel. Image processing was completed with Zen software (Zeiss).

Helium ion microscopy

Helium ion microscopy experiments were performed as described in (45). Specifically, fungi grown on various substrates were chemically fixed with 2% glutaraldehyde (Sigma–Aldrich) and dehydrated through a series of 10 ml step-gradients from 0% to 70% ethanol then centrifuged at 4°C (3,000g for 2 min). Samples were washed twice more with 10 ml of 100% ethanol for 15 min, then centrifuged and finally resuspended in 5 ml of 100% ethanol to remove any residual water. Fungal and/or plant biomass suspensions in 100% ethanol were gently extracted by wide-mouth pipet and placed onto stainless steel carriers for automatic critical point drying (CPD) using an Autosamdri-815 (Tousimis, Rockville, MD), with CO2 as a transitional fluid. The CPD-processed biomass was mounted onto aluminum stubs and sputter coated with approximately 10–20 nm of conductive carbon to preserve the sample surface information and minimize charge effects. Secondary electron images of the samples were obtained using an Orion helium ion microscope (HIM) (Carl Zeiss Microscopy, Peabody, MA) at 25 or 30 keV beam energy, with a probe current range of 0.1–1 pA. Prepared samples were transferred into the HIM via load-lock system and were maintained at $\sim 3 \times 10^{-7}$ Torr during imaging. Use of a low energy electron flood gun (∼500 eV) was applied briefly interlaced with the helium ion beam that enabled charge control to be maintained from sample to sample. The image signal was acquired in line-averaging mode, with 16 lines integrated into each line in the final image with a dwell time of 1 μs at a working distance range of 7–8 mm. Charge neutralization was applied to the sample after each individual line pass of the helium ion beam, which displaced charges on the surface and minimized charging effects in the final image. No post-processing procedures were applied to the digital images besides standard noise reduction, brightness, and contrast adjustment using Photoshop plugins.

III. Isolation and Characterization of Native Cellulosomes from *Neocallimastix californiae*

3.1 Introduction

Waste biomass (lignocellulose) produced from agricultural and municipal sources is an abundant, renewable resource that remains largely untapped for chemicals manufacturing. At the molecular level, lignocellulose is a composite material of three biopolymers – cellulose, hemicellulose, and lignin – which can be upgraded to higher value chemicals via a two-step process of catalytic depolymerization into monomers and conversion of those monomers to desired products. This process naturally occurs on an enormous scale, in which microbes deploy a suite of enzymes to hydrolyze lignocellulose into sugars and other breakdown products that are metabolized to fuel microbial growth.

Biological lignocellulose hydrolysis requires the coordinated action of a cocktail of enzymes that attack the material's different polymers to produce a mixture of soluble sugars and aromatic compounds for downstream conversion. A major outstanding goal for the bioprocessing community is to improve hydrolytic enzyme cocktail performance in industrial realizations of this process, as enzyme cost and hydrolysis sugar yield represent key cost drivers of biomass valorization process economics [118,119].

Anaerobic fungi, which inhabit the guts of herbivores, are lignocellulose deconstruction specialists that present a promising source for better lignocellulolytic enzyme cocktails. Many anaerobic fungi proliferate through a multi-staged life cycle in which motile, flagellated zoospores encyst on plant matter, growing root-like, enzyme-bearing rhizoid structures that penetrate and engulf biomass [120]. Like many species of anaerobic bacteria, anaerobic fungi produce cellulosomes, protein complexes that modularly tether together enzymes to enhance biomass hydrolysis [121]. Indeed, cellulosomes are estimated to harbor >80% of the cellulose degrading activity of anaerobic fungi, making them attractive for use in industrial scale biomass hydrolysis [62].

Fungal cellulosomes are thought to assemble via non-covalent interactions between enzyme-fused dockerin domains and cohesin domains repeated on a central scaffoldin

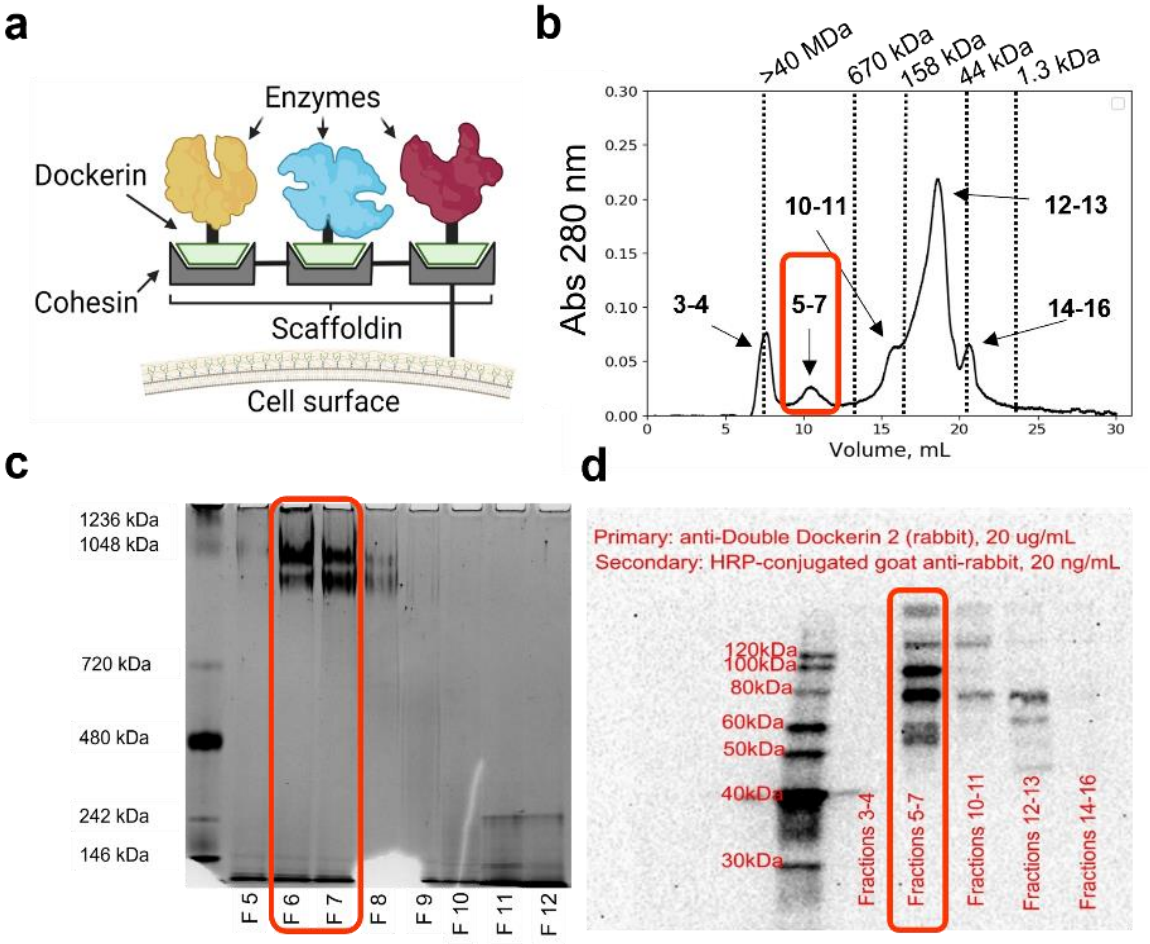


Figure 3.3. The affinity digest purification method produces fungal cellulosome samples that are high MW, dockerin-rich, enzyme complexes. (a) Fungal cellulosomes are thought to be modular, cell-localized complexes that assemble via non-covalent interactions between dockerin domains and cohesins. (b) SEC of a N. californiae cellulosome prep using a Superose 6 Increase column produces a clear high MW peak. Numbers indicate fraction numbers collected during a run. (c) Native PAGE of collected fractions, showing fractions 5-7 contain species with predicted MW in the MDa range. (d) SDS-PAGE and Western blot of pooled SEC fractions with an anti-dockerin antibody indicates fractions 5-7 contain many dockerin-containing species. A) Created with Biorender.

protein (Fig. 3.1a). While fungal dockerin has a known structure and is proven necessary for cellulosome incorporation [122], the identity of fungal cohesin remains elusive, making it impossible to reconstitute fungal cellulosome in recombinant systems for *in vitro* characterization and engineering [81].

Our understanding of fungal cellulosomes and their role in anaerobic fungal biology took a major leap forward with the sequencing of several *Neocallimastigomycota* genomes [82,93,123], which uncovered a large catalog of dockerin-encoding genes with functional domains including nearly one hundred different carbohydrate-active enzyme (CAZyme) families [124], proteases, serpins, kinases, phosphatases, and many domains of unknown function (mycocosm.jgi.doe.gov). A significant roadblock in developing fungal cellulosomes as components of industrial cellulase cocktails, however, is a limited understanding of fungal cellulosome biochemistry and structure. Such knowledge is crucial for engineering optimized lignocellulolytic enzyme systems that improve bioprocess performance [125].

An outstanding challenge towards characterizing fungal cellulosome structure and biochemistry has been the purification of active cellulosome complexes to high purity at high yield [62,82,100]. As a result, no experimentally supported model of fungal cellulosome structure exists, no high resolution catalog of cellulosome complex composition has been measured, and no detailed fungal cellulosome lignocellulose hydrolysis kinetics have been reported.

Here, we demonstrate a robust method for purifying cellulosome complexes from the anaerobic fungus *Neocallimastix californiae* that enables in depth characterization of cellulosome structure, composition, and biochemistry. We apply this method to isolate cellulosomes from *N. californiae* cultured on different biomass substrates, quantifying how the relative enzyme content and lignocellulose hydrolysis rate and yield change as a function of growth condition using mass spectrometry-based proteomics and enzyme activity measurements. Finally, we present the first cryo-electron tomogram of a fungal cellulosome complex, showing a “grapes on a bunch” structure for a *N. californiae* cellulosome containing several bound enzymes and a crystalline cellulose-interacting domain.

3.2 Results and Discussion

Validating a high-yield, reproducible method for N. californiae cellulosome isolation

Structural and biochemical characterization of fungal cellulosomes requires milligram quantities of pure sample, which were difficult to produce with published methods for fungal cellulosome isolation [82]. To purify fungal cellulosomes for downstream characterization, we evaluated several methods by three criteria – 1) isolation of distinct, high molecular weight (MW) protein complexes with activity against biomass, 2) enrichment for dockerin-containing proteins, and 3) method yield and purity. An adapted version of the affinity digest purification method developed for isolating cellulosomes from *Clostridium thermocellum* [126] proved best. Other methods we tested included cellulosome adsorption and desorption from Avicel [101] and an alcohol-based method for collecting lignocellulose-bound proteins [127] (data not shown). In the affinity digest method, anaerobic fungi are grown to stationary phase, and cellulosome complexes are adsorbed from the supernatant by adding phosphoric acid swollen cellulose (PASC). Weakly bound proteins are removed by washing the PASC. Rather than isolating cellulosomes by desorption, which is inefficient [101], the enzyme-PASC mixture is dialyzed against acidic buffer at elevated temperature. During dialysis, adsorbed cellulosomes solubilize the PASC adsorbent until

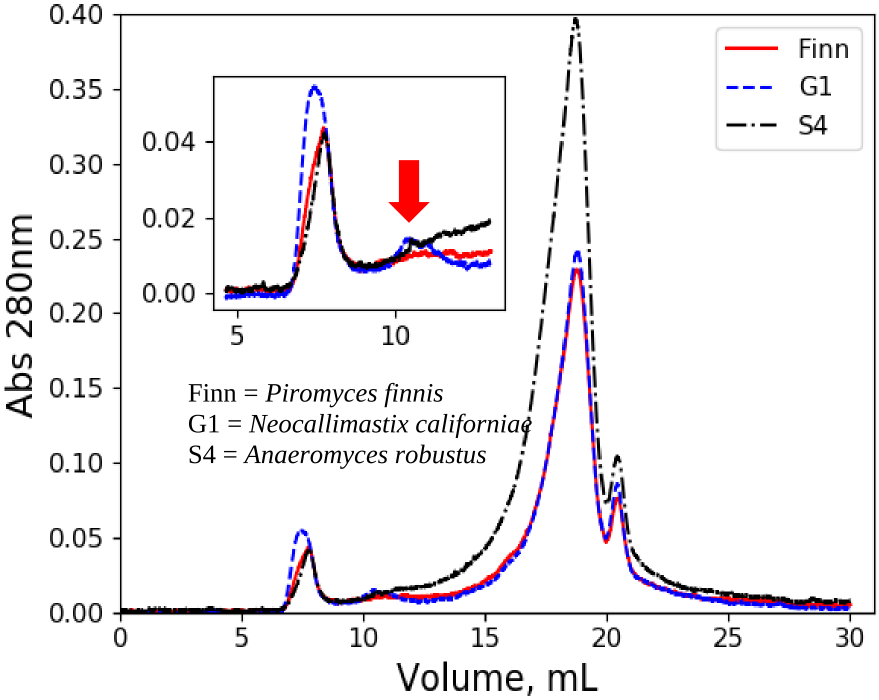


Figure 3.4. No clear high MW peak is observed by SEC when performing this cellulosome purification method on two other anaerobic fungal species.

they become free in solution, after which the soluble protein complexes are concentrated and separated from low MW species by size exclusion chromatography (SEC). Complexes prepared using this method met our criteria (Fig. 3.1b-d). From *N. californiae* grown on lignocellulose, this method produces about 6 mg purified complex per liter culture and robustly produces complexes with consistent SDS-PAGE profiles across trials (Fig. 3.3b). We only observed a clear high MW peak when using the affinity digest purification method on *N. californiae* cultures; *Anaeromyces robustus* and *Piromyces finnis* affinity digest purifications did not yield clear high MW species in SEC (Fig. 3.2).

N. californiae alters cellulosome composition in response to different substrates

Immunofluorescence microscopy studies of cellulosome localization by anaerobic fungi under different growth conditions uncovered that only immature zoospore cells display cellulosomes during growth on simple sugars such as glucose or cellobiose [120]. Furthermore, transcriptomic analysis of gene expression patterns across growth conditions suggests that cellulosome enzyme production varies when anaerobic fungi are challenged with substrates of different physical character or containing different sets of chemical bonds [128]. Towards elucidating fungal cellulosome composition-activity relationships, we hypothesized we could purify cellulosomes with different composition profiles by growing *N. californiae* on different substrates. To measure how *N. californiae* alters the composition of its cellulosomes in response to growth on substrates with different physicochemical properties, we isolated and characterized via shotgun proteomics cellulosomes from *N. californiae* grown on cellobiose, filter paper, and reed canary grass. Cellobiose (CB) is a soluble disaccharide with one β-1-4 glycosidic bond. Filter paper (FP) is nearly pure, insoluble cellulose of high crystallinity, containing mostly β-1-4 glycosidic bonds. Reed canary grass (RCG) is an insoluble lignocellulose similar to switchgrass, containing approximately 40% cellulose, 25% hemicellulose, 22% lignin, and 13% ash and other biomolecules [129].

All growth conditions yielded cellulosome complexes of about the same size (Fig. 3.3a), suggesting either that our purification method biases for certain cellulosome complexes or that the scaffoldins anchoring subunits into these various complexes have approximately the same number of dockerin binding sites. We observed consistent SDS-PAGE profiles between cellulosome purification trials and identified two major bands present across all growth conditions, with discernable changes in banding patterns among minority species (Fig. 3.3b). From shotgun proteomics of cellulosome preparations from cellobiose, filter paper, and RCG in biological triplicate, we identified in fungal cellulosomes catalytic domains from 14 unique Glycoside Hydrolase (GH) families, 3 unique Carbohydrate

Esterase (CE) families, and three other protein families ("distantly related to plant expansins," spore coat CotH proteins, and putative scaffoldins) with mean sequence coverages above 25%; several hits had no functional annotations besides possessing dockerin domains. Protein sequence coverage for individual database proteins with peptide-spectrum matches passing filter criteria ranged from 63% to <1% (Table A1).

We quantified the relative cellulosome content of different enzyme families using the normalized spectral abundance factor (NSAF) [130], which correlates with molar abundance. We identified five enzyme families whose relative compositions exceeded 10 mol % and found six enzyme families whose relative composition changes significantly between any two of the three growth conditions (Fig. 3.3c, Table A2). Consistent with previous reports [131], Glycoside Hydrolase (GH) family 48 reducing end-acting exocellulases constitute the major cellulosome component under all growth conditions, about 37-48 mol % of cellulosome enzymes. Putative endoglucanases (GH5,GH9, GH45) and non-reducing end-acting cellobiohydrolases (GH6) consistently make up another 21-27% of cellulosome subunits, implying the core exocellulase-endocellulase synergism from bacterial cellulosomes is present in fungal cellulosomes as well [132–134]. Based on the protein molecular weights, the prominent bands in Fig. 3.3b are likely GH48 and GH9 species (~80 kDa) and GH5 and GH6 species (~60 kDa). GH2 and GH3 domains, which often contain β-glucosidase activity, consistently make up another 1.2-1.5 mol % of cellulosome content, enabling these complexes to produce glucose as the final cellulose breakdown product instead of cellobiose, the final breakdown product of bacterial cellulosomes [135].

Expansins, non-catalytic proteins that loosen cell wall polymer matrices to facilitate their hydrolysis [136], comprise 5-12% of cellulosome enzymes. Consistent with the insoluble, high

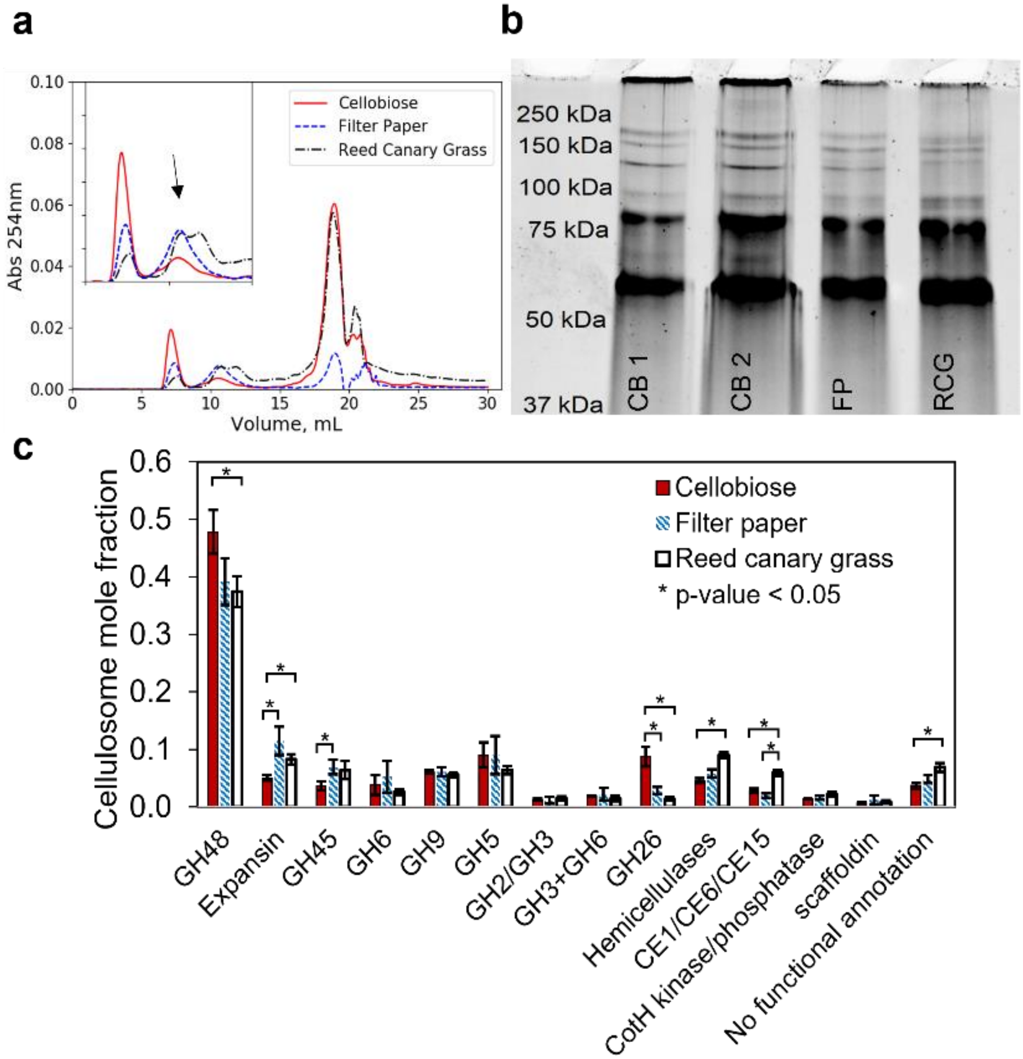


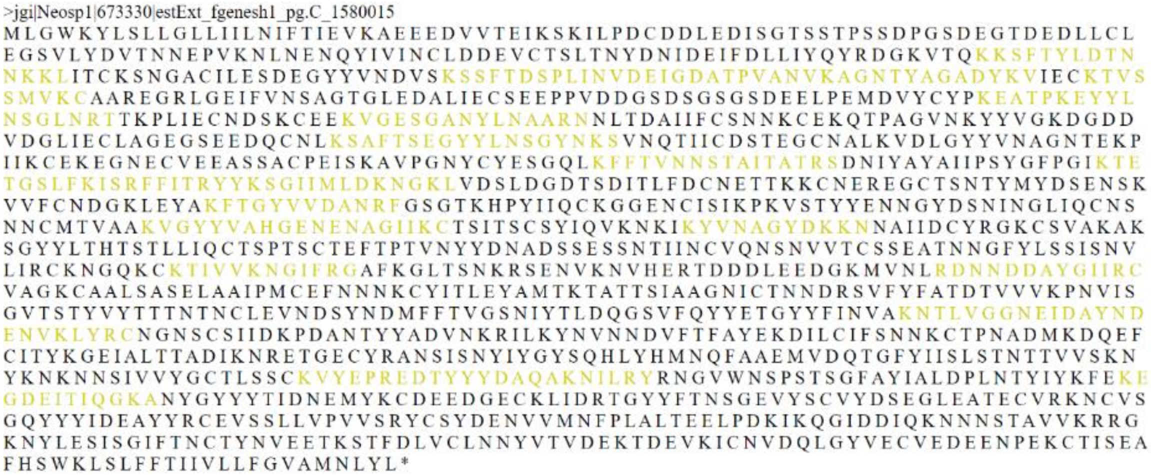
Figure 3.3. ***N. californiae*** **cellulosome composition changes as a function of growth substrate.** (a) SEC traces of cellulosome preps from N. californiae cultured on cellobiose, filter paper, and reed canary grass all produce a high MW species eluting around 10 mL. (b) The SDS-PAGE profile of cellulosome complexes purified from different growth conditions are largely similar. Here CB, FP, and RCG correspond to cellobiose, filter paper, and reed canary grass respectively. CB1 and CB2 represent two independent cellulosome purification experiments. (c) Estimated mole fraction of enzyme families in different cellulosome samples computed using normalized spectral abundance factors (NSAF). Error bars represent one standard deviation of three independent trials. Pairwise comparisons denoted with a * were statistically significant with $p < 0.05$ (unpaired Student's t-test; df calculated for each pair).

degree of polymerization of filter paper and reed canary grass, expansins are further enriched in cellulosomes isolated from FP and RCG cultures. Annotated expansins are notably absent in most bacterial cellulosomes; *Clostridium clariflavum* is one of the only anaerobic bacteria producing expansin-containing cellulosomes, and in this context, expansins significantly enhance cellulose hydrolysis rates [137,138].

A multi-domain protein family containing GH3 and GH6 domains with complementary cellobiohydrolase and β-glucosidase activities is prevalent in fungal cellulosomes across all growth conditions at 1.5-2 mol %. Multi-domain cellulases from other organisms such as *Caldicellulosiruptor bescii* are highly efficient cellulose degraders [43] suggesting these GH3+GH6 enzymes may be useful for industrial bioprocessing applications. Other database proteins with multiple catalytic domains detected in cellulosome samples included a GH9+GH43 dockerin-containing protein and a CE1+GH11 dockerin-containing protein (Table A1, ID 230365, 414424).

Despite a demonstrated preference by anaerobic fungi for degrading hemicellulose over cellulose [128], Hemicellulases (GH10,11,16,30,39,43,53,74) and carbohydrate esterases (CE1,6,15) constitute a relative minority in *N. californiae* cellulosomes, implying high specific activity or suggesting hemicellulase activity is prevalent among enzymes families traditionally annotated as cellulases. These enzyme families encompass a wide diversity of hemicellulase activities including endo- and exo-xylanase, xyloglucanase, arabinofuranosidase, galactanase, acetyl esterase, hydroxycinnamoyl esterase, and 4-O-methyl-glucuronyl methylesterase activities (cazy.org). As expected, hemicellulases and carbohydrate esterases were enriched in RCG cellulosomes (Fig. 3.3c). GH26 mannanases were uniquely enriched in cellobiose culture cellulosomes. Since only motile zoospores produce cellulosomes when grown on cellobiose [120], we speculate this mannanase may play a key role in directing where zoospores encyst on plant matter; mannan is the main structural component of hemicellulose in angiosperms (grasses), the main diet of the animals in which many anaerobic fungi have evolved [139].

The lowest abundance protein families with constant composition across all growth conditions were CotH kinases and phosphatases, and putative scaffoldins, a class of repeat-rich protein annotated in fungal genomes that show evidence of binding fungal dockerin [82]. Interestingly, CotH kinase-encoding genes are the second largest class of dockerin-containing genes across anaerobic fungal genomes, representing 18% of dockerin genes, second only to CAZyme dockerin genes (27%) (mycocosm.jgi.doe.gov). Their exact role in cellulosome biology remains to be elucidated, but a dockerin-fused protein homologous to CotH from *Bacillus subtilis* is a known component of the *Clostridium thermocellum* cellulosome posited to play a structural role in cellulosome formation [140].The putative scaffoldins are clearly present in all samples, and mapping peptides to the sequence suggests these large proteins were intact in the samples (Fig. 3.4).



```
>jgi|Neosp1|673330|estExt_fgenesh1_pg.C_1580015
MLGWKYLSLLGLLIILNIFTIEVKAEEEDVVTEIKSKILPDCDDLEDISGTSSTPSSDPGSDEGTDEDLLCL
EGSVLYDVTNNEPVKNLNENQYIVINCLDDEVCTSLTNYDNIDEIFDLLIYQYRDGKVTQKKSFTYLDTN
NKKLITCKSNGACILESDEGYYVNDVSKSSFTDSPLINVDEIGDATPVANVKAGNTYAGADYKVIECKTVS
SMVKCAAREGRLGEIFVNSAGTGLEDALIECSEEPPVDDGSDSGSGSDEELPEMDVYCYPKEATPKEYYL
NSGLNRTTKPLIECNDSKCEEKVGESGANYLNAARNNLTDAIIFCSNNKCEKQTPAGVNKYYVGKDGDD
VDGLIECLAGEGSEEDQCNLKSAFTSEGYYLNSGYNKSVNQTIICDSTEGCNALKVDLGYYVNAGNTEKP
IIKCEKEGNECVEEASSACPEISKAVPGNYCYESGQLKFFTVNNSTAITATRSDNIYAYAIIPSYGFPGIKTE
TGSLFKISRFFITRYYKSGIIMLDKNGKLVDSLDGDTSDITLFDCNETTKKCNEREGCTSNTYMYDSENSK
VVFCNDGKLEYAKFTGYVVDANRFGSGTKHPYIIQCKGGENCISIKPKVSTYYENNGYDSNINGLIQCNS
NNCMTVAAKVGYYVAHGENENAGIIKCTSITSCSYIQVKNKIKYVNAGYDKKNNAIIDCYRGKCSVAKAK
SGYYLTHTSTLLIQCTSPTSCTEFTPTVNYYDNADSSESSNTIINCVQNSNVVTCSSEATNNGFYLSSISNV
LIRCKNGQKCKTIVVKNGIFRGAFKGLTSNKRSENVKNVHERTDDDLEEDGKMVNLRDNNDDAYGIIRC
VAGKCAALSASELAAIPMCEFNNNKCYITLEYAMTKTATTSIAAGNICTNNDRSVFYFATDTVVVKPNVIS
GVTSTYVYTTTNTNCLEVNDSYNDMFFTVGSNIYTLDQGSVFQYYETGYYFINVAKNTLVGGNEIDAYND
ENVKLYRCNGNSCSIIDKPDANTYYADVNKRILKYNVNNDVFTFAYEKDILCIFSNNKCTPNADMKDQEF
CITYKGEIALTTADIKNRETGECYRANSISNYIYGYSQHLYHMNQFAAEMVDQTGFYIISLSTNTTVVSKN
YKNKNNSIVVYGCTLSSCKVYEPREDTYYYDAQAKNILRYRNGVWNSPSTSGFAYIALDPLNTYIYKFEKE
GDEITIQGKANYGYYYTIDNEMYKCDEEDGECKLIDRTGYYFTNSGEVYSCVYDSEGLEATECVRKNCVS
GQYYYIDEAYYRCEVSSLLVPVVSRYCSYDENVVMNFPLALTEELPDKIKQGIDDIQKNNNSTAVVKRRG
KNYLESISGIFTNCTYNVEETKSTFDLVCLNNYVTVDEKTDEVKICNVDQLGYVECVEDEENPEKCTISEA
FHSWKLSLFFTIIVLLFGVAMNLYL*
```

Figure 3.4. Mapping peptides to scaffoldin sequence suggests the intact protein is present. The sequence of 673330, the most abundant scaffoldin protein detected by proteomics, is shown. Yellow letters indicate peptides mapping to that part of the sequence were present in the proteomics dataset.

Composition changes are not fully explained by gene expression shifts

A reasonable hypothesis explaining why cellulosome composition changes under different growth conditions would be that changes in gene expression account for composition shifts. To answer this question, we reanalyzed transcriptomic data previously published by our group [128], measuring transcripts mapping to the same enzyme families by which we grouped proteins in the proteomics data. We lacked transcriptomic data for *N. californiae* grown on filter paper and could only make direct comparisons between cellobiose-grown and RCG-grown *N. californiae*. Of the five enzyme families that experience shifts in relative composition (p-value < 0.05, two-sided t-test, df = 4) quantified by proteomics, GH26, hemicellulases, and carbohydrate esterases see concomitant shifts in

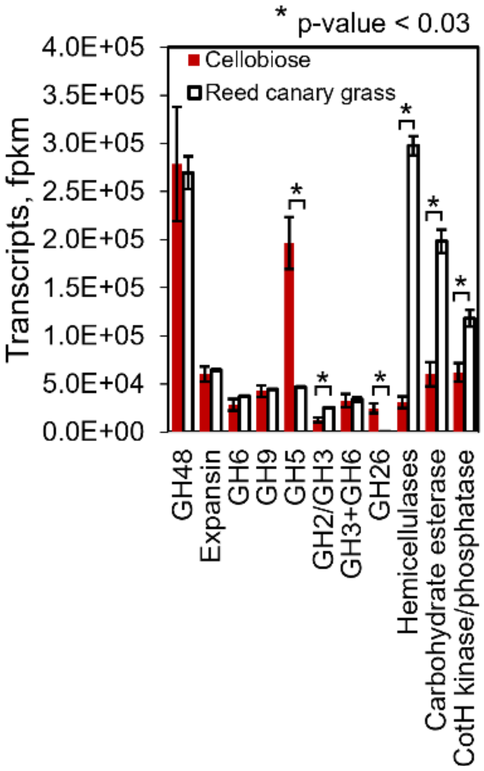


Figure 3.5. Gene expression changes do not universally explain changes in cellulosome composition between N. californiae grown on cellobiose vs reed canary grass. Summated transcripts for the same enzyme families as in (a) from N. californiae grown on cellobiose and reed canary grass. Data are presented as mean ± one standard deviation for N=3 biological replicates. Asterisks indicate a p-value < 0.03 for comparing sample means with an unpaired two-sided Student's t-test (degrees of freedom computed for each pair).

transcript count (Fig. 3.5). GH48 and expansin enzyme families have no change in gene expression to explain the protein-level cellulosome composition shift. For the CotH kinase/phosphorylase enzyme family, a 1.9-fold change in gene expression is not reflected in the cellulosome composition data. Statistically significant gene expression changes for family GH5 and GH2/GH3 enzymes were not reflected in cellulosome proteomics data.

Activity profiles distinguish cellulosomes isolated from different growth conditions

A major interest of the bioprocessing field is the development of enzyme cocktails with faster kinetics and improved activity against specific substrates. These properties are a function of the kinetics of the enzymes themselves, but also the composition and stoichiometry of enzymes within cellulosome complexes. An understanding of fungal cellulosome composition-activity relationships is critical to engineering cellulosomes for bioprocessing applications, but this is challenging because fungal cellulosomes with defined composition cannot be reconstituted heterologously. Since proteomics analyses indicated cellulosomes purified from *N. californiae* grown on different carbon sources differ in composition, we sought to identify cellulosome composition-activity relationships by characterizing the lignocellulose hydrolysis activity of native cellulosomes purified from *N. californiae* grown on cellobiose, filter paper, and reed canary grass.

We used nanostructure-initiator mass spectrometry (NIMS) [141] to measure simultaneously a range of breakdown products during 70 hour hydrolysis of three substrates – crystalline cellulose (Avicel), insoluble beechwood xylan, and ionic liquid pretreated switchgrass (IL-SG). Hydrolysis yields on all substrates varied substantially among three biological replicates of each cellulosome preparation, potentially caused by variable cellulosome enzyme content or by subunit activity losses during purification, which may be inconsistent across independent purification trials (Fig. 3.6). Technical replicate

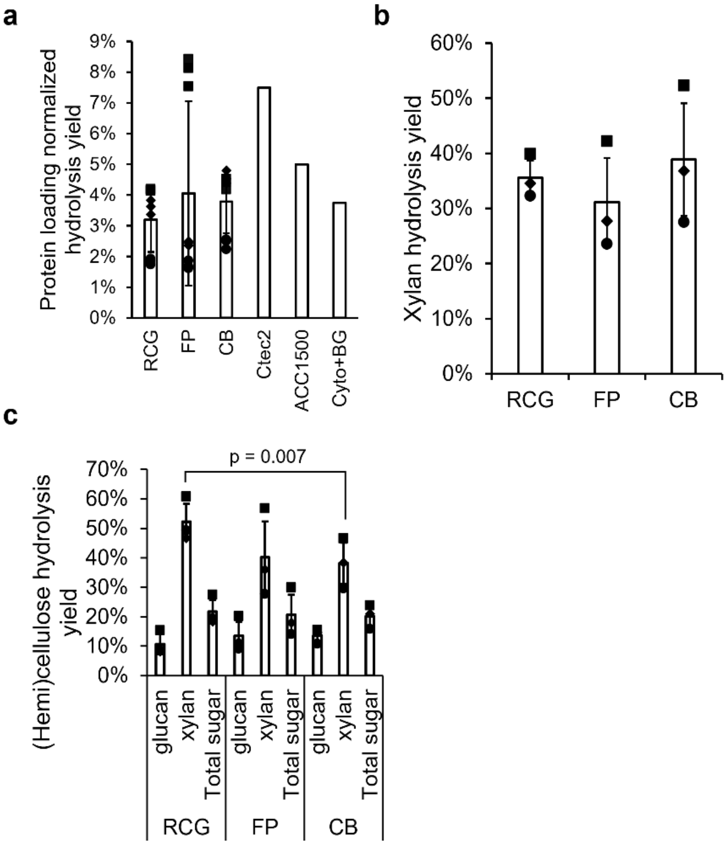


Figure 3.6. Hydrolysis yields after 70 hours are similar among cellulosomes purified from different grown conditions. RCG, FP, and CB represent cellulosome samples purified from N. californiae grown on reed canary grass, filter paper, and cellobiose respectively. (a) Total Avicel hydrolysis yield normalized by the mg protein supplied per g biomass, with computed values for three commercial cellulase cocktails, Ctec2, ACC1500, and Cyto+BG for comparison. Commerical cellulase cocktail yields were computed from reference 32. (b) Total beechwood xylan hydrolysis yield. (c) Total hydrolysis yield of glucan and xylan from ionic liquid pretreated switchgrass. In panel a), data are presented as mean ± one standard deviation for N=9 (technical triplicate measurements of 3 biological replicates). In b) and c), data are presented as mean ± one standard deviation of single measurements of N=3 biological replicates. Each shape corresponds to a specific biological replicate, i.e. the RCG square data points in a)-c) correspond to the same protein sample.

measurements for each biological replicate indicate this variability did not result from experimental error in activity experiments (Fig. 3.6a). Initial hydrolysis rates, which should be directly proportional to enzyme concentration assuming constant rate constants, varied 3.7-fold on average among cellulosome biological replicates – very large differences to attribute to protein loading variability (Fig. 3.7).

Fungal cellulosome preparations displayed Avicel hydrolysis yields on par with select commercial cellulase preparations when normalized for protein loading (Fig. 3.6a) [142]. Beechwood xylan hydrolysis yields measured in this work were comparable to commercial xylanase preparations, which hydrolyzed 35% of beechwood xylan after 20h at 37°C, though a direct comparison is not possible without knowing the exact protein loading used to test commercial xylanase (Fig. 3.6b) [143]. Fungal cellulosomes freed on average 21% of

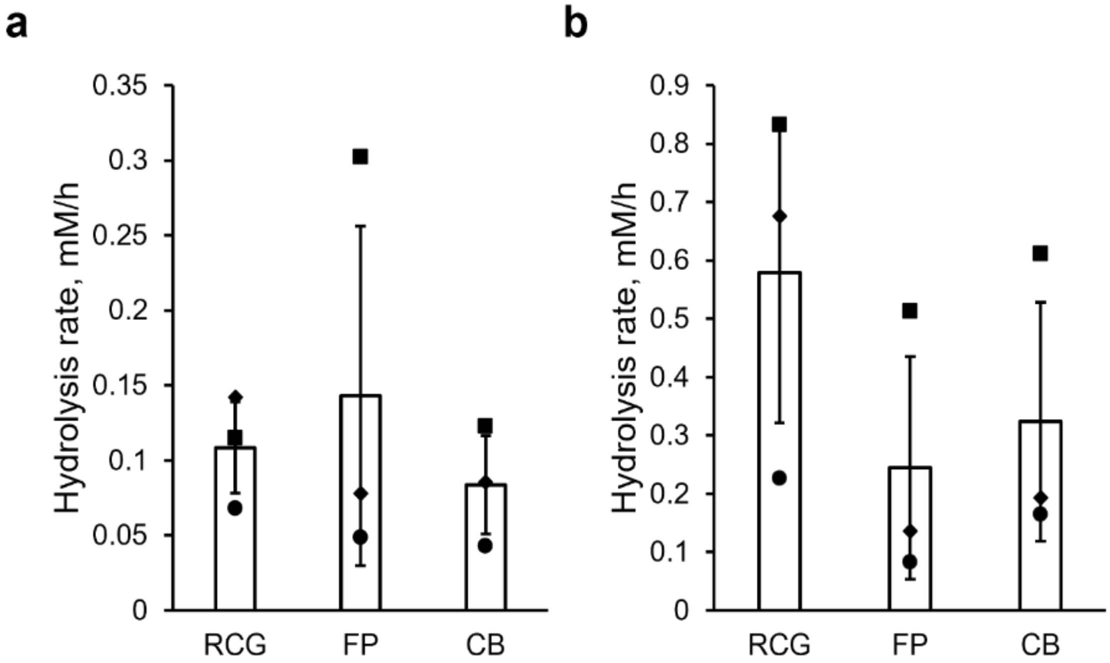


Figure 3.7. Initial hydrolysis rates exhibit high variability among biological replicates of cellulosome samples. The different shapes correspond to the same cellulosome sample replicates as in Figure S3 and Figure 4 in the main text. (a) Initial hydrolysis rate on Avicel. (b) Initial hydrolysis rate on beechwood xylan. In both charts, initial rates are computed as the slope of the sum of hydrolysis product concentrations (mM) vs time from 0 to 240 minutes.

sugars from IL-SG cellulose and xylan, with average cellulose hydrolysis yields across cellulosome samples of 13% and average xylan hydrolysis yields of 44% (Fig. 3.6c).

Across all cellulosome samples, xylan hydrolysis yields exceed cellulose hydrolysis yields by a minimum of three-fold, both in the context of purified biopolymers and in pretreated switchgrass (Fig. 3.6c, Fig. 3.8). This is consistent with previous reports detailing a degradative "preference" for hemicellulose over cellulose exhibited by anaerobic fungal supernatants [128]. A clear composition-activity dependence emerges in comparing switchgrass hydrolysis performance of CB and RCG cellulosomes. RCG cellulosomes demonstrated significantly higher switchgrass hemicellulose hydrolysis yield but comparable switchgrass cellulose hydrolysis compared to CB cellulosomes, likely due to enrichment of hemicellulases and carbohydrate esterases in RCG complexes (Fig. 3.3c, 3.6c). This relationship does not describe purified xylan hydrolysis by CB and RCG cellulosomes, perhaps because the hemicellulose within switchgrass is more substituted with diverse sugars, requiring a greater diversity of enzymes to deconstruct, while purified xylan is mainly a homopolymer of xylose. No cellulosome isolate displayed an apparent advantage in hydrolyzing cellulose when tested on Avicel or IL-SG, a surprising result given the fairly dramatic changes in GH48 and expansin content (Fig. 3.3c).

The distribution of final breakdown products from Avicel was similar across all samples – [glucose (G1)] > [cellobiose (G2)] > [cellotriose (G3)], consistent with the presence of β-glucosidase in each sample, but the relative fraction of glucose was lower for CB vs FP and RCG cellulosomes ($p = 0.047$ and $p = 0.088$ respectively, Fig. 3.9a). The initial rate ratio of glucose to cellobiose production during Avicel hydrolysis was also lower for CB cellulosomes, further suggesting a relative lack of β-glucosidase activity (Fig. 3.10a).

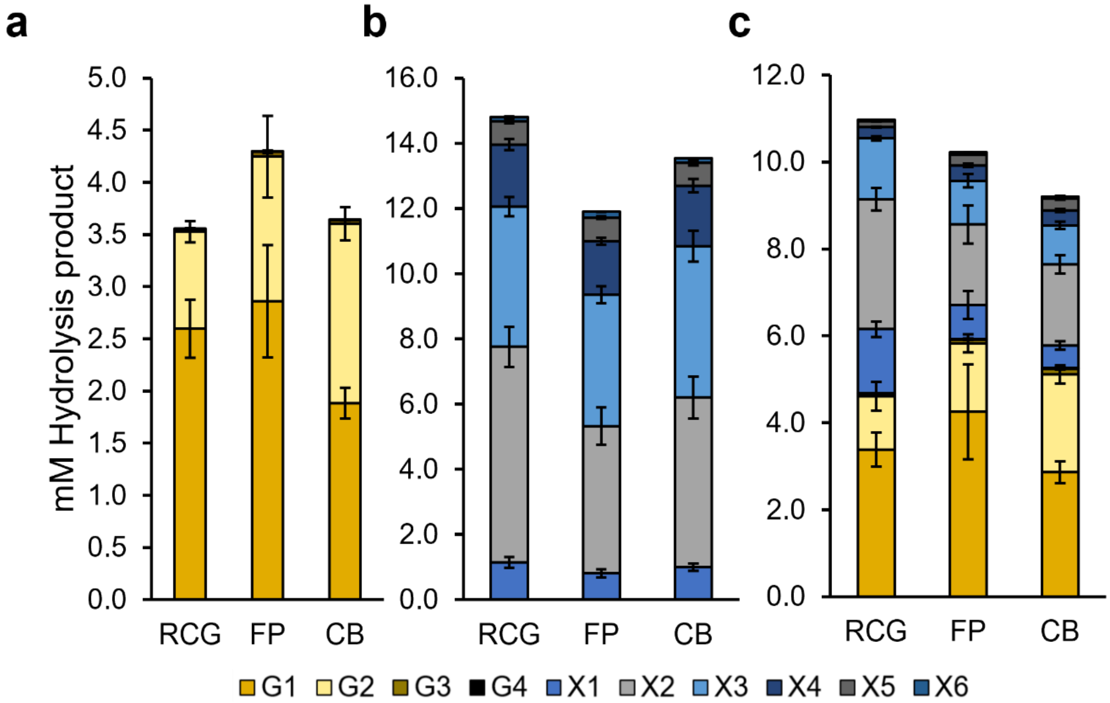


Figure 3.9. Cellulosomes with varied enzyme content produce distinct hydrolysis product distributions. RCG, FP, and CB represent cellulosome samples purified from N. californiae grown on reed canary grass, filter paper, and cellobiose respectively. Plotted are endpoint concentrations of each measured hydrolysis product from Avicel hydrolysis (a), beechwood xylan hydrolysis (b), and ionic liquid pretreated switchgrass hydrolysis (c). Data are presented as the mean ± SEM of N=9 (technical triplicates of three biological replicates) for a) and b) and N=3 (single measurement of three biological replicates) for c). In all panels, G1, G2, G3, G4 = glucose, cellobiose, cellotriose, cellotetraose, and X1,X2,…,X6 = xylose, xylobiose, …, xylohexaose.

Comparing the raw G1 and G2 production rates indicates the lower G1/G2 ratio for CB cellulosomes comes from lower G1 production rates, not higher G2 production rates (data not shown). The composition data from Figure 3.3c, however, does not show a significant change in GH2/3 family enzymes, which contain many known β-glucosidases, to explain this. One possible explanation is that enzymes annotated as belonging to other families possess multiple enzyme activities including β-glucosidase; many of the GH families are known to encompass several types of (hemi)cellulase activity, such as cellobiohydrolase and β-glucosidase activities (cazy.org).

The relative composition differences among the three cellulosome preparations generally did not affect the endpoint xylan degradation product distributions, besides a lower xylobiose concentration in FP relative to RCG samples (p = 0.03) (Fig. 3.9b). What is

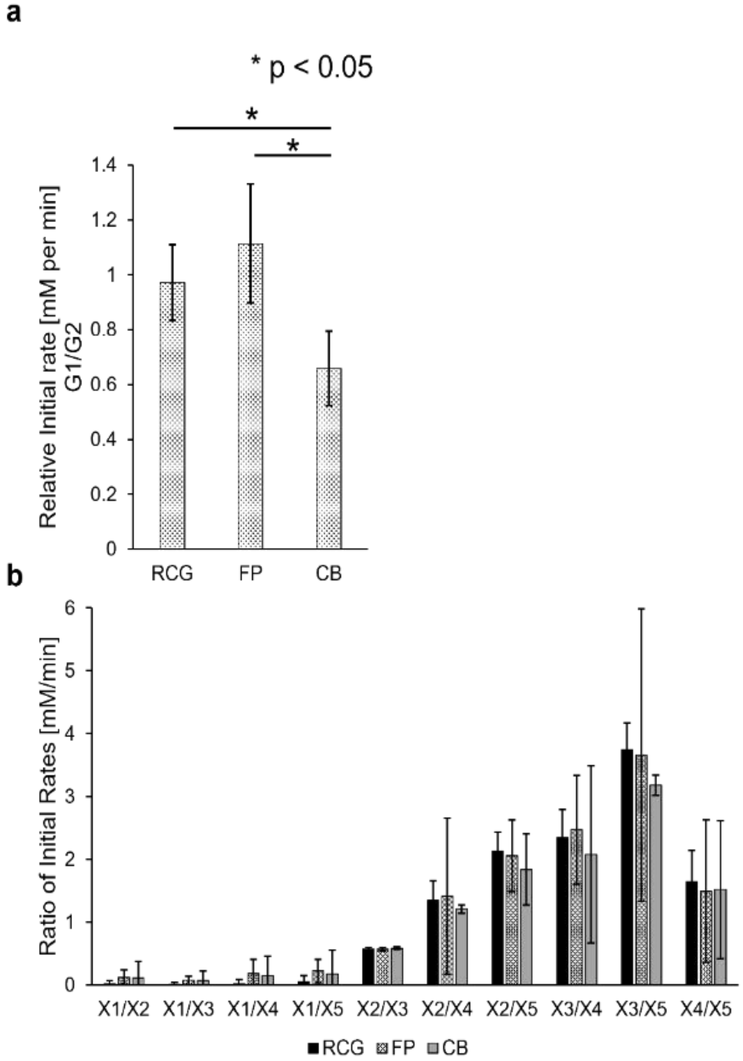


Figure 3.10. Hydrolysis product generation rates vary on Avicel but not on xylan. RCG, FP, and CB represent cellulosome samples purified from N. californiae grown on reed canary grass, filter paper, and cellobiose respectively. (a) Ratio of initial glucose production rate to cellobiose production rate (mM/min). Error bars represent one standard deviation of biological triplicates. (b) Initial production rate ratios for xylose (X1) and other xylooligomers. All rates were estimated from molar concentration vs time data in units of mM/min. Data are presented as mean ± one standard deviation for N=3 biological replicates.

surprising, however, is that fungal cellulosomes from *N. californiae* accumulated much higher concentrations of xylooligomers, especially xylobiose (X2) and xylotriose (X3), relative to xylose (X1) during xylan degradation. Relative initial rate ratios for xylan hydrolysis products were broadly consistent across cellulosome samples, with xylotriose as the fastest accumulating product for all cellulosome preparations (Fig. 3.10b). These findings are inconsistent with the observed changes in relative cellulosome xylanase content, which we would have expected to produce different xylan hydrolysis activity patterns.

After IL-SG hydrolysis, the major cellulose and hemicellulose breakdown products across cellulosome samples were glucose and xylobiose (Fig. 3.9c). Due to the high biological variabilty in hydrolysis yield of cellulosome samples isolated from the same growth conditions, few comparisons of endpoint product concentrations met accepted statistical significance benchmarks for hypothesis testing. However, comparing relative breakdown product concentrations (mole fractions) illuminated biochemical differences among the different cellulosome samples.

Relative glucose content was lower for CB and RCG vs FP cellulosomes ($p = 0.03$ and $p = 0.02$ respectively) and cellobiose content was higher for CB relative to RCG and FP cellulosomes ($p = 0.02$ and $p = 0.04$ respectively), again suggesting a relative lack of β-glucosidase activity in CB cellulosomes. Xylose content was also lower in CB and FP cellulosome product pools compared to RCG ($p = 0.008$ and $p = 0.056$ respectively), indicative of less complete switchgrass hemicellulose hydrolysis. A more granular analysis of hemicellulase enzyme family content within the different cellulosome samples suggests a ~7.5x lower GH43 composition in CB and FP cellulosomes relative to RCG cellulosomes may explain this result (Table A2); GH43's are annotated as β-xylosidases, which cleave xylobiose into two xylose units (cazy.org).

The large variability observed for biological replicates of the same cellulosome samples makes statistically robust correlation of changes in activity metrics to changes in enzyme composition challenging. However, assuming the cellulosome biological replicates we analyzed reflect the behavior of those kinds of cellulosomes in general, we analyzed how mean hydrolysis rate and yield metrics during the breakdown of IL-SG correlated with cellulosome composition (Fig. 3.11). As expected, glucan hydrolysis yield and xylan hydrolysis yield are strongly and positively correlated with the cellulose-acting and hemicellulose-acting enzyme content of fungal cellulosomes respectively. Exocellulase content was highly correlated with the initial G2 production rate, consistent with the fact that exocellulases produce cellobiose as their product. Endoglucanase and β-glucosidase content were highly correlated with initial glucose production rate. No similar types of connections between xylan hydrolysis rate metrics and composition of particular hemicellulases.

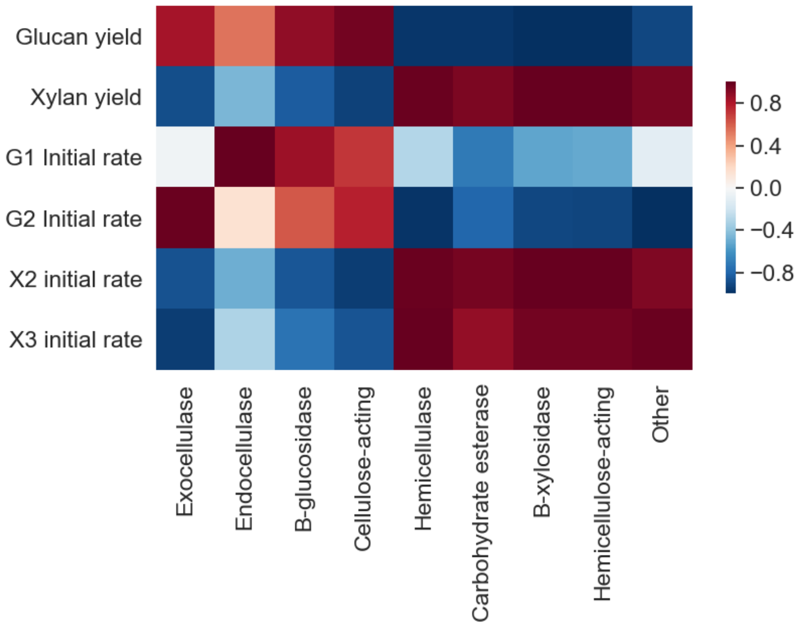


Figure 3.11. Correlation matrix of IL-SG hydrolysis rate and yield metrics with cellulosome enzyme content. The heat map scale represents the Pearson correlation between the two quantities labeling each row and column. Here G1 and G2 represent glucose and cellobiose, and X2 and X3 represent xylobiose and xylotriose, respectively.

Cryo-electron tomography resolves a fungal cellulosome complex

A longstanding goal of the anaerobic fungal community has been to solve the structure of a fungal cellulosome, from which to learn how these molecular machines assemble and function. Their heterogeneous, flexible nature and proclivity for sticking to polymeric substrates like filters makes native fungal cellulosomes difficult to work with and challenging to image with any structural biology technique. After many rounds of optimizing sample and grid preparation conditions, we observed by cryo-electron microscopy (cryo-EM) native fungal cellulosomes isolated from *N. californiae* grown on reed canary grass, revealing species of diverse morphologies resembling clumps of "grapes on a vine" (Fig. 3.12a-b). No distinct 2D classes could be obtained even after several rounds of 2D classification, indicating the heterogeneous nature of the complexes. We hypothesized that adding nanocellulose to purified cellulosomes may recruit complexes to the cellulose interface and facilitate imaging. Crystalline nanocellulose of defined length (10-20 µm) were frozen at 0.1 - 0.2 % concentration to visualize individual fibers of cellulose. The motion corrected images show thin flat fibers of cellulose with clear background (Fig. 3.12c-d). To template cellulosome complexes onto nanocellulose fibers, nanocellulose solution was incubated with RCG cellulosomes before grid deposition. Subsequent imaging showed cellulosomes decorating the nanocellulose at several places, even "crosslinking" separate cellulose fibers together (Fig. 3.12e-f). A zone of clearance was seen around the fibrils which indicated that the cellulosomes had indeed been recruited to the cellulose interface and not merely juxtaposed.

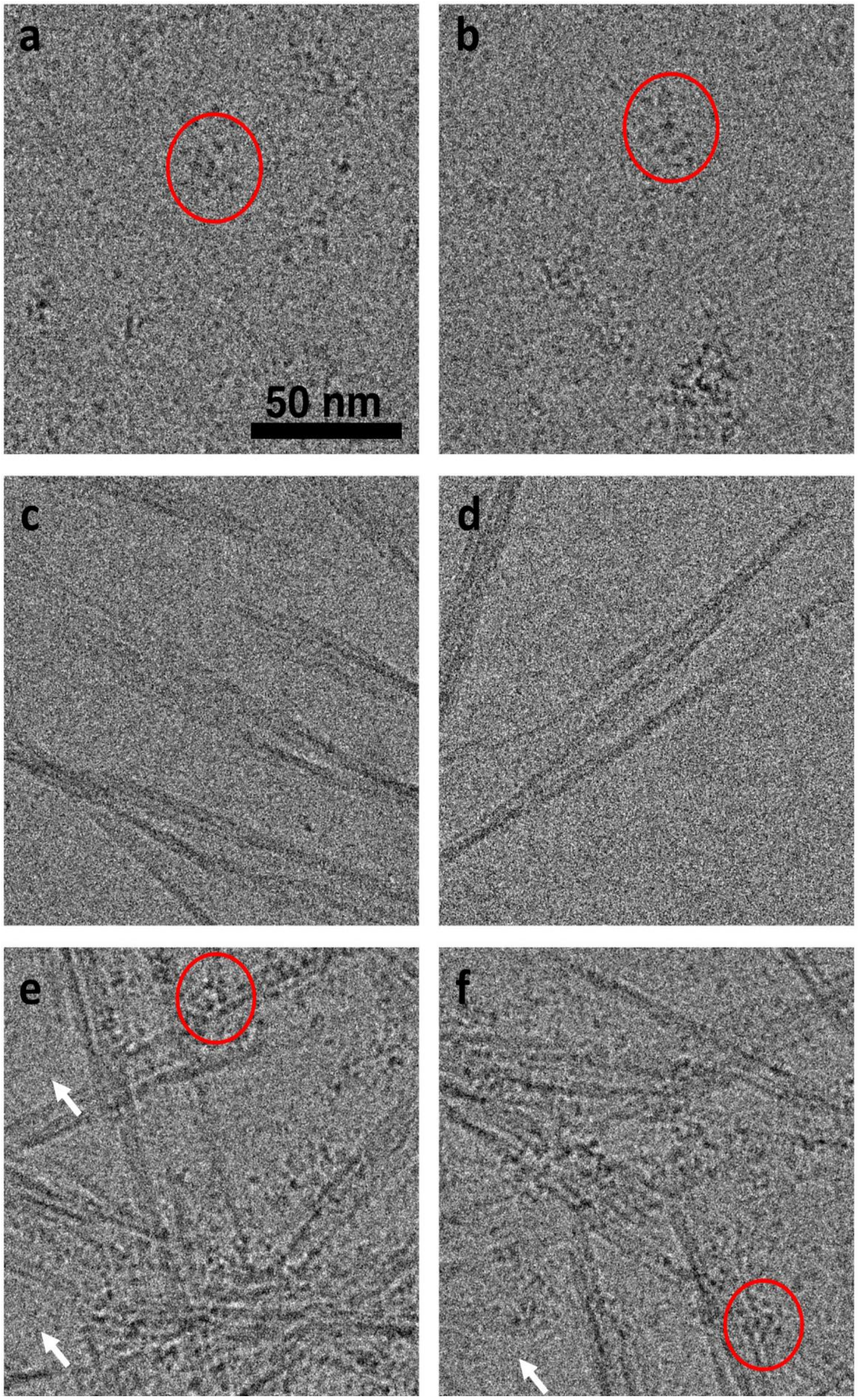


Figure 3.12. Cryo-EM images of cellulosomes recruited to nanocellulose substrate visualized – A) and B) Cryo-EM images of RCG cellulose show putative heterogenous species of various oligomeric arrangements like a grapes on a bunch arrangement (red circle) and D) Cryo-EM images of crystalline nanocellulose (control) E) and F) Cellulosomes seen decorating the nanocellulose fibers when incubated together indicating (red circles). A zone of clearance is seen around the nanocellulose free of any other cellulosome species (white arrow) (Scalebar = 50 nm for all images).

We applied cryo-electron tomography (cryo-ET) as an alternative technique to generate a cellulosome structure despite the heterogeneity of complex structures apparent in the sample. From a ± 60° tilt series of a grid containing cellulosome and nanocellulose, we observed several distinct, multidomain cellulosomes attached to cellulose fibers (Fig. 3.13a-b). A representative complex was selected, and the corresponding volume was extracted for further processing (Fig. 3.13c). The volume was low pass filtered to 40 Å and contrast inverted per standard practice for better visualization in ChimeraX. The visualized complex consisted of several globular domains roughly 6-8 nm in size and was clearly attached to the fiber at one end, indicating recruitment and binding rather than juxtaposition of the complex to the substrate (Fig. 3.13d-e). At least two proteins with different shapes, including one containing a visible structure resembling a fungal dockerin, are distinguishable in the tomogram, demonstrating the incorporation of different enzymes into a single complex.

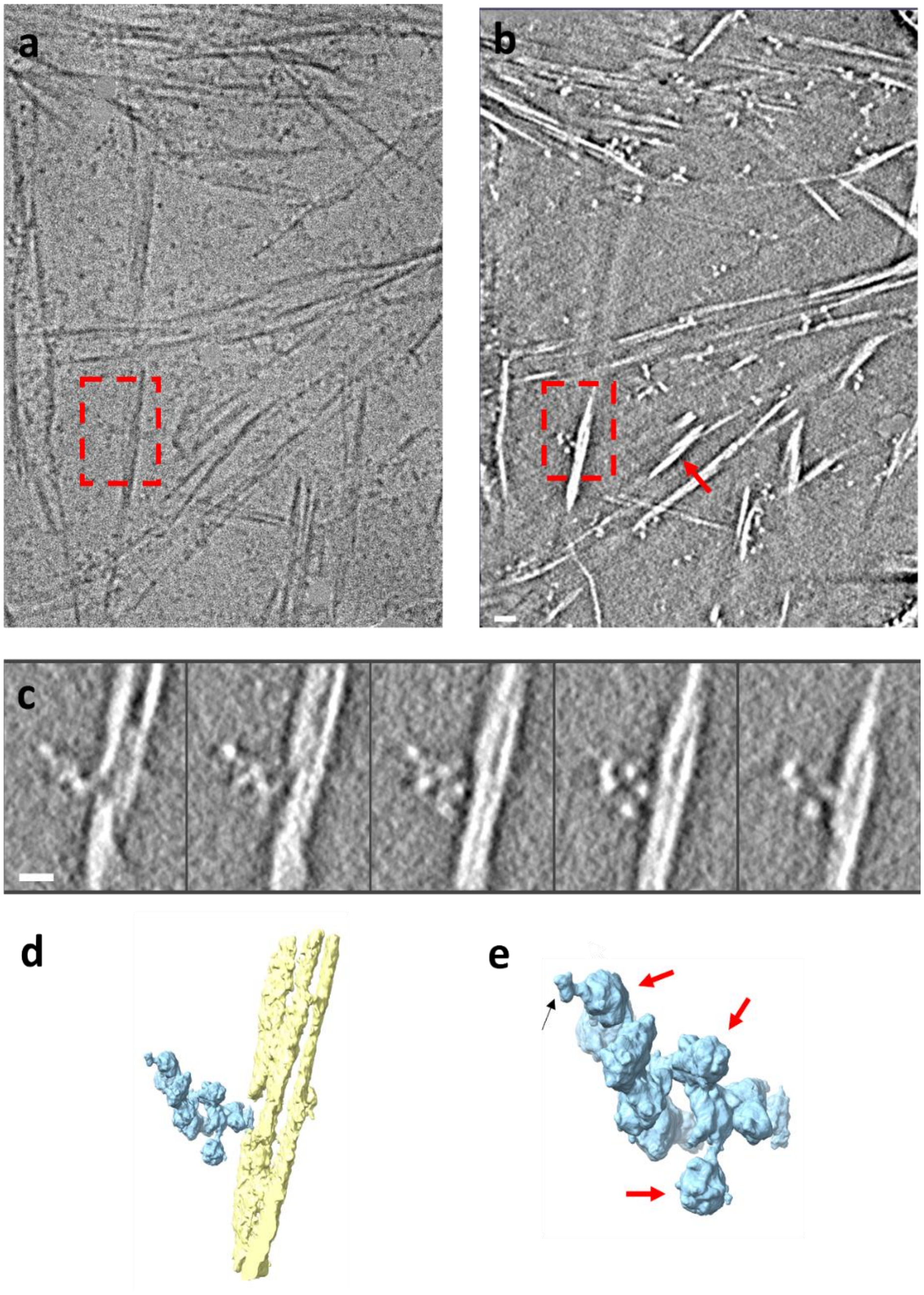


Figure 3.13. Cryo-ET analysis of RCG cellulosome complex attached to cellulose nanofiber. A) A 0° tilt image from the tilt series collected on RCG cellulosome-cellulose complex. Dashed red boundary indicates the region of interest. B) Virtual slice through the corresponding reconstructed 3d volume with inverted contrast, showing several long thin, flat cellulose fibers (red arrows). The dashed red boundary highlights an isolated RCG cellulosomes complex attached to the cellulose fiber. (scale bar = 20 nm) C) A multiple z slices through the area highlighted in (B), where densities of the cellulose and cellulosomes complex are inverted to white (scale bar = 10 nm) . D) A volume of 700x700x250 pixel was extracted and rendered in ChimeraX to visualize the interaction between the cellulose fiber (yellow) and cellulosomes complex (cyan). E) A zoomed in view of the multiple domains of the RCG cellulosome complex (annotated with red arrows) including one with a visible structure resembling a fungal dockerin (black arrow).

3.3 Conclusions

The major goal of this work was to characterize purified fungal cellulosomes to better understand their composition, structure, and biochemical activity. We first demonstrated a method that reliably produces purified cellulosome complexes from the fungus *Neocallimastix californiae* with high yield and purity. The same method did not produce clear high MW species when used to purify cellulosomes from anaerobic fungi of the genera *Piromyces* and *Anaeromyces*. Previous work on *Piromyces* sp. E2 shows a clear high MW complex could be purified from concentrated culture supernatant when the organism was grown on defined M2 medium [101], and another group found that supernatants from *Neocallimastix frontalis* cultures only contained a high MW peak visible by SEC when the fungus was grown in defined M2 medium [100], so cellulosome isolation from the culture supernatant could be growth media-dependent for the other species we investigated. Alternatively, anaerobic fungi may produce both secreted and cell-anchored cellulosomes of which we are only harvesting the secreted fraction; perhaps *Neocallimastix californiae* secretes cellulosomes at higher concentrations than *Piromyces finnis* or *Anaeromyces robustus*. Why our method only works on one species remains a challenging question to answer without a clear understanding of cellulosome assembly and secretion mechanisms.

This method produces cellulosome complexes with similar composition across independent purification trials, as visualized by SDS-PAGE and quantified by shotgun proteomics, making it a useful workhorse for continued investigation of *N. californiae* cellulosome structure and function. From these data, we resolved, at the highest resolution yet, the composition of *N. californiae* cellulosome complexes, and by repeating this method with cellulosomes purified from *N. californiae* grown on different substrates, we measured how complex composition changed with growth condition. *N. californiae* cellulosomes possess a core suite of exocellulases (GH48, GH6), endocellulases (GH5, GH45, GH9), and expansins, with β-glucosidases (GH2/GH3) that enable these complexes to efficiently hydrolyze cellulose and produce glucose as the major product. The abundance of expansins and incorporation of β-glucosidases differentiate these cellulosomes from bacterial systems characterized to date. Characterizing individual members of these major fungal cellulosome families via heterologous expression will illuminate how these individual enzyme families contribute to lignocellulose hydrolysis, providing valuable, mechanistic insight into fungal cellulosome activity.

Five enzyme families experience composition changes between any two of the three growth conditions, with the largest magnitude changes occurring for GH26 mannanases, hemicellulases, and carbohydrate esterases, which also undergo large changes in gene expression as quantified by RNA-seq. Changes in cellulosome composition did not always agree with changes in gene expression however, which may have several explanations - 1) post-transcriptional mechanisms control enzyme production under different growth conditions, 2) post-translational mechanisms control enzyme incorporation into complexes (e.g. varied cellulosome binding thermodynamics among enzyme families or steric restrictions on complex composition), or 3) our cellulosome purification method biases for particular complexes. Characterization of mini bacterial cellulosomes with trivalent scaffoldins and three different dockerin-fused enzymes uncovered that only 3/10 possible unique assembly states are observed, indicating sterics and other physical forces drive random cellulosome assembly towards complexes with specific compositions [79]. Similar mechanisms of discrimination likely affected the cellulosome composition observations in this work, but interrogating this directly remains out of reach until fungal cohesin's identity is revealed, enabling heterologous reconstitution of fungal cellulosomes *in vitro*.

To characterize the lignocellulolytic activity of fungal cellulosome complexes, we measured breakdown products over time during hydrolysis of several substrates. The activity data overall show that *N. californiae* cellulosomes are effective at hydrolyzing lignocellulose, especially hemicellulose. Simultaneously measuring many breakdown products highlighted that cellulose and hemicellulose degradation by fungal cellulosomes seems to proceed differently. Larger cellooligomers hardly accumulate during Avicel hydrolysis, suggesting cellobiohydrolase/exoglucanase and β-glucosidase activities dominate. Our observation that β-glucosidases constitute only 1-2 mol % of fungal cellulosome content implies these enzymes have very high specific activities, in line with prior work [62]. Quantifying breakdown product distributions and relative product generation rates highlighted that β-glucosidase activity is lacking in cellobiose-generated cellulosomes compared to filter paper and reed canary grass. However, no obvious change in GH2/GH3 content explains this, suggesting that glucose-producing enzyme activity is present in other enzyme families.

During xylan hydrolysis, larger xylooligomers up to xyloheptaose were detected and xylotriose and xylobiose accumulated much more rapidly and to much higher concentrations than xylose, suggesting endoxylanases dominate. This was less pronounced during switchgrass hydrolysis, likely because switchgrass hemicellulose backbone polymers are more diverse and substituted compared to beechwood xylan, which resembles a homopolymer of xylose (Megazyme part no. P-XYLNBE). Mostly incomplete conversion of xylooligomers to xylose by cellulosomes begs the question of whether xylobiase activity is mainly contributed by a freely diffusing enzyme or if larger xylooligomers are transported into anaerobic fungal cells for metabolism.

Switchgrass hydrolysis was the only case for which differential activity profiles and composition measurements among the three cellulosome samples aligned. CB cellulosomes freed less xylose and less xylooligomer overall from switchgrass hemicellulose compared to RCG cellulosomes, consistent with observed decreases in overall hemicellulase and carbohydrate esterase content. We hypothesize these composition changes produce activity differences on switchgrass but not purified xylan because the diversity of chemical bonds in switchgrass hemicellulose necessitates the presence of diverse hemicellulases and carbohydrate esterases, while purified xylan only requires exo- and endo-xylanases for hydrolysis. While total hemicellulase and CE content increases ~2x in RCG vs CB cellulosomes, endoxylanase content only increases by 48% (Table A2).

While our results clearly support the importance of incorporating hemicellulases and carbohydrate esterases to improve lignocellulose hydrolysis, we were unable to derive any composition-activity relationships that shed light on the contribution that major families like GH48, GH5, GH9, and expansins make to lignocellulose degradation by fungal cellulosomes. Altering growth conditions did not change the cellulosomal composition of these families enough and no differences in cellulose hydrolysis capacity could be tied to composition. Future work characterizing the individual cellulosome components in isolation to identify the most active ones will be key to developing fungal cellulosomes into bioprocessing tools. Of particular interest are enzymes from GH48, universally the major cellulosome component, Expansins, a relatively rare cellulosome member among cellulosome-producing microbes, enzymes from GH26, which were highly enriched in cellobiose cellulosomes, and multi-domain GH3+GH6 enzymes. Ideally, fungal cellulosomes could be reconstituted recombinantly *in vitro* to understand synergistic interactions among cellulosome components and derive composition-activity relationships to guide future cellulosome engineering, but this cannot be done without knowledge of the core domains mediating cellulosome assembly.

Technical challenges isolating fungal cellulosome complexes with sufficient purity and optimizing cryo-EM grid preparation, as well as microscope technology, were significant barriers to obtaining structural information about fungal cellulosomes in the past. Cryo-EM and cryo-ET analyses illustrated for the first time a “grapes on a vine” morphology for native cellulosome complexes from *N. californiae* similar to cellulosomes produced by anaerobic bacteria [144]. Reconstructed volumes from cryo-ET show the incorporation of distinct proteins, likely dockerin-containing enzymes from different families, into single complexes, the first real evidence that fungal cellulosomes resemble the models we use to represent them. No clear set of interacting structures is discernable from our cryo-ET model, but fitting predicted atomistic structures of proteins from the presented proteomics data to densities in the cellulosome reconstruction may narrow the set of potential protein-protein interactions facilitating fungal cellulosome assembly. At the least, the structural biology methodology presented here represents a critical first step towards elucidating the details of fungal cellulosome assembly through structural biology techniques. Overall, these results provide much new insight into how cellulosomes from *Neocallimastix californiae* degrade lignocellulose and provide a blueprint for translating these molecular machines into useful technologies for bioprocessing.

Overall, these results provide much new insight into how cellulosomes from *Neocallimastix californiae* degrade lignocellulose and provide a blueprint for translating these molecular machines into useful technologies for bioprocessing.

3.4 Materials and Methods

Statistical tests

If not explicitly stated, statistical comparisons were performed using an unpaired, two-sided Student's t-test of biological triplicate data assuming unequal variance.

Culturing anaerobic fungi

Neocallimastix californiae was grown in sealed glass bottles or Hungate tubes in medium C [117] supplemented with carbon source (1% w/v) under a 100% CO_2 headspace. For routine passaging, growing cultures were diluted 1:10 into Hungate tubes containing fresh media and 1% (w/v) reed canary grass every 4-5 days. *N. californiae* was passaged three times into medium C with a new carbon source before inoculating a large scale culture for cellulosome harvesting to ensure culture phenotype matched the growth condition.

For cellulosome harvesting, 300 – 1000 mL of culture with 1% (w/v) substrate in sealed septum bottles was inoculated with a preculture 1/10th the final volume and the headspace brought to 0 psig. Culture growth was monitored by measuring accumulated headspace pressure [145], and headspace gas was evacuated to 0 psig every 24 hours for 6-8 days.

Cellulosome harvesting and purification

6-8 day old cultures were transferred to centrifuge bottles or Falcon tubes and centrifuged at 3,000x g for three minutes in a fixed angle rotor to pellet cells and residual substrate. The supernatant was then transferred to new centrifuge bottles and the pH adjusted to 7.0 with 9N NaOH. Secreted cellulase concentration was estimated by Bradford assay (ThermoFisher Scientific cat. 23238), subtracting out the signal from uninoculated media with the same substrate. Cellulosomes were then purified from fungal supernatant using an adapted version of the affinity digest protocol developed for purifying cellulosomes from *Clostridium thermocellum* [126].

Phosphoric acid swollen cellulose (PASC), prepared as described in [126], was added to pH-adjusted fungal supernatant at a concentration of 4 mg PASC (dry weight) per mg cellulase. The mixture was incubated on a rocking platform at 4°C for one hour to adsorb cellulosomes and other cellulose-binding proteins out of solution. PASC and adsorbed proteins were pelleted by centrifugation at 12,000x g for 10 minutes in a fixed angle rotor and the supernatant was discarded. The resulting pellet was resuspended in 20 mL of dialysis buffer (10 mM MES buffer pH 6.4) and transferred to a 30 mL 10 kDa MWCO Slide-A-lyzer dialysis cassette (ThermoFisher Scientific cat. 66830) for dialysis at 39°C against 10 mM MES buffer pH 6.2. Dialysis buffer was changed approximately every hour until the PASC was completely solubilized (~4-8 hours). The dialyzed solution was further clarified by centrifugation at 12,000x g for 10 minutes to pellet remaining insolubles, and the clarified solution was concentrated using a 300kDa MWCO PES filter (Sartorius cat. VS2051) by repeated centrifugation at 4,000x g and 4°C in a swing bucket rotor until the concentration rate appeared to slow considerably. Concentrated sample was aliquoted and snap-frozen in a metal block for long-term storage at -80°C.

Cellulosome analysis by protein gel electrophoresis

SDS-PAGE: 25 μg of cellulosome protein in 1x Laemmli SDS-PAGE sample buffer was incubated at 90°C for 30 minutes prior to loading on a 8% Tris-Glycine SDS-PAGE gel.

Electrophoresis was performed at 110V for 80 minutes, followed by SYPRO Ruby (Bio-Rad Laboratories) staining following the manufacturer's instructions.

Native PAGE: SEC fractions containing the cellulosome peak were pooled and concentrated approximately 40-fold and 20 µL was mixed with 20 µL Laemmli sample buffer without SDS or reducing agent. The mixture was then loaded on a 3-8% Tris-acetate gel (Invitrogen). Electrophoresis was performed with the gel tank on ice in a 4°C cold room for 2.5 hours, 160V. Bands were visualized by Coomassie staining.

Size exclusion chromatography

500 µL of affinity digest-purified cellulosome sample was loaded on a BioLogic DuoFlow chromatography system (Bio-Rad Laboratories, Hercules, CA) equipped with a Superose 6 Increase 10/300 GL column (Cytiva Life Sciences) kept at 4°C. PBS pH 7.4 was used as mobile phase. 1 mL fractions were collected and pooled to capture the cellulosome peak before concentration with a 0.5 mL 30 kDa MWCO PES filter (ThermoFisher Scientific cat. 88502). Concentrated samples were brought to 50% glycerol for storage at -20°C or 0% glycerol for storage at -80°C. Sample protein concentrations were measured in technical triplicate using the BCA assay with BSA as standard (ThermoFisher Scientific cat. 23225). Routine sample characterization was performed by SDS-PAGE and/or Native PAGE as described above.

Mass spectrometry proteomics sample preparation

Urea and Dithiothreitol (DTT) were added to cellulosome samples to a final concentration of 8M and 5 mM respectively before samples were incubated at 37°C for 1 hour to fully denature sample proteins. Iodoacetamide was then added to a final concentration of 40 mM and the samples were incubated as before, but in darkness. Samples were diluted eight-fold with 1.14 mM $CaCl_2$ in 50 mM ammonium bicarbonate and trypsin was added in a 1:50 (trypsin:protein, w:w) ratio. Trypsin digestion was performed for 3 hours at 37°C with shaking at 850 rpm before samples were frozen and stored at -70C. Solid Phase Extraction (SPE) was performed on the samples using Ultra Microspin C18 columns from The Nest Group (Ipswich, MA). Columns were conditioned with 200 uL of methanol followed by 200 uL of 0.1% trifluoroacetic acid (TFA) in water. The samples were acidified to 0.1% TFA and then applied to each column followed by 400 uL of 95:4.9:0.1 H2O:MeCN:TFA. Samples were eluted into 1.5-mL microcentrifuge tubes with 80:19.9:0.1 MeCN:H2O:TFA and concentrated in a vacuum concentrator to 30 uL. Peptide concentrations were measured by BCA assay (ThermoFisher Scientific) and peptide samples were diluted to 0.1 µg/ µL prior to LC-MS analysis.

Mass spectrometry proteomics data collection and analysis

A Waters nano-Acquity dual pumping UPLC system (Milford, MA) was configured for on-line trapping of a 5 µL injection at 5 µL/min for 5 min with reverse-flow elution onto the analytical column at 200 nL/min using 0.1% formic acid in water as mobile phase (mobile phase A). The analytical column was slurry packed in-house using Waters BEH C18 1.7 um particles (Waters Chromatography, Drinagh, Ireland) into a 25 cm PicoFrit 360um od x 75um id column (New Objective, Littleton, MA) and held at 45°C during use by a 15 cm AgileSleeve flexible capillary heater (Analytical Sales and Services, Inc., Flanders, NJ). The trapping column was slurry packed in-house using Jupiter 5µm C18 particles (Phenomenex, Torrance, CA) into a 4cm long piece of 360µm od x 150µm id fused silica (Polymicro Technologies, Phoenix, AZ) and solgel fritted at both ends. Mobile phases consisted of (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile with the following gradient profile (min, %B): 0, 1; 10, 8; 105, 25; 115, 35; 120, 75; 123, 95; 129, 95; 130, 50; 132, 95; 138, 95; 140, 1.

MS analysis was performed with a Fusion Lumos mass spectrometer (Thermo Scientific, San Jose, CA) outfitted with a modified Nanospray Flex ion source (Thermo Scientific, San Jose, CA). The ion transfer tube temperature and spray voltage were 250°C and 2.2 kV, respectively. Data were collected for 120 min following a 27 min delay from sample injection. FT-MS spectra were acquired from 300-1800 m/z at a resolution of 120k (AGC target 4e5). FT-MS/MS spectra were acquired using HCD (collision energy 32) in data dependent acquisition mode for 3 seconds at an isolation width of m/z 0.7 using a first fixed mass of m/z 110 and a resolution of 50k (AGC target 1.25e5).

Mass spectrometry data were analyzed with MSGF+ using a target-decoy approach [146] and a protein database consisting of common contaminant proteins, Mycocosm-verified dockerin-containing proteins from the *N. californiae* genome (also called Catalog genes), and annotated scaffoldins from the *N. californiae* genome. Spectrum-peptide mapping used a 20 ppm parent mass tolerance, trypsin-specific digestion, allowing unlimited missed cleavages, and a variable post-translational modification of oxidized methionine and acetylated protein N-termini, with up to 2 post-translational modifications per peptide. The resulting data from MSGF+ was filtered and reformatted using MzidToTsvConverter (github.com/PNNL-Comp-Mass-Spec/Mzid-To-Tsv-Converter/) with an E-value cutoff of 0.01. Results were further filtered by a custom Python script using the MSGF+ reported Q-value to produce a final dataset of peptide-spectrum matches (PSMs) with false discovery rate <1%.

Protein sequence coverage and relative protein abundance calculated as the normalized spectral abundance factor (NSAF) [130] were computed from the filtered data using a custom Python script (https://github.com/slillington/proteomics-data-analysis) and Microsoft Excel. Coverage was defined as the % of amino acids in a database protein sequence detected in peptides present in a MS dataset. Database proteins were grouped by their domain annotations (e.g. presence of a particular catalytic domain like GH48) to measure changes in cellulosome enzyme family composition.

RNA-seq data re-analysis

The details of RNA-seq data collection are detailed in reference [128]. To make the data consistent with the proteomics data acquired in this work, annotated transcripts from the full RNA-seq dataset were filtered to only contain those that were dockerin-containing. A pivot table in Excel was used to sum fragments per kilobase million (fpkm) values for dockerin-containing transcripts by enzyme family (e.g. those containing a GH48 domain) for comparison to the proteomics mol% data.

Nanostructured Initiator Mass Spectrometry (NIMS) enzyme kinetics measurements

Hydrolysis reactions were performed in 50 µL 100 mM sodium acetate buffer pH 5.5 with 1 mg substrate and 2.5 µg cellulosome. Sample protein concentrations determined by triplicate BCA assay (ThermoFisher Scientific) measurements were as follows: RCG1 – 1.70 ± 0.1 mg/mL ,RCG2 – 0.80 ± 0.08 mg/mL, RCG3 – 2.12 ± 0.08 mg/mL, FP1 – 0.86 ± 0.09 mg/mL, FP2 – 0.70 ± 0.03 mg/mL, FP3 – 1.42 ± 0.08 mg/mL, CB1 – 1.25 ± 0.05 mg/mL, CB2 – 1.08 ± 0.05 mg/mL, CB3 – 1.41 ± 0.05 mg/mL. Avicel was purchased from Sigma Aldrich (part no. 11365) and beechwood xylan was purchased from Megazyme (part no. P-XYLNBE). Ionic liquid pretreated switchgrass was prepared in house at the Joint

Bioenergy Institute and the analyzed composition measured by Microbac Laboratories (Boulder, CO) was as follows: Glucan – 50.17 mass %, Xylan – 18.23 mass %, Lignin – 13.01 mass %, Galactan – 0.00 mass %, Arabinan – 4.77 mass %, Mannan – 0.00 mass %, Ethanol extractives – 2.3 mass %, Water extractable/others – 9.49 mass %.

During time course measurements, reducing sugars were derivatized for NIMS analysis by post-enzymatic reaction using oxime tagging techniques at time 0, 5 min, 10 min, 20 min, 30 min, 60 min, 2 h, 4 h, 6 h, 22 h, 46 h, 70 h [147]. [U]-13C glucose and [U]-13C xylose (Omicron Biochemicals) were included in the derivatization process and used as internal standards for oligosaccharide quantification. Synthesis of the *O*-alkyloxyamine fluorous-tagged NIMS probe and assays of oxime-tagging in the presence of internal standards were performed as described previously [147].

At each timepoint, a 2 µL aliquot of the reaction mixture was transferred into a Eppendorf vial (0.2 mL) containing 6 µL of 100 mM glycine acetate, pH 1.2, 1.0 µL of a 2.5 mM aqueous solution of [*U*]-^{13}C glucose and [*U*]-^{13}C xylose, 2 µL of CH_3CN, 1 µL of MeOH, 1 µL of the NIMS probe (100 mM in 1:1 (v/v) H_2O:MeOH), and 0.1 µL of aniline. The resulting mixture was incubated at room temperature for 16h. All NIMS analyses were performed on NIMS chips, fabricated as described previously [141,148] using a Bruker UltrafleXtreme MALDI TOF-TOF mass spectrometer (Bruker Daltonics, Bremen, Germany). A grid drawn manually on the NIMS chip using a diamond-tip scribe helped with spotting and identification of sample spots in the spectrometer. A 0.2 µL aliquot of the above oxime-NIMS solution was spotted onto the surface of the NIMS chip and removed after 30 s. Chips were loaded using a modified standard MALDI plate. FlexControl and FlexAnalysis were used for acquisition and data analysis. Spectra were recorded in positive reflector mode with a laser power of 20%. The instruments were calibrated using either the NIMS internal standards (oxime probe, oxime-tagged [U]-13C glucose, oxime tagged [U]-13C xylose) or Anaspec Peptide Calibration mixture 1 (Anaspec, Fremont, CA). Data were acquired by summing up 30000 laser shots in 2000 shot steps, sampling 15 regions locations per spot. Signal intensities were identified for the ions of the tagging products.

Hydrolysis product generation rates were quantified as the slope of a least squares linear fit to molar concentration time series data measured by NIMS. Hydrolysis yields were computed as the mass or sum of masses of hydrolysis products generated divided by the original mass of substrate.

Cryo-electron microscopy

Sample preparation for cryo-electron microscopy

Cellulosomes grown on reed canary grass as a substrate (RCG cellulosomes) were clarified by centrifuging at 17,000 g for 1h at 4°C to pellet the residual cellular debris. The sample was concentrated to ~500 uL in a 100 kDa MWCO Amicon spin concentrator. The sample was loaded onto an AKTA Pure FPLC system stored at 4°C using a Superose 6 Increase 10/300. Fractions corresponding to 1 mDa were concentrated using a 30 kDa MWCO Amicon filter; the sample purity was verified by SDS and Native PAGE and sample was used for cryo-EM analyses. Cellulose nanofibers (referred to as nanocellulose) defined length (10-20 µm) was purchased as a 6% (w/v) stock solution (Nanografi) and used at a final concentration of 0.1 – 0.2 % for cryo-EM. Grids were prepared by loading 3 µL of a ~0.1 to 0.2 % nanocellulose solution or RCG (1mg/ml) was loaded on Quantifoil grids (Q1/2, 300 mesh). To demonstrate RCG recruitment to the substrate, the nanocellulose solution (0.1-0.2% final concentration) was incubated with RCG cellulosomes for 20 mins at room temperature and then loaded on grids. Samples were vitrified by blotting for 3 – 3.5 s and plunge freezing in liquid ethane using a Leica EMGP2 and were stored under liquid nitrogen till further use.

Cryo-EM imaging analyses

Screening and data collection was performed by loading grids 300 keV Titan Krios G3i (Thermo Fisher). Movies were collected on a K3 direct electron detector using a 20 ev slit width on a Bioquantum energy filter (Gatan Inc). All datasets were collected using the standard EPU software. Movies were collected at 130,000x magnification in super resolution mode resulting in a pixel size of 0.3398 Å. An exposure of 1.4 – 1.65 s and a defocus range of −0.5 to −2.5 µm was used resulting in a total dose of 45 to 52 $e^{-}/Å^{2}$. All movies were processed using cryoSPARC Live and cryoSPARC [149]. Motion correction and CTF estimation were performed using default parameters. Images were binned 2x by Fourier cropping and visualized using Fiji [150].

Cryo-ET analyses

Using SerialEM [151], single axis tilt series were collected from -60° to + 60° in 3° increments at a magnification of 84,000 x resulting in a pixel size of 0.55 Å in super resolution mode. Using a target defocus of -3.5 µm, each tilt was recorded as movies with 2 sub frames using the same microscope-camera combination as the single particle data. Additional contrast was provided by using a Volta Phase Plate and an exposure of 0.24 s per movie. Motion correction was performed using *MotionCor2* [152]. Tilt series were binned 2x and tomograms were calculated using the SIRT function in IMOD [153]. Tomograms were further low pass filtered to help in visualization. A small volume of 700 x 700 x 250 voxels was extracted, and contrast inverted for better visualization of a representative cellulosome assembly bound to nanocellulose. All visualizations were done using ChimeraX [154].

IV. Stimuli-responsive protein complexes inspired by fungal cellulosomes

4.1 Introduction

Enzymes drive the complex network of chemical transformations that mediates all critical life processes. To enhance and direct the activity of enzyme systems, nature evolved strategies to spatiotemporally organize enzymes into multi-functional, ordered complexes (reviewed in [8,9,70]). Exemplifying its importance, enzyme complexation as a strategy for efficient multi-step chemical transformation has evolved in unique biological contexts, including natural product biosynthesis [71], biomass depolymerization [72], and signal transduction [73,74].

As it does for enzyme systems that naturally incorporate into complexes, enzyme colocalization can dramatically enhance product fluxes when heterologous enzyme sets are artificially colocalized, making it an attractive technology for use in industrial biocatalysis. Improved biocatalytic performance from enzyme colocalization has been demonstrated in a variety of contexts with different colocalization systems [8,9]. Designer cellulosomes assembled with bacterial dockerin and cohesin enhanced hydrolysis of waste biomass into sugars 10-fold [155,156]. Tetravalent complexes of mevalonate biosynthesis enzymes assembled using SH3 domain-ligand pairs increased mevalonate flux 77-fold [157]. When immobilized on a DNA scaffold, the glucose oxidase/horseradish peroxidase pair exhibits 48-fold higher activity [158]. While existing components have demonstrated many successes, there may not be a universal parts set for synthetic protein complex construction, as the mechanisms of rate enhancement are inconsistent and complex [10]. Additionally, certain enzymes may poorly tolerate specific fusion partners required to mediate assembly or do not function well in synthetic complexes with particular spatial architectures [11,12]. Thus, adding new parts to the toolkit for constructing synthetic protein complexes remains highly desirable for designing optimized biocatalytic enzyme systems for wide-ranging applications.

An emerging goal in synthetic biology is to design protein systems whose activity and self-assembly can be manipulated post-translationally. The performance and output of multi-enzyme complexes are dictated chiefly by the intrinsic kinetics, stoichiometry, and local spatial arrangement of their constituent proteins [155,157,159]. Indeed, many natural enzyme complexes post-translationally regulate complex self-assembly to dynamically direct or throttle fluxes through reaction networks, enabling cells to sense and respond to their environment [160–162]. As such, stimuli-responsive control over protein complex assembly, stoichiometry, and/or subunit arrangement presents a powerful mechanism for controlling the collective function of synthetic multi-protein systems (Fig. 4.1b,d), which has applications in metabolic engineering [163], drug development [164], and cell programming [159,165].

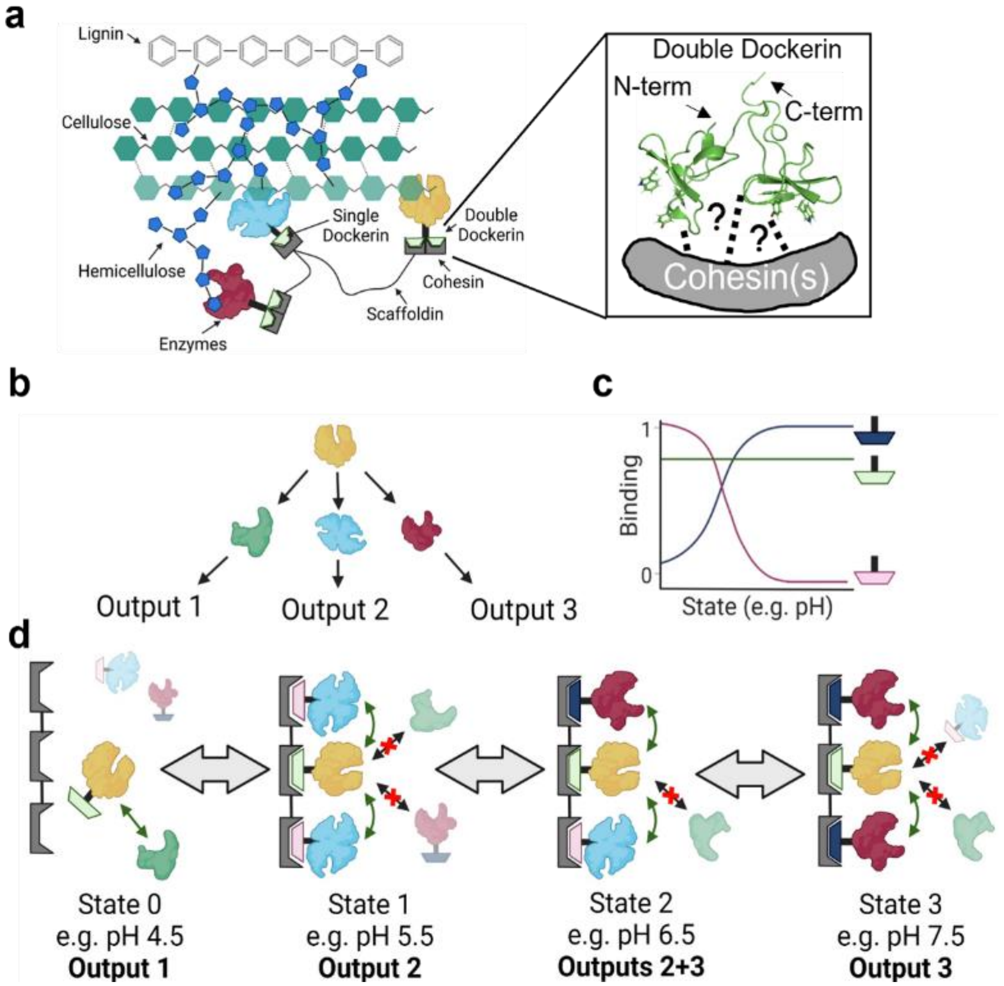


Figure 4.1. Fungal cellulosomes as a template for synthetic protein complexes with stimuli-controlled composition. A) Cellulosomes assemble diverse enzymes into complexes via modular parts, dockerin and cohesin, which are grafted onto many proteins to mediate their assembly. However, the identity of fungal cohesin is unknown to date, precluding reconstitution of fungal cellulosomes for characterization or engineering. B) Networks of enzymes performing sequential reactions mediate cellular metabolism and signaling and often contain branch points, in which each branch produces a different final product. C) The composition and functional output of protein complexes are directly related. One way to achieve stimuli-responsive composition uses a suite of modular interaction domains that exhibit different binding behaviors, such as opposite binding pH dependencies. D) Such a system in principle enables rapid, reversible control over the functional output of a multi-enzyme system by modulating an environmental stimulus like pH. Created with Biorender.com

The majority of modular nanotechnologies produce static assemblies not easily manipulated by environmental triggers [166–171]. Tools for synthetic, stimuli-responsive self-assembly have been developed to respond to environmental changes in pH [164], temperature [172], light [163,173], and the presence of specific molecules [174]. However, these approaches, which trigger two-state assembly/disassembly (an ON and OFF state), under-utilize the capacity of protein complexes to produce more than two outputs; even simple assemblies can produce >2 distinct outputs if different compositions dominate under different environmental states (Fig. 4.1b,d). The challenge not addressed by current stimuli-responsive self-assembly technologies is that such behavior requires programming protein-protein interactions that process the same environmental change in different ways (Fig. 4.1c). These new tools for building protein complexes that predictably and reversibly transition among unique assembly states as a function of the environment would greatly expand our ability to program sophisticated functionalities into protein systems for synthetic biology applications.

Cellulosomes produced by anaerobic fungi (*Neocallimastigomycota*) found in the guts of large herbivores present an attractive framework for multi-state, stimuli-responsive protein assembly that is unexplored for engineering applications (Fig. 4.1a). These complexes are exceptionally efficient at hydrolyzing biomass [103] but the nature of their biochemical activity is poorly understood because fungal cellulosomes cannot be reconstituted *in vitro*. Cellulosomes are thought to assemble via modular, noncovalent interactions between enzyme-fused dockerin domains and cohesin domains repeated on a central scaffoldin protein, as bacterial cellulosomes do [46], but the sequence and structure of fungal cohesin remain unknown [175,176]. Across anaerobic fungal genomes, dockerin sequences are fused to the N- or C-termini of 99 different Pfam protein families [83–85], and numerous fusions of fungal dockerins to proteins from other organisms facilitate incorporation of the functional protein into the cellulosome [86]. Furthermore, the dockerin fold accommodates high sequence diversity, implying high functional engineerability without disrupting binding activity [83–85]. Thus, fungal dockerin is an ideal part for synthetic protein assembly provided a suitable binding partner is discovered or engineered.

Here, we synthesized a synthetic cohesin analog, a nanobody that binds polypeptides encoding tandem dockerin domains (termed a double dockerin), with which to build synthetic protein complexes inspired by fungal cellulosomes. Using a combination of atomistic and coarse-grained molecular modeling approaches, we predict and validate a structure for the nanobody-dockerin complex and design a tethered nanobody “scaffoldin” capable of forming heterotrimers with dockerin-fused enzymes – the first reconstitution of dockerin-fused fungal enzyme complexes. Towards constructing synthetic fungal cellulosomes with stimuli-dependent composition, we apply yeast surface display to develop dockerin variants that conditionally bind nanobody in response to pH change, a relevant biological stimulus [177], and identify the hotspot histidines affecting this phenotype. Further testing of two variants in soluble form supports the conclusion that the evolved binding pH dependence is an intrinsic molecular property and not a surface display artifact. Overall, our findings showcase a new framework for synthetic enzyme complex formation with fungal cellulosome parts that has broad utility in enhancing and directing the activity of diverse biocatalytic cascades.

4.2 Results and Discussion

Since the sequence and structure of cohesin, fungal dockerin's native binding partner, remains unknown, we developed a synthetic dockerin binding partner to build protein complexes containing dockerin-containing proteins. A dockerin-specific nanobody was an attractive cohesin substitute given the small size, conserved structure, clear complementarity determining regions (CDRs) that participate in binding, and compatibility with heterologous expression that nanobodies exhibit [178]. With the goal of isolating a panel of nanobodies against different cellulosome epitopes, we raised nanobodies against purified cellulosomes isolated from *Neocallimastix californiae*. Epitope-containing proteins were identified by affinity chromatography using immobilized nanobody.

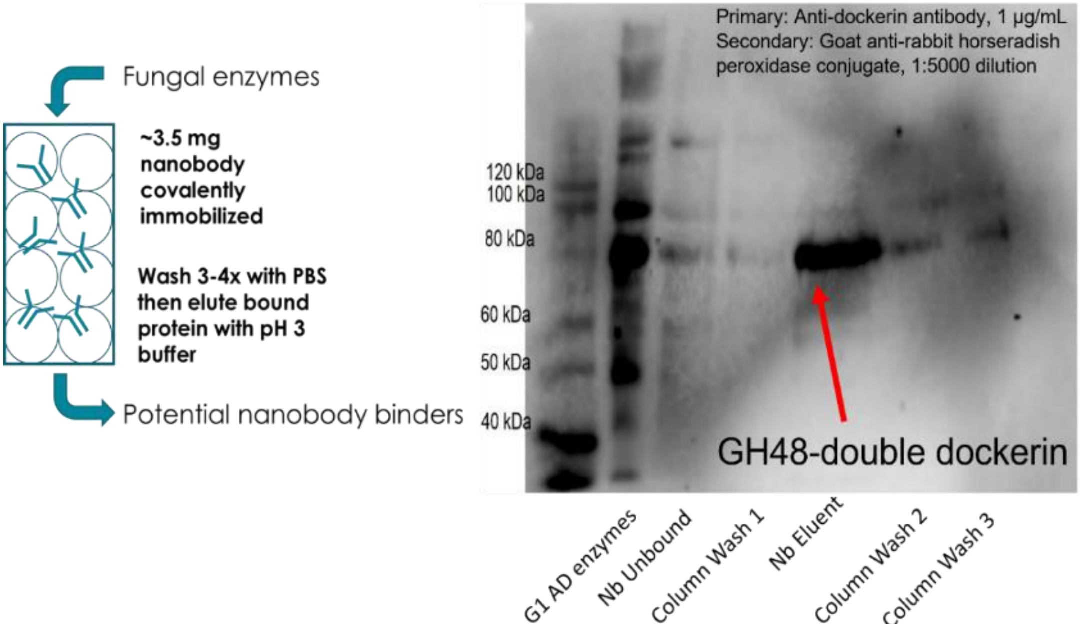

Fungal enzymes
~3.5 mg nanobody covalently immobilized
Wash 3-4x with PBS then elute bound protein with pH 3 buffer
Potential nanobody binders
Primary: Anti-dockerin antibody, 1 µg/mL
Secondary: Goat anti-rabbit horseradish peroxidase conjugate, 1:5000 dilution
120 kDa
100 kDa
80 kDa
60 kDa
50 kDa
40 kDa
GH48-double dockerin
G1 AD enzymes
Nb Unbound
Column Wash 1
Nb Eluent
Column Wash 2
Column Wash 3

Figure 4.2. Binding partners for a nanobody isolated from alpacas immunized against native cellulosome complexes were identified via affinity chromatography. (Left) Nanobody was immobilized on agarose beads in a column that was exposed to a mixture of cellulosome proteins from a Neocallimastix californiae culture. (Right) Samples from the starting enzyme mixture, washes, and the final elution were run on an SDS-PAGE gel and blotted with anti-double dockerin antibody. The strong band at ~80 kDa was previously identified as a double dockerin-containing GH48. G1 AD enzymes – a crude mixture of cellulosome proteins from a N. californiae culture; Nb unbound – column flowthrough of G1 AD enzymes; Column wash X – flowthrough after washing column with PBS; Nb Eluent – proteins that remained bound to the column throughout washing but were eluted by pH 3 buffer.

One nanobody, NbE6, appeared to bind the major cellulosome component Cel48A [131], an exocellulase with a double dockerin at its C-terminus (Fig. 4.2). Enzyme-linked immunosorbent assay (ELISA) binding affinity measurements with purified recombinant protein containing a double dockerin domain (doc5600) with an N-terminal TrxA tag confirmed NbE6 binds this double dockerin with a $K_d \cong 20\ nM$ (Fig. 4.3b). NbE6 did not bind recombinant protein containing only a single dockerin domain, and native mass spectrometry (MS) confirmed NbE6 binds double dockerin with 1:1 stoichiometry (Fig. 2b, Fig. A1).

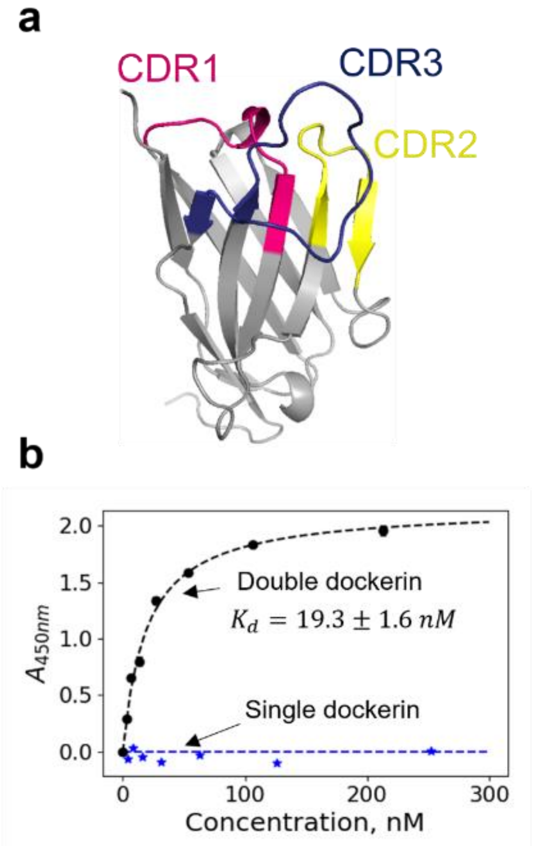


Figure 4.3. A double dockerin nanobody serves as a cohesin replacement with which to build synthetic cellulosomes. A) The RoseTTaFold structure for the nanobody contains the conserved nanobody structure with three complementarity determining regions (CDRs). B) ELISA between purified nanobody and recombinant double dockerin shows high affinity binding between domains. Purified single dockerin shows little nanobody binding activity. Error bars (smaller than the data points) represent SEM of n = 2. Uncertainty represents one standard deviation in the parameter estimate.

Without a co-crystal structure, we employed a multi-step protein docking and molecular simulation workflow, combined with experimental point mutagenesis, to map the interaction interface of NbE6 and double dockerin (Fig. 4.4a). Briefly, structure predictions for doc5600 and NbE6 served as the basis for protein-protein docking. Since the double dockerin is known to be very flexible [179], we generated alternative conformations that may better model its configuration upon binding NbE6 using temperature replica exchange molecular dynamics (REMD). Dominant doc5600 conformers were docked to NbE6 using ClusPro, a

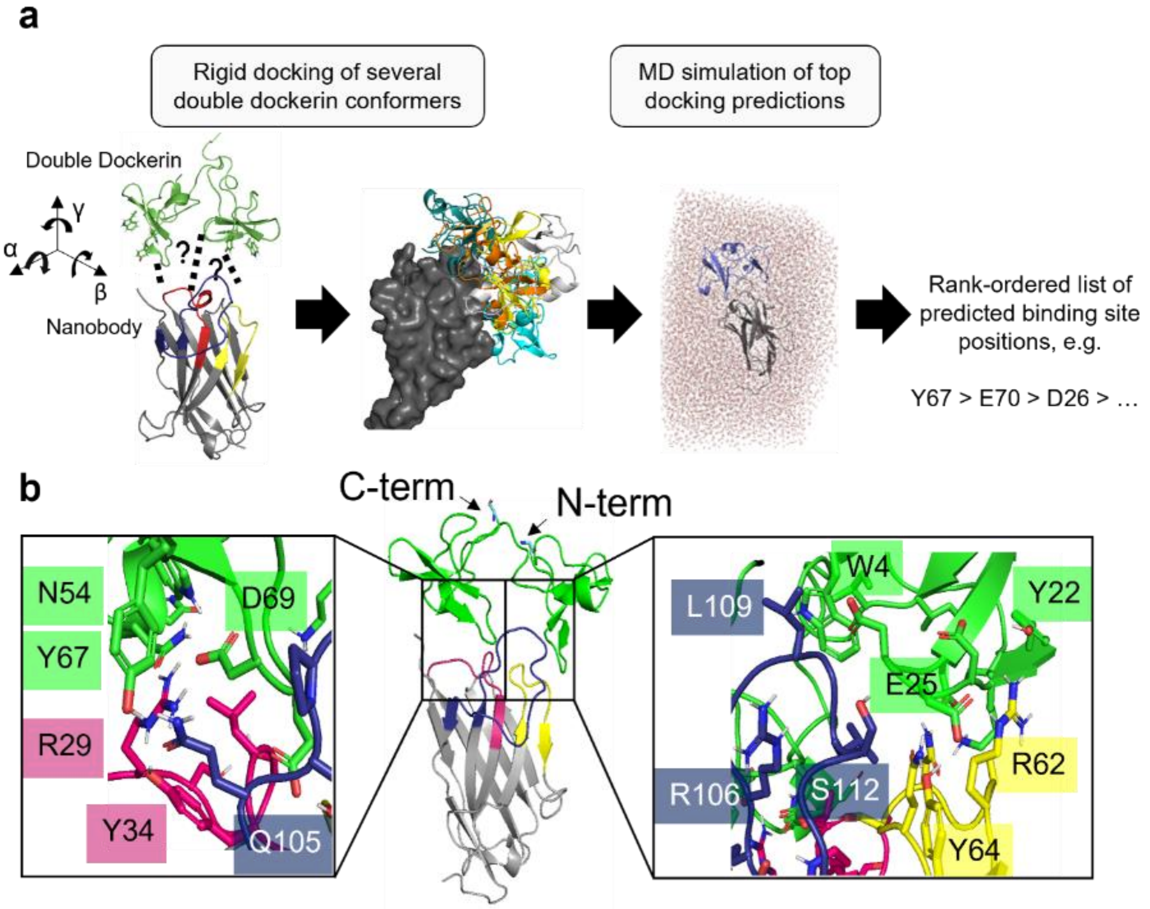

a
Rigid docking of several double dockerin conformers
MD simulation of top docking predictions
Double Dockerin
γ
α
β
Nanobody
?
?
Rank-ordered list of predicted binding site positions, e.g.
Y67 > E70 > D26 > …
b
C-term
N-term
N54
D69
Y67
R29
Y34
Q105
W4
L109
Y22
E25
R62
R106
S112
Y64

Figure 4.4. Simulation workflow predicts the nanobody-dockerin binding interface. A) Protein-protein docking is a challenging problem in which the interaction space is searched by rotating and translating one protein around the other. Our workflow started with rigid protein-protein docking of several dockerin conformers to NbE6 with ClusPro. The best docking poses for each conformer then underwent MD simulations for better sampling of flexible regions. Protein-protein contact data were compiled from MD simulations to generate a rank-ordered list of predicted binding interface dockerin positions. B) The model that best agreed with modeling workflow predictions is consistent with a canonical protein interaction interface with hydrophobic amino acids at the center surrounded by a dense network of salt bridges and hydrogen bonding networks. Predicted key contact-forming positions are annotated.

rigid docking server [180–182] to predict structures of the NbE6-double dockerin complex. To better sample backbone flexibility, the best ClusPro prediction for each NbE6-doc5600 conformer pair underwent explicit solvent MD simulation. We hypothesized that doc5600 residues that frequently contact NbE6 across simulations of all complex models most likely comprised the binding interface.

Compiled MD simulation data identified 22 double dockerin residues likely involved in binding NbE6 (Fig. 4.4b, Table 4.1). Of these 22, we cloned and measured the dissociation constant of four point mutants – Y22V, W28F-W74F, Y67V, and D69A – to validate model predictions. We chose these mutants because they cover sites from both folded domains of doc5600 and are predicted to minimally affect the doc5600 folding free energy while changing the chemical character at that site [183]. Y67V had a significantly larger dissociation constant compared to WT doc5600 ($p = 0.0072$, unpaired two-sided t-test) and a dissociation constant could not be fit for the D69A data, validating the importance of these two positions in binding NbE6; Y22V and W28F-W74F showed little change, perhaps suggesting only hydrophobic character is important at these positions (Fig. 4.5). The workflow-generated model consistent with the experimental data in which most of the 22 predicted binding residues formed contacts with NbE6 is shown in Fig. 4.4b.

Table 4.1. Summed contact counts between library mutated doc5600 and NbE6 residues computed from MD simulations. Each row corresponds to a doc5600 residue, each column to a NbE6 residue.

	NbE6 Res																														
doc5600 Res	**R113**	**L109**	**R62**	**W116**	**L111**	**S112**	**T114**	**Y64**	**R106**	**G110**	**G108**	**L33**	**N59**	**P107**	**G115**	**A61**	**T63**	**Q57**	**D118**	**Q105**	**Y65**	**G117**	**S32**	**S58**	**R29**	**R35**	**G60**	**I56**	**S30**	**F119**	**A55**	**D104**
E70	306	143	213	218	46	174	203	58	28	58	1				39		10	13	87			43					21	24		2		
Y67	290	374	158	2	190	60	85	52	18	50	74			104	14	2	63		73		39	14					3					1
E25	203	142	152	66	91	77	80	160	80	72		97	62		22			66					86	53		45			16		1	
K41	163	3	156	268	119	191	287	101		12			40		47	13		73			4											
K23		221	192	31	120	27	3	96	194	84	23	214	1	21		22	26			123	8	1			32							
W4	155	180	93	20	178	129	106	55		52			53		2	32		31			27											
Y66	120	179	26		66			1		12	59			38	2				12	25	8											
T68	168	258	104		47	65	14	20	130	59	36			15				7	6													
N17	4	61	204	174	37	39	52	108		6			86		18	29	4	19	1		1	20		3								
D69	182	84	79	32	7	99	76	29	1	26					2	3	34		56		5	29										
D26	145		110	49	196	72	8	57	33	15			55					20						7								
L8	324	6	29	190		4	27	7	1						44																	
W81	108	152	1						33	1	99	7		85					3	51					1							
T6		141	6		128	15				60			8																			
Y10	80			69		13	92	60							35						24					2						
D24	83	60	38	3	18	30		18	44	35		78											19	1								

Our model suggests that residues from all three NbE6 CDR’s form contacts with doc5600, including 13 polar contacts primarily involving NbE6 Gln, Tyr, and Arg sidechains. While the static model has two salt bridges (NbE6 R62 – doc5600 E25 and NbE6 R29 – doc5600 D69), simulation data suggests that NbE6 R106 and R113 may form additional salt bridges with doc5600 Asp and Glu side chains (Table 4.1). The NbE6-doc5600 complex is likely further stabilized by burying hydrophobic residues, including doc5600 W4 and NbE6 P107, L109, and L111. As expected, neither terminus of doc5600 is predicted to be near the binding interface, as NbE6 was raised against cellulosome proteins containing double dockerins fused to other domains.

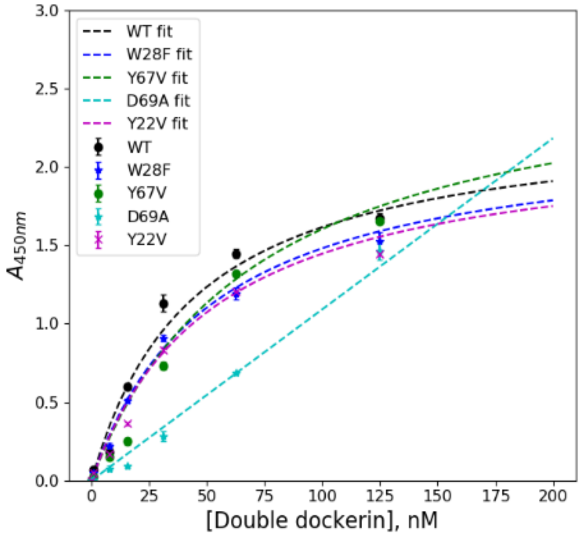


Figure 4.5. ELISA-based dissociation constant measurements for WT doc5600 and point mutants. Error bars represent SEM of N=2 at each concentration. K_d measurements were as follows: WT - $\mathbf{44.5 \pm 6.9}$ **nM**; W28F - $\mathbf{52.1 \pm 11.0}$ **nM**; Y67V - $\mathbf{71.2 \pm 6.9}$ **nM**; D69A – no fit; Y22V - $\mathbf{53.3 \pm 141}$ **nM**.

The canonical model of a cellulosome contains a central scaffoldin protein comprising multiple, modular, dockerin-binding domains that assemble many dockerin-containing enzymes into a single entity (Fig. 4.1a). Towards building synthetic fungal cellulosomes that assemble via dockerin-nanobody interactions, we designed a synthetic “scaffoldin” composed of tethered NbE6 domains. As a proof of concept, we developed a scaffoldin with two NbE6 domains that forms heterotrimeric complexes (Fig. 4.6a).

Biophysical attributes of the scaffoldin are key to enabling complete cellulosome formation at biologically relevant protein concentrations, controlling complex composition, and to enabling synergistic activity between complementary cellulosomal enzymes [79,184]. To guide scaffoldin design, we considered binding site availability as a metric easily

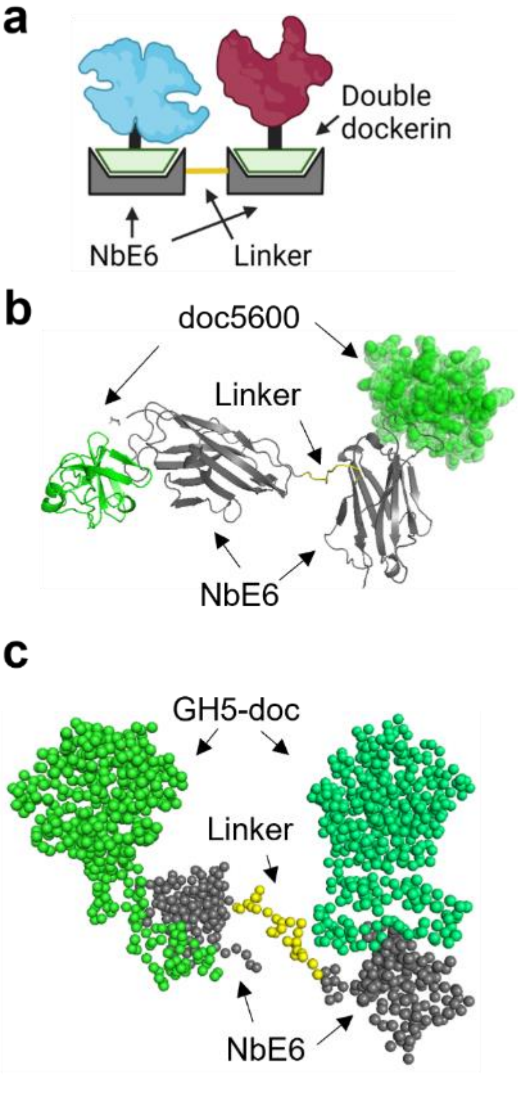


Figure 4.6. Tethered NbE6 scaffoldin design guided by simulation. A) Cartoon schematic of a synthetic fungal cellulosome system. B) A snapshot from an atomistic REMD simulation of a tethered NbE6 scaffoldin construct with two doc5600 molecules docked to each binding site. Linker shown in yellow. C) A snapshot from a coarse-grained, trimeric system comprising a tethered Nb scaffoldin bound by two doc5600-GH5 fusion proteins.

measurable with MD. Natural bacterial scaffoldins have cohesins linked by flexible, Pro/Thr-rich linkers that range in length from as few as four amino acids to over one hundred [185]. Thus, we computationally screened tethered nanobody scaffoldins with Gly/Ser linkers of various lengths to investigate how linker length affected predicted dockerin binding site availability using atomistic REMD simulations.

REMD simulation data were collected for two scaffoldin systems, one with a 9-mer Gly/Ser linker and one with a 20-mer Gly/Ser linker between two NbE6 domains, to predict how binding site availability was affected by increasing linker length. To estimate binding site availability, simulation trajectories were post-processed, "docking" a doc5600 structure to each NbE6 binding site at each trajectory frame and checking for steric overlaps between dockerins and scaffoldins (interatomic distances <7 angstroms) (Fig. 4.6b). While the 9-mer Gly/Ser linker scaffoldin had two unblocked binding sites only 18.7% of the time, a 20-mer Gly/Ser linker scaffoldin was fully binding competent 58% of the time, suggesting tethered nanobody scaffoldin architectures should have linkers of 20 residues or longer.

Simulating full trimeric systems of tethered nanobody scaffoldins with bound dockerin-fused enzymes provided the best way to predict synthetic cellulosome properties for a given scaffoldin design. However, these systems are large (trimeric systems are about 1200 amino acids) requiring extensive computational resources and advanced sampling techniques to simulate atomistically. To circumvent this challenge, we employed relative entropy coarse graining [186] to faithfully simulate the behavior of atomistic scaffoldin designs. This bottom-up coarse graining replaces the many atoms within amino acids with single coarse grained "beads" whose physical interactions are optimized to match the configurational probability distributions of the atomistic system (Fig. 4.6c).

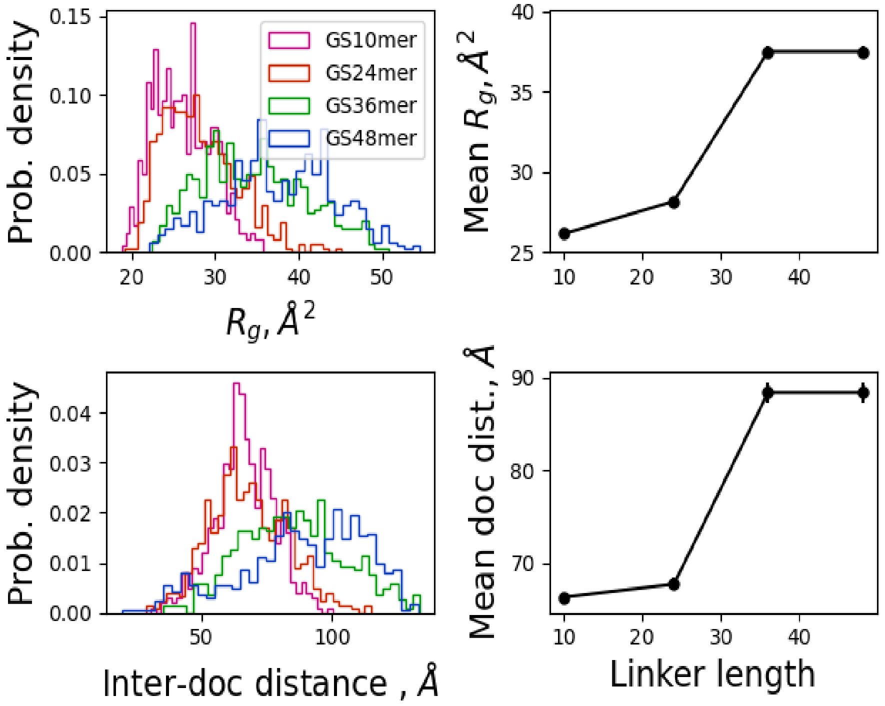


Figure 4.7. Radius of gyration and interdockerin distance distributions for tethered NbE6 scaffoldin designs as a function of the Gly-Ser linker length between NbE6 domains. Linkers with lengths ranging from 10-48 amino acids were tested. Line plots are presented as the mean ± the standard error of the mean.

Coarse-grained simulations of scaffoldin designs with Gly/Ser linkers ranging from 10-48 amino acids predicted a step-wise increase and plateau in scaffoldin R_g and bound inter-dockerin distance for linkers 36 residues or longer (Fig. 4.7), consistent with experimental observations for synthetic bacterial mini-scaffoldins [184]. Since longer linkers are predicted to improve binding competency, we experimentally produced and tested the Gly/Ser 36-mer tethered nanobody scaffoldin (TetNb-36) to confirm its ability to form trimeric complexes.

Complex formation between purified TetNb-36 and TrxA-fused doc5600 was assessed by non-denaturing SDS-PAGE. Due to its pI of 9.3, TetNb-36 on its own does not run into a tris-glycine gel under standard conditions (pH 8.3), while TrxA-doc5600 migrates far into the gel (Fig. 5). When TetNb-36 and TrxA-doc5600 are mixed at 200 nM doc5600 concentration, a clear, higher molecular weight (MW) band emerges, and band intensity increases significantly when the TrxA-doc5600 concentration is increased to 1 μM. The two

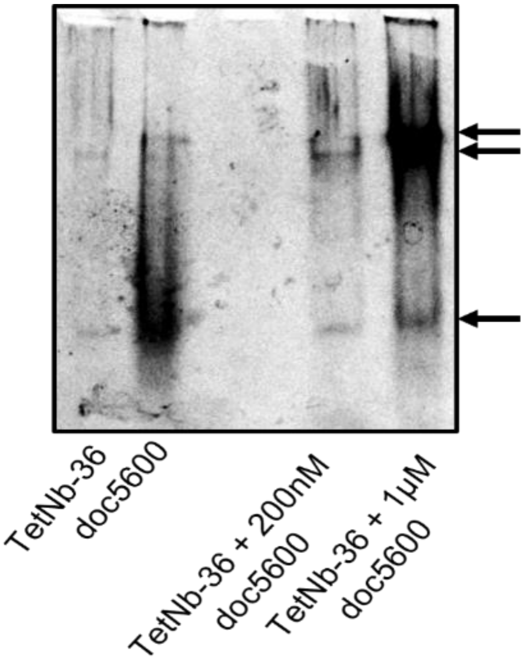


Figure 4.8. Non-denaturing SDS-PAGE and TetNb-36 mixtures with doc5600 shows clear binding. doc5600 clearly migrates from the lower band to a band that migrates more slowly through the gel, indicative of complex formation with TetNb-36. The larger species exhibit a double banding pattern suggestive of a heterodimeric and heterotrimeric species, annotated with arrows. Band intensity of both species, especially the upper band increases with increasing doc5600 concentration.

lanes with TetNb-36 and TrxA-doc5600 mixtures show two bands close in MW, suggesting the formation of dimeric and trimeric species.

Native mass spectrometry (MS) of TetNb-36 and dockerin mixtures provided an unambiguous detection method for heterooligomeric species. Unfortunately, TetNb-36 and TrxA-doc5600 mixtures did not produce resolvable complexes by native MS, but complexes with TetNb-36 and a double dockerin-containing Glycoside Hydrolase family 5 (GH5) enzyme produced clear peaks corresponding to heterodimer and heterotrimer species, confirming the TetNb-36 construct can form trimeric complexes as designed (Fig. A2).

Having demonstrated synthetic protein complex formation based on the NbE6-doc5600 pair, we sought to transform these domains into parts for constructing complexes with stimuli-responsive assembly and composition. We focused on developing a set of NbE6-dockerin pairs with varied pH-dependent binding behaviors, as varied binding pH dependencies are necessary to enable predictable, pH-dependent protein complex composition (Fig. 4.1d). Without an experimental NbE6-double dockerin structure, we chose a library screening strategy informed by our structure prediction workflow in which 22 predicted interface double dockerin positions were combinatorially mutated to histidine to impart pH-dependent binding behavior (Fig. 4.9) [187,188].

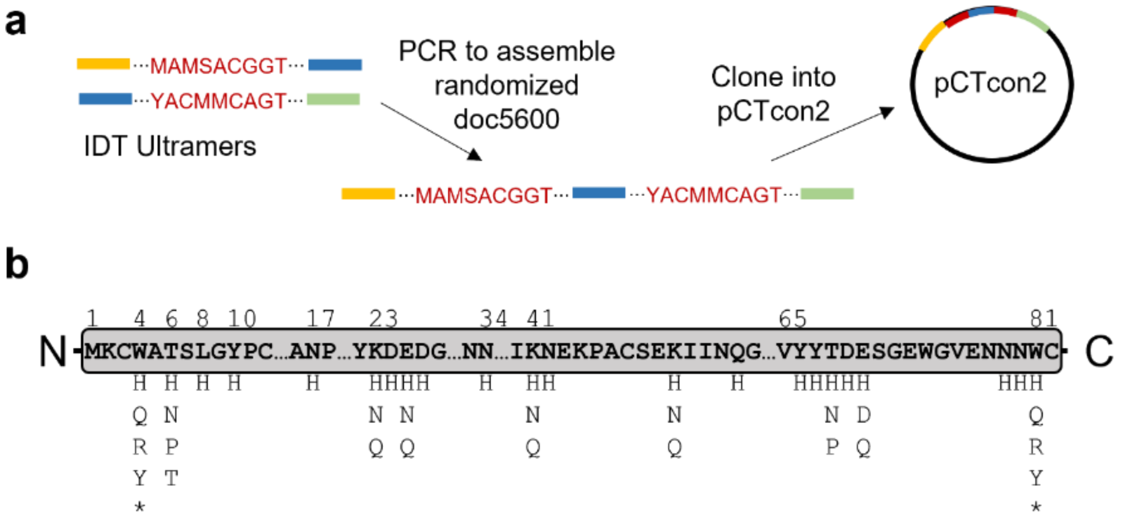

a
···MAMSACGGT···
···YACMMCAGT···
IDT Ultramers
PCR to assemble randomized doc5600
···MAMSACGGT··· ···YACMMCAGT···
Clone into pCTcon2
pCTcon2
b
1 4 6 8 10 17 23 34 41 65 81
N
MKCWATSLGYPC...ANP...YKDEDG...NN...IKNEKPACSEKIINQG...VYYTDESGEWGVENNNWC
C
H H H H H HHHH H HH H H HHHHH HHH
Q N N N N N N D Q
R P Q Q Q Q P Q R
Y T Y
* *

Figure 4.9. Combinatorial Histidine library construction. A) Schematic of library construction procedure with overlapping, randomized IDT Ultramer oligos. B) Representation of the library diversity mapped onto the WT doc5600 sequence shown in gray. Amino acids listed under a given position represent sequence possibilities at that position encoded in the library. * is a STOP codon.

To rapidly screen doc5600 variants for pH-dependent binding, we transformed the library into EBY100 *Saccharomyces cerevisiae* cells for screening by yeast surface display [189]. Transformation efficiencies limited our screening to only ~10% (1.3 x 10^8 variants) of the theoretical library size (1.8 x 10^9 variants), but we hypothesized that, since ~96% of library variants would theoretically have at least 5 histidine residues, we would identify variants with desired phenotypes among the 10^8 we screened.

We used a multi-step, magnetic cell sorting (MACS) screen that utilized both biotinylated and unlabeled NbE6 to select for tight binding at neutral pH and no binding at acidic pH (Fig. 4.10) [187]. Five rounds of MACS were performed, with the first round only selecting for binding at pH 7.2 to remove doc5600 variants that poorly display in yeast or that are non-functional. Hypothesizing that changing the pH's used during MACS selection

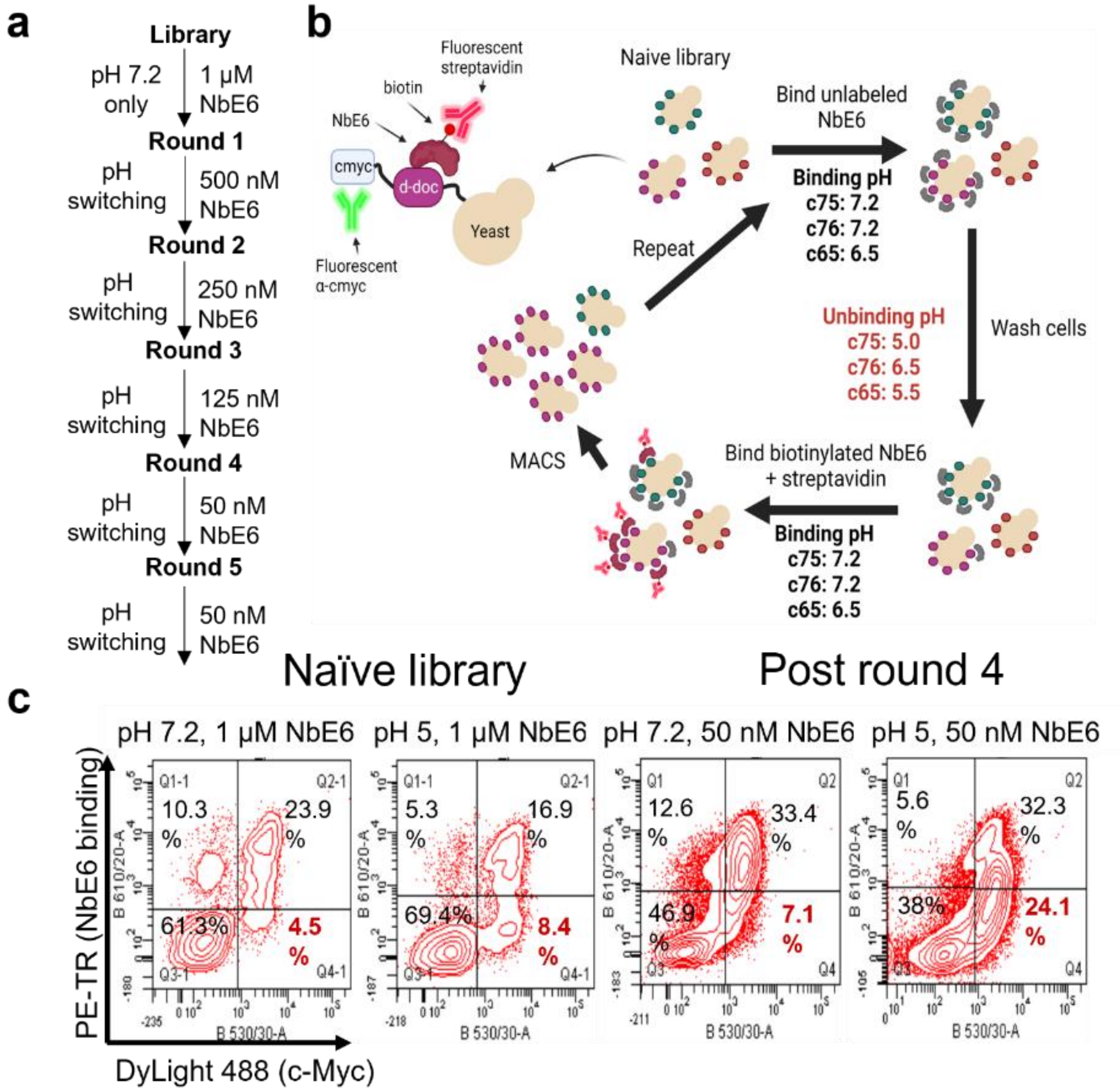
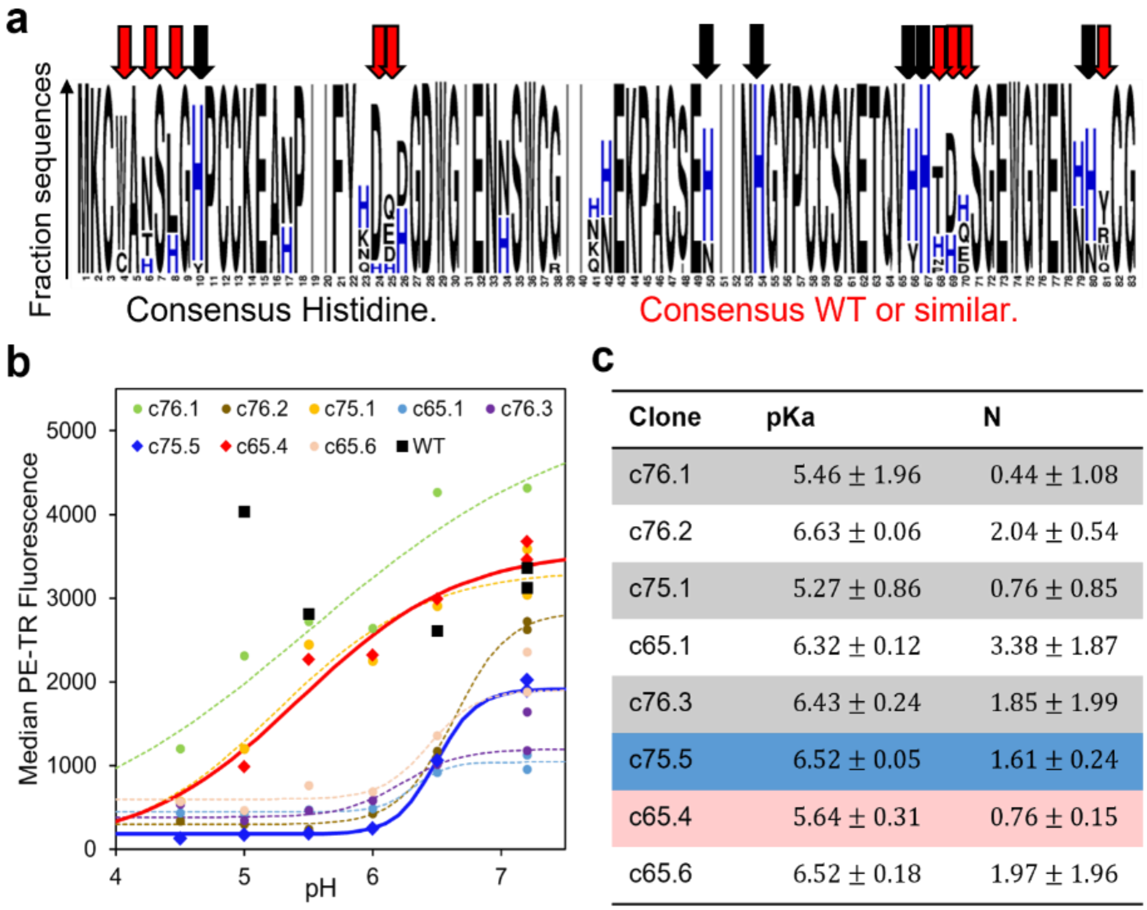


Figure 4.10. A MACS screen for pH-dependent NbE6 binding. A) Screening schedule for pH-dependent double dockerins. After a first round of positive-only selection, biotinylated NbE6 concentrations were sequentially lowered over five rounds of multi-step screens to select for switch-like binding. B) A schematic of the screening procedure that enriches for pH-dependent binders. Three parallel campaigns with different selection pH's were performed, denoted as c75, c76, and c65. C) Enrichment for pH-dependent binders between the naïve library and post round 4 library is apparent from a large increase in the change of the Q4 population fraction between pH 7.2 and pH5 (bolded and red). Panel B was created with Biorender.com.

would select for clones with different pH dependent binding behaviors, we performed three selection campaigns in parallel – one with positive selection at pH 7.2 and negative selection at pH 5.0 (c75), one with positive selection at pH 7.2 and negative selection at pH 6.5 (c76), and one with positive selection at pH 6.5 and negative selection at pH 5.5 (c65).Sanger sequencing of 22 clones across the three screening campaigns after the fifth selection round showed the emergence of consensus sequence characteristics conferring tight NbE6 binding at neutral pH and possibly binding pH-dependence (Fig. 4.11a). Among sequenced clones, nine positions were consensus WT or similar (W4, T6, L8, D24, E25, T68, D69, E70, and W81), implying a critical role for that amino acid chemical character either in the dockerin fold or in maintaining high affinity binding to NbE6. Five double dockerin sites were mutated to His in >50% of clones – Y10, K50, N54, Y66, and Y67, implicating these sites in binding pH dependence. Sequencing consensus agreed with experimental point mutation data (Fig. 4.5), and overall showed that double dockerin tolerates substantial sequence modifications while maintaining NbE6 binding, an encouraging result for future efforts focused on engineering the double dockerin domain for other means.

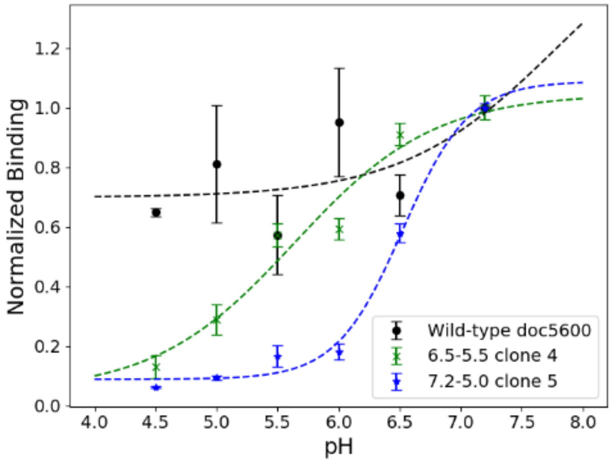
We measured by flow cytometry NbE6 binding as a function of pH for eight sequenced clones. NbE6 binding levels when incubated with 50 nM biotinylated NbE6 were measured across six pHs and the resulting data were fit to the following model to extract switch parameters: $F(pH) = \left(\frac{F_{max}}{1+10^{n(pH_0-pH)}}\right) + F_{min}$ (Eqn. 1), where F is fluorescence signal, n is



Clone	pKa	N
c76.1	5.46 ± 1.96	0.44 ± 1.08
c76.2	6.63 ± 0.06	2.04 ± 0.54
c75.1	5.27 ± 0.86	0.76 ± 0.85
c65.1	6.32 ± 0.12	3.38 ± 1.87
c76.3	6.43 ± 0.24	1.85 ± 1.99
c75.5	6.52 ± 0.05	1.61 ± 0.24
c65.4	5.64 ± 0.31	0.76 ± 0.15
c65.6	6.52 ± 0.18	1.97 ± 1.96

Figure 4.11. Characterized clones display a range of binding pH-dependencies. A) Sequence logo plot of 23 clones enriched after five rounds of MACS screening. Arrows indicate consensus His or WT amino acids. B) Enriched clones across the three screening campaigns show a clear binding pH dependence not exhibited by WT doc5600 (black squares) when incubated with 50 nM biotinylated NbE6. "c76" indicates the clone was isolated from screening for binding at pH 7.2 and unbinding at pH 6.5. "c75" – screening campaign with binding at pH 7.2, unbinding at pH 5.0; "c65" – screening campaign with binding at pH 6.5, unbinding at pH 5.5. Clones c65.4 (red bold) and c75.5 (blue bold) were chosen for further investigation into key features conferring different pH sensitive binding phenotypes. Plotted are N=1 measurements per pH, except for pH 7.2 where N=2. Median fluorescence values are computed for dockerin displaying cells only (α-cmyc DyLight 488 signal > 800). C) Fit pH_0's and Hill coefficients (n) for characterized clones computed using scipy. Data were fit to the equation: $F(pH) = F_{min} + \frac{F_{max}}{1+10^{-n(pH-pKa)}}$ using nonlinear least squares in SciPy with parameter estimate uncertainty equal to one standard deviation. C75.5, C65.4, and WT doc5600 were fit to N=3 replicates per pH collected over two independent trials to ensure reproducibility (Figure 4.12).

the Hill coefficient, and pH_0 is the transition pH (Fig. 4.11b). The eight double dockerin variants demonstrated a range of switch parameters, with transition pH's ranging from 5.2 to 6.6 and Hill coefficients ranging from <1 to ~3, while WT doc5600 showed no clear pH dependence (Fig. 4.11c). Results were consistent across multiple independent trials (Fig. 4.12). There was no clear evidence that changing the selection conditions selected for specific Hill coefficient or pH_0 ranges as hypothesized, but more variants from each campaign should be characterized to confirm these conclusions.



	Hill coefficient, n	Binding midpoint, pH_0
Wild-type	-	-
6.5-5.5 clone 4	0.76 ± 0.15	5.64 ± 0.31
7.2-5.0 clone 5	1.61 ± 0.24	6.52 ± 0.05

Figure 4.12. Doc5600 WT and mutant variant binding profiles as a function of pH are reproducible. Plotted data are means ± one standard deviation of N=3 replicates measured across two independent trials. Data were fit to the equation $\boldsymbol{F = F_{min} + \frac{1}{1+10^{-n(pH-pH_0)}}}$ using nonlinear least squares. Fit parameters are means ± one standard deviation.

An important question is whether specific histidine mutations predict binding pH dependence, or if only the total number of histidines, regardless of position, is relevant. To answer this question, we analyzed data from Fig. 6 using LASSO regression. The response variable was either the fit Hill coefficient or the dynamic range $(F_{pH=7.2}/F_{pH=5.0})$, and each of the 22 histidine-mutated positions in the library represented an independent variable that was "1" if histidine and "0" if not. The total number of histidines in each sequence was included as another independent variable. While total histidine count was most important to predicting Hill coefficient, H8 and H26 had the only positive coefficients for predicting

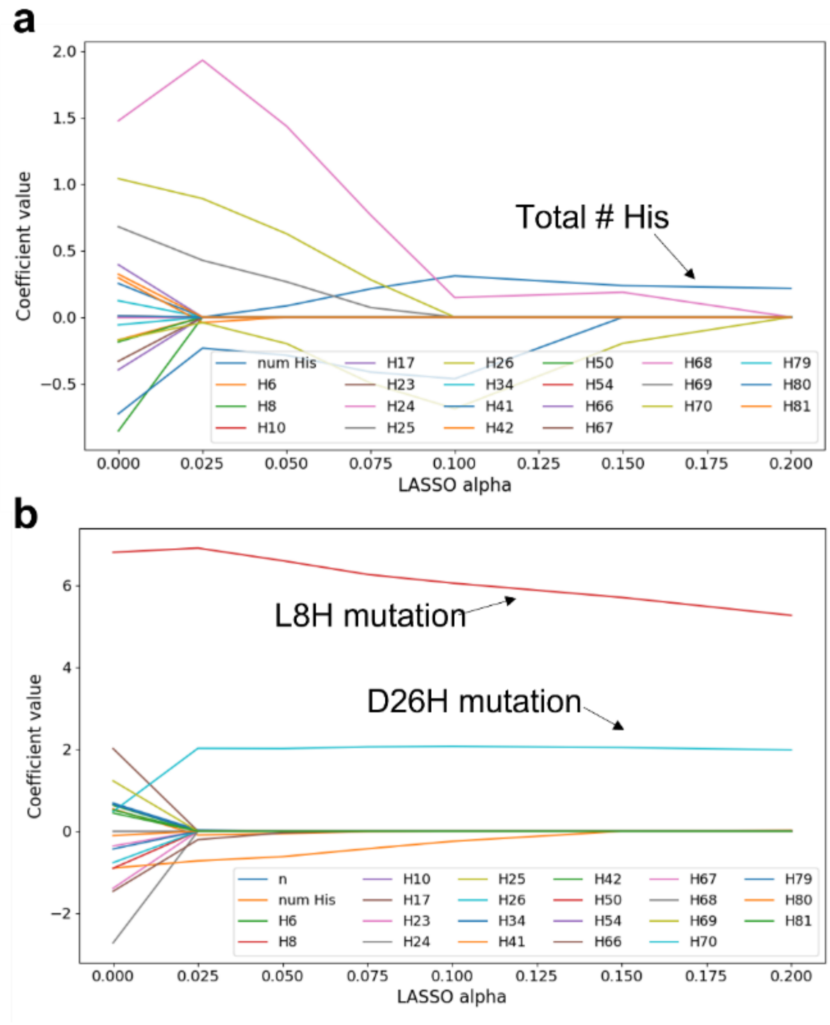


Figure 4.13. a) Independent variable coefficients as a function of LASSO alpha parameter for predicting Hill coefficient (n). b) Independent variable coefficients as a function of LASSO alpha parameter for predicting binding dynamic range.

dynamic range with non-zero α, suggesting there are position-specific effectors of pH dependence (Fig. 4.13).

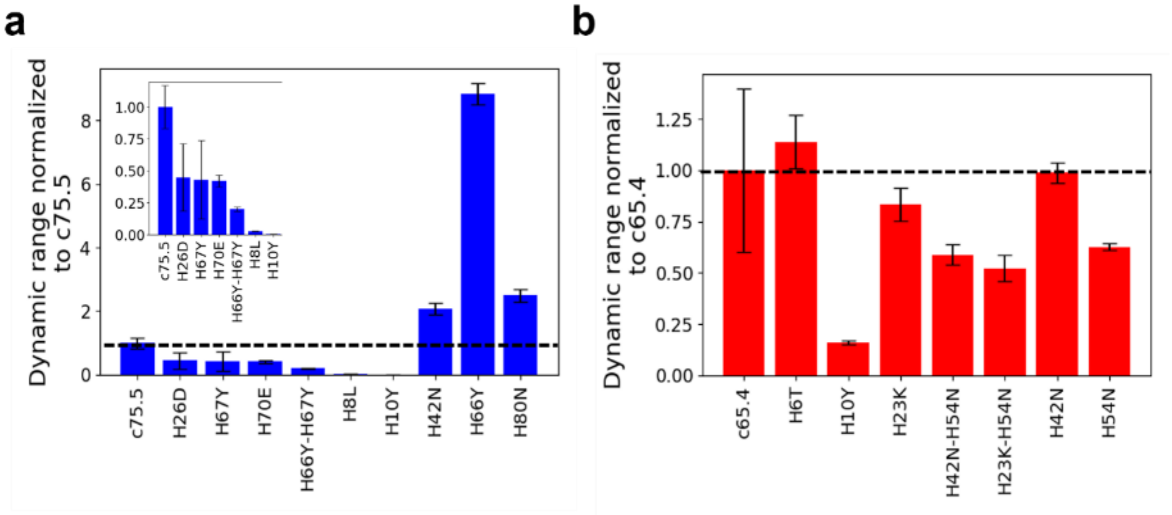
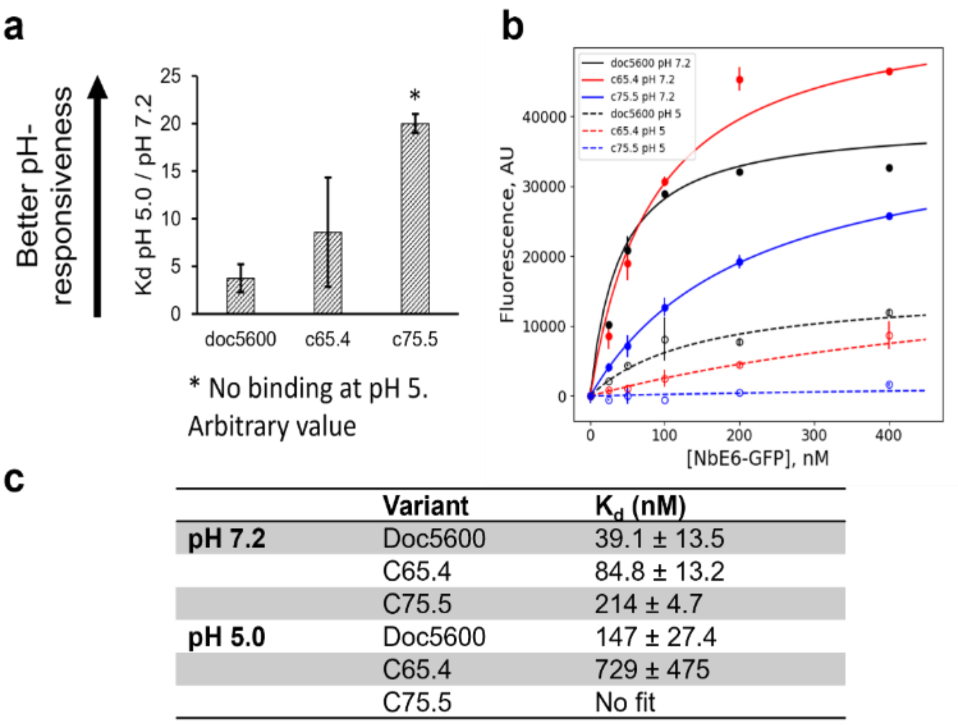


Figure 4.14. Point mutagenesis of His residues identifies those conferring pH sensitivity. Measured changes in dynamic range (measured binding signal at pH 7.2 divided by that at pH 5) for each mutant relative to c75.5 (A) or c65.4 (B). The raw dynamic ranges of c65.4 and c75.5 were 6.8 and 160 respectively. See Figure S8 for raw binding curves. Error bars represent propagated uncertainty computed from standard deviations of N = 2 at each pH.

We hypothesized that histidines conferring binding pH sensitivity could be identified if reverting that residue to the WT amino acid decreased binding pH responsiveness. For further analysis, we chose two dockerin variants, c65.4 and c75.5, with different transition pH's and Hill coefficients (thick curves in Fig 6b). Single point mutants and some double mutants of c65.4 and c75.5 were cloned into pCTcon2 to measure changes in binding dynamic range ($F_{pH=7.2}/F_{pH=5.0}$) with mutation. Our results indicate that only one or two histidine positions in each clone (Y10H/N54H in c65.4 and L8H/Y10H in c75.5) impart most of the binding pH sensitivity on the yeast surface (Fig. 7, Fig. S12). Generally, removing histidines reduced the dynamic range as expected, but four mutations – c65.4 H6T and c75.5 H42N, H66Y, and H80N - boosted dynamic range by increasing binding signal at neutral pH; c65.4 H23K and H42N had little effect. N54 is predicted to sit at the doc5600-NbE6 binding interface in the model shown in Fig. 4.4 potentially driving pH responsiveness via pH-dependent charge repulsion against NbE6 R29. L8 and Y10 are not at the binding interface in Fig. 4.4, though L8 contacted NbE6 R113 more frequently than any other doc5600 residue in simulations, suggesting it could also contribute to pH dependence via the same mechanism as N54 (Table 4.1). Y10H's mechanistic contribution to pH dependence is unclear, perhaps destabilizing the doc5600 folding free energy at low pH instead of affecting the binding free energy.

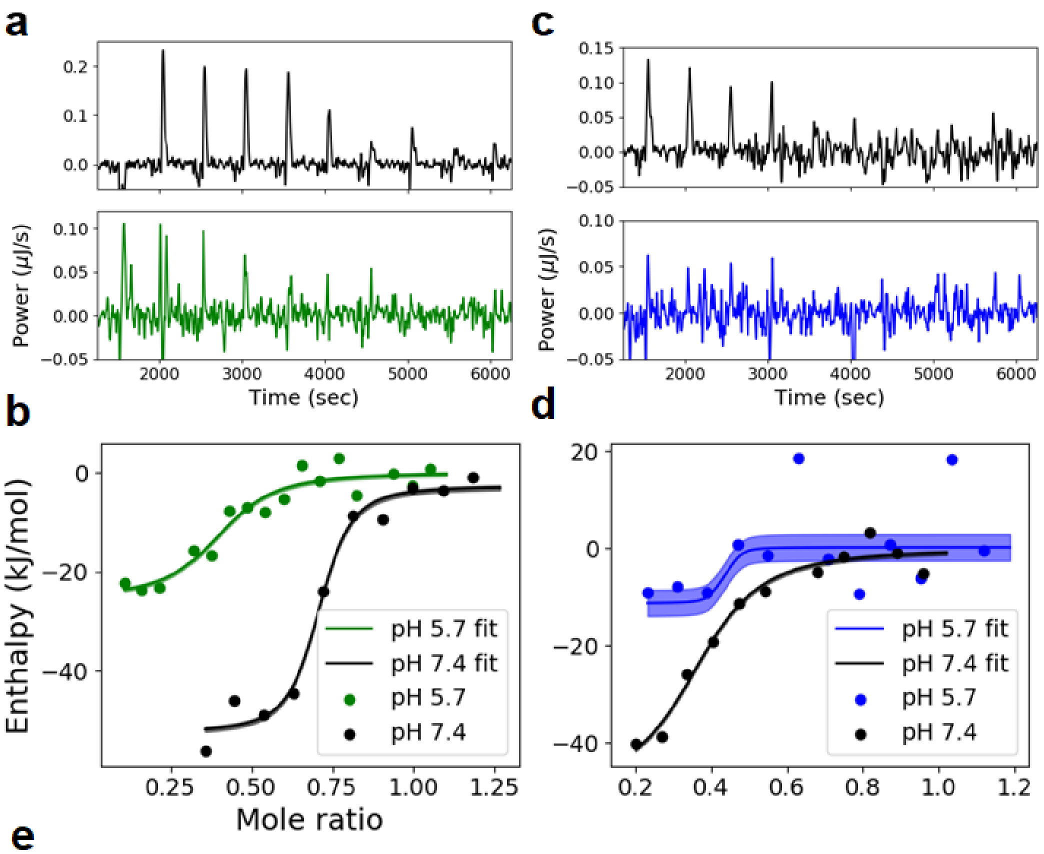
Since our desired application for dockerin-NbE6 pairs with pH-dependent binding is for soluble protein complex formation, we tested whether c65.4 and c75.5 maintained NbE6 binding pH dependence off the yeast cell surface using plate-based assays (Fig. 4.15). When we measured equilibrium binding affinities for immobilized c65.4, c75.5, and WT doc5600



	Variant	K_d (nM)
pH 7.2	Doc5600	39.1 ± 13.5
	C65.4	84.8 ± 13.2
	C75.5	214 ± 4.7
pH 5.0	Doc5600	147 ± 27.4
	C65.4	729 ± 475
	C75.5	No fit

Figure 4.15. Purified, MACS-enriched variants display pH-dependent nanobody binding when immobilized on a plate surface. Binding of GFP-fused NbE6 to dockerin variants immobilized on a microplate surface was monitored by GFP fluorescence. A) Fold change in the measured K_d at pH 5.0 vs pH 7.2 for WT doc5600, c65.4, and c75.5 computed from equilibrium binding curves (B). An equilibrium constant could not be fit for c75.5 in B) so 20 was set as an arbitrary value. C) Fit parameters extracted from B) provided as mean ± SD for N=2 replicates per pH.

with soluble GFP-tagged NbE6 at pH 7.4 and pH 5.0, we saw the expected results (Fig. 4.15a-b). We confirmed that nanobody binding pH dependence is an intrinsic molecular property of c75.5 by measuring titration curves at neutral and acidic pH by isothermal



e

	Ka (M-1)	N	ΔH (kJ/mol)	ΔS (J/molK)
doc5600 pH 7.4	$2.0 \pm 5.7\ x\ 10^8$	0.67 ± 0.03	-50.1 ± 8.4	-9.2
doc5600 pH 5.7	$1.9 \pm 21\ x\ 10^7$	0.39 ± 0.08	-25.8 ± 22.5	53.0
c75.5 pH 7.4	$2.2 \pm 6.1\ x\ 10^7$	0.35 ± 0.07	-50.0 ± 32.6	-27.2
c75.5 pH 5.7	$1.0 \pm 0.0\ x\ 10^{9*}$	0.40 ± 8.01	-11.5 ± 161	133.8

Figure 4.16. Isothermal calorimetry confirms c75.5 maintains binding pH dependence in solution. A) Power vs time data for WT doc5600 at pH 7.4 (black, top subplot) and pH 5.7 (green, bottom subplot) when titrated with NbE6. B) Titration data and fit curves from which to estimate thermodynamic parameters. C) Power vs time data for c75.5 at pH 7.4 (black, top subplot) and pH 5.7 (blue, bottom subplot). D) Titration data and fit curves. Shaded areas denote ± one SD from nonlinear least squares fit. E) Table of fit parameters reported as the mean ± one SD of 5000 fit trials. ΔS is calculated from the mean K_a and ΔH values. *This was the maximum K_a value allowed during fitting.

calorimetry (ITC) (Fig. 4.16). Consistent with plate-based results, c75.5 shows binding at pH 7.4 and markedly reduced binding at pH 5.7, compared to WT doc5600 which produces binding curves at both pH's. While equilibrium binding affinities are not fit with confidence from these data, binding enthalpies are, showing clear increases in ΔH_{bind} at low compared to neutral pH consistent with less favorable binding.

Binding pH sensitivity of WT doc5600 observed in Figures 4.15 and 4.16 but not in yeast surface display results (Fig. 4.12) suggests the 6x His tags partially contribute to this behavior (surface displayed dockerins had no His tags, but NbE6 had a C-terminal 6x His tag). However, the near complete ablation of nanobody binding at pH 5.7 of c75.5 relative to WT doc5600 provides strong evidence that c75.5's pH dependent binding phenotype is real. Exploiting these proteins' distinct nanobody binding affinities as a function of pH presents a promising way to construct protein complexes that predictably assume different assembly states as a function of the environment.

4.3 Conclusions

Fungal cellulosomes present an attractive template for synthetic protein complexes designed for broad synthetic biology applications ranging from chemicals production [163] to cellular reprogramming [165]. However, an inability to reconstitute fungal cellulosome complexes *in vitro* without knowing the native dockerin binding partner's identity has precluded their biotechnological development [175]. In this study, we developed a small, modular domain that binds double dockerin-containing proteins with high affinity. Using this domain, we designed synthetic, multimer-forming "scaffoldins," providing the first opportunity to leverage the large, diverse bank of natural, fungal double dockerin-containing proteins [83] and functional chimeras in building synthetic protein complexes. Using an integrated computational modeling and experimental high-throughput protein engineering approach, we developed this modular interacting pair into a suite of binding domains with varying pH-dependent binding behavior. This library of interacting domains represents a new toolkit for engineering protein complexes with multi-state, pH-responsive assembly behavior when binding domains with different pH-dependence are combined. How pH-responsive assembly behavior affects multi-step catalysis in reaction systems with microscale pH gradients presents one interesting application of our tools. Furthermore, our hybrid computational-experimental approach integrating structure-guided residue/hotspot identification and high throughput mutagenesis and selection is readily extendable to engineering other sets of binding domains with binding triggers other than pH. Overall, this work presents a framework for building stimuli-responsive protein complexes and puts forth a blueprint for engineering these systems for specific applications.

4.4 Materials and Methods

Statistics and curve fitting

Unless otherwise stated, binding data were fit by nonlinear least squares with SciPy's curve_fit library. Reported fit parameter uncertainties were computed as the standard deviation of each parameter estimate.

Cellulosome isolation for nanobody generation

Neocallimastix californiae was cultured on Medium C with 1% (w/v) reed canary grass as substrate as described in [190]. 6-8 day old cultures were transferred to centrifuge bottles or Falcon tubes and centrifuged at 3,000x g for three minutes in a fixed angle rotor to pellet cells and residual substrate. The supernatant was then transferred to new centrifuge bottles and the pH adjusted to 7.0 with 9N NaOH. Secreted cellulase concentration was estimated by Bradford assay (ThermoFisher Scientific cat. 23238), subtracting out the signal from uninoculated media with the same substrate. Cellulosomes were then purified from fungal supernatant using an adapted version described below of the affinity digest protocol developed for purifying cellulosomes from *Clostridium thermocellum* [126].

Phosphoric acid swollen cellulose (PASC), prepared as described in [126], was added to pH-adjusted fungal supernatant at a concentration of 4 mg PASC (dry weight) per mg secreted cellulase. The mixture was incubated on a rocking platform at 4°C for one hour to adsorb cellulosomes and other cellulose-binding proteins out of solution. PASC and adsorbed proteins were pelleted by centrifugation at 12,000x g for 10 minutes in a fixed angle rotor and the supernatant was discarded. The resulting pellet was resuspended in 20 mL of dialysis buffer (10 mM MES buffer pH 6.4) and transferred to a 30 mL 10 kDa MWCO Slide-A-lyzer dialysis cassette (ThermoFisher Scientific cat. 66830) for dialysis at 39°C against 10 mM MES buffer pH 6.4. Dialysis buffer was changed approximately every hour until the PASC was completely solubilized (~4-8 hours). The dialyzed solution was further clarified by centrifugation at 12,000x g for 5 minutes to pellet remaining insoluble, and the clarified solution was concentrated using a 300kDa MWCO PES filter (Sartorius cat. VS2051) by repeated centrifugation at 4,000x g and 4°C in a swing bucket rotor until the concentration rate appeared to slow considerably. Concentrated sample was aliquoted and snap-frozen in a metal block before long-term storage at -80°C.

500 µL of affinity digest-purified cellulosome sample was loaded on a NGC Quest 10 chromatography system (Bio-Rad Laboratories, Hercules, CA) equipped with a Superose 6 Increase 10/300 GL column (Cytiva Life Sciences) kept at 4°C to purify high molecular weight cellulosome complexes from the crude affinity digest preparation. PBS was used as

mobile phase. 1 mL fractions were collected and pooled to capture the ~1 MDa cellulosome peak before concentration with a 0.5 mL 30 kDa MWCO PES filter (ThermoFisher Scientific cat. 88502). Concentrated samples were flash frozen and stored at -80°C. Sample protein concentrations were measured using the BCA assay (ThermoFisher Scientific cat. 23225).

1 mg of purified *N. californiae* cellulosome was sent to the University of Kentucky Center for Molecular Medicine's Nanobody Production Core as part of their fee-for-service nanobody production service [191]. Nanobody E6 was enriched after several rounds of panning against purified cellulosome *in vitro*. The sequence was cloned into a pMES expression vector with a C-terminal 6x-His tag for recombinant production and purification in *E. coli*.

Nanobody and Tethered Nanobody protein production

pMES4-NbE6 was transformed into *E.coli* BL21 (DE3) for protein production. pMES4-TetNb-36 LB media was inoculated at a starting OD_{600} of 0.05 and incubated at 37°C with shaking at 250 rpm. When the culture reached an OD_{600} of 0.5, IPTG was added to a final concentration of 0.5 mM and the culture was incubated for an additional 18 hours at 30°C before harvesting. Cells were harvested by centrifuging in a swing bucket rotor at 4000x g for 10 minutes before resuspension in PBS plus 10 mM imidazole (pH 7.4) with Pierce protease inhibitor following the manufacturer's instructions. 0.5 mm zirconia beads were added to 10% the liquid volume and the resuspended cells plus beads were vortexed for 30 seconds 10 times to lyse cells. Cell debris was pelleted by centrifugation in a fixed angle rotor at 10,000x g for 10 minutes. Nanobody was then purified from the soluble cell lysate by immobilized metal affinity chromatography following the manufacturer's instructions. Typical protein yield was ~6 mg/L culture for 50 mL cultures in 250 mL Erlenmeyer flasks.

As required for FACS analysis, biotinylated nanobody was obtained using an EZ-Link^{TM} NHS-PEG4 biotinylation kit (ThermoFisher Scientific cat. No. 21455).

Nanobody affinity chromatography

3.5 mg purified nanobody was immobilized on a 15 mL agarose column using an AminoLink Plus Immobilization Kit (Thermo Fisher cat. No. 20394). To determine the nanobody binding partner, a mixture of cellulosome proteins from the supernatant of a *N. californiae* culture was passed over the column. The column was washed four times with three column volumes of PBS (pH 7.2) each. Two washes with 0.1M glycine buffer (pH 3.0) eluted proteins that bound the nanobody. Fractions from each step of the affinity chromatography protocol were analyzed by SDS-PAGE and Western blot using an anti-double dockerin antibody [120] to identify proteins that bind the nanobody.

Nanobody ELISA

100 µL of 10 µg/mL purified nanobody in 0.1M sodium carbonate (pH 9) buffer were immobilized on a 96-well ELISA plate by incubation at 4°C overnight. Plates were then blocked with 200 µL of PBS + 2% bovine serum albumin (BSA) + 0.05% Tween 20 (blocking buffer) for 1 hr. The liquid was aspirated and 100 µL of purified TrxA-strep-doc5600 at concentrations ranging from 0 nM to 225 nM diluted in blocking buffer was added to each well. Plates were then incubated for 30 minutes at room temperature before the doc5600 solution was aspirated. Wells were then washed three times with 200 µL of wash buffer (PBS + 0.05% Tween 20). To detect bound TrxA-strep-doc5600, 100 µL of 1:5000 diluted Streptactin-Horseradish Peroxidase Conjugate (Bio-Rad Laboratories cat. 1610381) was added to each well and incubated for 1 hour at 4°C. Wells were then washed three times with wash buffer and streptactin levels were measured on a Tecan M1000 plate

reader using TMB chromogen solution following the manufacturer’s instructions (ThermoFisher Scientific cat. N301). Dissociation constants were estimated by fitting experimental data to a one-site Langmuir binding model.

Nanobody surface plasmon resonance (SPR)

SPR measurements were collected on a BIACORE 3000 (GE Healthcare) equipped with a CM5 sensor chip. Pure NbE6 at 10 μg/mL in 10 mM sodium acetate, pH 5.3 was immobilized using amine coupling chemistry to one flow cell at a density of 500 RU with an adjacent flow cell left as reference. To collect equilibrium binding data, doc5600 W28F in SPR buffer (10 mM HEPES, 150 mM NaCl, 3 mM EDTA, 0.005% Tween-20, pH 7.4) was flowed through both flow cells at 5 μL/min for 60 min. Response data were collected once every second. The surface was regenerated by flowing regeneration buffer (0.25% w/v SDS, 10% v/v glycerol, 5 mM DTT) through the flow cell at 50 μL/min for 30 seconds. These regeneration conditions were verified to maintain surface binding capacity over 5 injections (quantified as unchanging R_{max} for a single doc5600 W28F concentration). Equilibrium R_{max} values were plotted as a function of concentration and fit to a Langmuir isotherm to estimate K_d.

Yeast strains and culturing conditions

Saccharomyces cerevisiae strain EBY100 (MATa AGA1::GAL1-AGA1::URA3 ura3-52 trp1 leu2-delta200 his3-delta200 pep4::HIS3 prbd1.6R can1 GAL) (ATCC) was used for all experiments conducted in this study. The background strain was maintained in YPD medium. Strains harboring plasmid pCTcon2 were maintained in synthetic dextrose medium (2% w/v dextrose) supplemented with casamino acids lacking uracil, adenine, and tryptophan (SD-CAA) [192]. To induce surface display, cells were first grown overnight in SD-

CAA at 30°C with shaking. Following overnight incubation, cells were subcultured in SD-CAA (0.5% w/v dextrose) at an initial OD_{600} of 1. Once the culture reached an OD_{600} of 2-3, cells were resuspended in synthetic galactose medium (SG-CAA - 2% w/v galactose and 0.2% w/v dextrose) supplemented with casamino acids lacking uracil, adenine, and tryptophan at an OD_{600} of 1. Induction was typically carried out for 18-20 hours.

Combinatorial histidine library construction and transformation

Combinatorial substitution of histidines throughout the doc5600 sequence was achieved by purchasing from IDT (Coralville, IA) two overlapping oligonucleotides (Fwd ultramer and Rev ultramer in Table 4.2) with machine mixes at specific nucleotide positions. For cloning into the pCTcon2 backbone [192] – a generous gift from Dane Wittrup (Addgene plasmid #41843) - for yeast surface display, each oligo contained 48 bp of homology to pCTcon2. A simple Python program optimized the nucleotide sequence (including specific machine mix options) such that, for every mutated amino acid position, the nucleotide sequence maximizes the probability of encoding Histidine or the wild-type amino acid at that position. The full length, randomized doc5600 library was amplified by a two-step PCR using Phusion polymerase. In the first PCR (25 cycles, T_m = 64°C, extension time of 15 seconds), the two overlapping oligonucleotides, Fwd ultramer and Rev ultramer, served as primers for self-extension to produce dsDNA encoding the entire doc5600 with flanking homology to pCTcon2. The resulting 360 bp band was gel extracted and purified, serving as template DNA for a second PCR with primers SL61 and SL62 to generate microgram quantities of randomized insert for yeast transformation. Digested pCTcon2 backbone was prepared by double restriction digestion with NheI-HF and BamHI-HF enzymes as described in [193].

To prepare competent EBY100 yeast cells for transformation, 150 mL YPD in a 500 mL baffled shaker flask was inoculated at an OD_{600} of 0.5 and incubated in a 30°C shaker at 200rpm until the culture reached an OD_{600} of ~2. Cells were then harvested by centrifugation at 2000x g and 4°C for 5 minutes before being gently resuspended in 75 mL sterile 100mM lithium acetate. Once resuspended, sterile, freshly prepared 1M DTT was added to a final concentration of 10mM and the suspension was incubated in the 30°C shaker for 10 minutes. Cells were pelleted at 2000x g and 4°C for 5 minutes and the supernatant discarded. Cell pellets were then washed with 75 mL ice-cold electroporation buffer (1M sorbitol + 1mM CaCl2), pelleted at 2000x g and 4°C for 5 minutes, and resuspended by repeated pipetting in 750 uL ice-cold electroporation buffer. To each pre-chilled, 2mm electroporation cuvette was added 250 uL competent cells plus 5 ug of digested backbone and 15 ug insert (6 cuvettes in parallel). Electroporation was performed in a Bio-Rad GenePulser XCell system using a square wave protocol at 500V with a single 15ms pulse. Immediately after electroporation, 2mL of pre-warmed YPD was added to each cuvette and the contents transferred to a 14mL round bottom culture tube. Transformed cells were pooled and recovered for 1 hour at 30°C prior to pelleting at 900x g and resuspension in 1 mL pre-warmed SD-CAA media. Transformation efficiency was estimated by plating $1{:}10^5$, $1{:}10^6$, and $1{:}10^7$ dilutions and counting colonies, yielding a library size estimate of 1.3×10^8. To estimate the background population containing undigested pCTcon2 backbone, the above process was performed with digested backbone only and dilutions of 1:100 and 1:1000 were plated. The remaining transformed cells were transferred to 600 mL pre-warmed SD-CAA media (100 mL per electroporation) and shaken at 250rpm and 30°C until saturated. 10^{10} cells were passaged into 500 mL fresh SD-CAA at an OD_{600} of 0.5 and again grown to saturation. Cells were then harvested to make library aliquots containing 10^9 cells in SD-CAA + 15% glycerol for storage at -80°C.

Magnetic cell sorting (MACS)

500 mL of SD-CAA was inoculated with one library cryostock and grown overnight at 30°C. The following day, 10^9 cells were subcultured into 100 mL SD-CAA (0.5% dextrose) and grown to a culture OD_{600} of 2. 10^9 cells were then pelleted and resuspended in 100 mL SG-CAA (as prepared in [192]) before being returned to the 30°C incubator for 20 hours of induction. 1 x 10^9 induced cells were pelleted and washed twice with PBS + 0.1% BSA (PBSB) at pH 7.2, then resuspended in 1 mL PBSB, pH 7.2 with 1 µM unlabeled NbE6 and incubated with end-over-over rotation at room temperature for 30 minutes. Cells were pelleted by centrifugation at 13,000x g for 30 seconds and washed once with PBSB, pH 7.2. Cells displaying functional double dockerin variants at this point in the protocol are saturated with unlabeled NbE6. To select for pH-dependent binding, cells were then resuspended in PBSB at a lower pH and washed with low pH PBSB twice. Total exposure time to low pH buffer did not exceed 5 minutes to ensure stringent selection. Cells were transferred back to PBSB at pH 7.2 and incubated with a pre-mixed mixture of biotinylated NbE6 (final concentration 500 nM) and 10 µL per 10^9 cells of Pierce streptavidin beads (ThermoFisher Scientific cat. 88816) for 2 minutes. A DynaMag^{TM}-2 magnet (ThermoFisher Scientific cat. 12321D) was used to separate and wash the beads and bound cells. Three rounds of washes with 500 µL PBSB, pH 7.2 were performed before the beads were resuspended in 500 µL SD-CAA and transferred to 15 mL SD-CAA with ampicillin (100 µg/mL). Remaining library diversity was estimated by plating 1:10^4 and 1:10^5 dilutions to be 1.5 x 10^7 variants after round 1.

Subsequent rounds of MACS were scaled down to screen at least ~10x the estimated library diversity. In Round 2, 10^8 cells were screened and 10^7 cells were screened in each subsequent round. To select for tighter NbE6 binding in addition to pH-sensitivity, the concentration of biotinylated NbE6 and the volume of streptavidin beads were sequentially lowered over 4 rounds of screening. The concentration of biotinylated NbE6 dropped from 500 nM to 250 nM, 125 nM, and finally 50 nM in the final screening round. The volume of streptavidin beads was decreased from 10 μL to 7 μL to 4 μL to 2 μL.

LASSO Regression analysis of flow cytometry data

Sequences with measured Hill coefficient, pH_0, and binding dynamic range were featurized by recording, for each of the 22 library mutation positions, 1 if the sequence contained His at that position and 0 if not. The total number of histidines was included as an additional independent variable, and the predicted response variable was either dynamic range or Hill coefficient. LASSO regression was performed in Python with scikit-learn with $\alpha = 0.1$.

Nanobody-double dockerin complex model generation

Atomistic structures for the nanobody and doc5600 were generated by RoseTTaFold [194] with high confidence since both sequences have homologs in the PDB (2j4m and 4gft). Since doc5600 comprises two folded dockerin domains connected by a flexible linker, the top RoseTTaFold doc5600 model was subjected to 500 ns of implicit solvent temperature REMD to identify dominant alternative doc5600 configurations that may better match how doc5600 looks when bound to NbE6. REMD simulations were performed using in house developed software [195] run on the OpenMM MD engine [196]. 20 replicas with temperatures ranging from 300 – 450K were simulated using the ff14SBonlysc forcefield with the igb8 implicit solvent model, no SASA force field term, and no non-bonded interaction cutoff. 5 swap attempts between neighboring replicas were performed every 200 ps. Replicas were propagated with constant NVT using a 2 fs timestep. Constant temperature was maintained using an Andersen thermostat with velocities randomized every 500 steps. Bonds involving H-atoms were constrained with SHAKE. All other parameters were set to default values.

The first 490 ns of simulation time were used for equilibration, and representative equilibrium doc5600 structures were obtained by analyzing the final 10 ns of simulation with an agglomerative k-means clustering algorithm with an inter-cluster distance cutoff of 2.0 Å.

Complex models were generated by docking doc5600 to the nanobody using ClusPro in antibody mode, with residues not in the three complementary determining regions (CDRs) of the nanobody appropriately masked [181,182]. The most populated cluster's centroid structure was saved for further analysis by molecular dynamics.

Nanobody-double dockerin interface prediction

Explicit solvent simulations of nanobody-double dockerin complex models were performed using OpenMM [196] on GPUs. Systems were set up using tleap to generate Amber topology and coordinate files [197]. Protein systems were solvated in periodic boxes with rigid, explicit water such that the minimum distance between a solute atom and the box edge was 8.0 Å. Na+ or Cl- atoms were added to neutralize the system charge. NPT Simulations at 300K and 1 atm were performed using the FF14SB force field [198] with TIP3P water model [199], a 2 fs time step, and a 1 nm non-bonded interaction cutoff. A Langevin integrator with a time constant of 1 ps^{-1} and a Monte Carlo barostat applied every 125 time steps were used to maintain constant temperature and pressure. Electrostatic forces were computed using PME.

To keep the double dockerin and nanobody close in space, a parabolic restraint with a force constant of 1 kcal/mol/Å^2 was applied to the centroids of both proteins if the distance between centroids exceeded the sum of the proteins' radii of gyration. Systems were minimized using a conjugate gradient minimization algorithm then equilibrated for 200 time steps at 300K before production NPT MD simulations of 200 ns were performed in triplicate.

Ranked lists of double dockerin residues most likely involved in nanobody binding were obtained by measuring contact frequencies from explicit solvent simulations of three different complex models generated by ClusPro. Contact frequencies between doc5600 and nanobody residues were computed from the last 50 ns of every explicit solvent simulation trajectory using Cpptraj. Here, a contact was counted in a frame if the centers of mass of two amino acids were within 3.5 angstroms. The results for all simulations were combined to produce a rank-ordered list of doc5600 residues ranked by their measured contact frequencies with any nanobody residue. The complex model that contained the most highly ranked contacts served as our best model for the double dockerin-nanobody complex.

Production of soluble doc5600 variants

Point mutants of doc5600 were cloned into the pET-32a backbone using overlap extension PCR as described in [200] using primers SL31-SL38. PCR products were transformed into E. coli BL21 for protein expression. C65.4 and c75.5 were subcloned from the respective pCTcon2 plasmids into pET-32a using primers SL87-SL92. To produce protein, LB media was inoculated at a starting OD_{600} of 0.05 and incubated at 37°C with shaking at 250 rpm. When the culture reached an OD_{600} of 0.5, IPTG was added to a final concentration of 0.5 mM and the culture was incubated for an additional 18 hours at 30°C before harvesting. Cells were harvested by centrifuging in a swing bucket rotor at 4000x g for 10 minutes before resuspension in PBS plus 10 mM imidazole (pH 7.4) with Pierce protease inhibitor following the manufacturer's instructions. 0.5 mm zirconia beads were added to 10% the liquid volume and the resuspended cells plus beads were vortexed for 30 seconds 10 times to lyse cells. Cell debris was pelleted by centrifugation in a fixed angle rotor at 10,000x g for 10 minutes. Proteins were then purified from the soluble cell lysate by immobilized metal affinity chromatography following the manufacturer's instructions. Typical protein yield was about 3 mg/L culture for 50 mL cultures in 250 mL Erlenmeyer flasks.

Flow cytometry analysis of libraries and isolated clones

0.2 OD_{600}*mL of induced cells were washed once with 500 μL of PBSB then incubated in 100 μL PBSB containing biotinylated NbE6 at the stated concentration (between 20 and 100 nM) and anti-c-myc DyLight 488 antibody (Fisher Scientific cat. 5081547) at 1:100 dilution for 30 minutes at room temperature. Cells were then washed once with 500 μL PBSB at the same pH and incubated for 15 minutes on ice with 100 μL PBSB plus streptavidin PE-TR conjugate (ThermoFisher Scientific cat. SA1017) in a 1:100 dilution. Labeled cells were washed twice with 500 μL PBSB before analysis on a Becton Dickson FACSAria flow cytometer with a 488nm laser. Fluorescence in the 520 nm and 610 nm channels measured double dockerin display labels and bound biotinylated NbE6 respectively. Data were analyzed using the FACSDiva software.

Generation of point mutants of c75.5 and c65.4

Synthetic dsDNA encoding point mutants of clone c75.5 and c65.4, listed in Table 4.2, were purchased from Twist Bioscience (South San Francisco, CA). As in the Ultramers^{TM}

used for yeast surface display library construction, 48 bp homology to pCTcon2 was added to each end of the c75.5 and c65.4 sequences to facilitate cloning into EBY100 via homologous recombination. Synthetic dsDNA was co-transformed with NheI-BamHI digested pCTcon2 into EBY100. Sequences were verified by colony PCR and Sanger sequencing prior to experimental characterization.

96-well plate-based analysis of isolated clones

0.3 OD_{600}*mL of induced cells were added to each well and washed twice with 100 μL of PBSB at the desired pH by centrifuging at 4°C at 3000 rpm for 3 minutes and resuspending. Washed cells were then incubated with 100 μL of PBSB at the desired pH with 50 nM biotinylated NbE6 and anti-c-myc DyLight 488 antibody (Fisher Scientific cat. 5081547) at 1:200 dilution for 30 minutes at room temperature with shaking. Cells were then washed once with 200 μL PBSB at the desired pH and incubated for 15 minutes with 100 μL PBSB plus streptavidin PE-TR conjugate (ThermoFisher Scientific cat. SA1017) in a 1:100 dilution. Labeled cells were washed twice with 200 μL PBSB at the desired pH before resuspension in 150 uL PBSB and analyzed on a Tecan M1000 plate reader. Fluorescence signals at 488 nm excitation, 515 nm emission to quantify double dockerin display level and 488 nm excitation, 615 nm emission to quantify NbE6 binding were measured. A normalized fluorescence quantity computed as $\mathrm{F_{normalized}} = (\mathrm{F_{615nm}} - \mathrm{F_{615nm\,,blank}})/\mathrm{F_{515nm}}$ was used to measure NbE6 binding vs pH. $F_{615nm,blank}$ was measured from cells only labeled with Strepavidin PE-TR to subtract out signal from non-specific Streptavidin binding and autofluorescence. “Wild type” dockerin mutants were recharacterized each trial to enable comparison across trials.

Tethered NbE6 scaffoldin MD simulation

All-atom simulations were performed using an in house replica exchange program with OpenMM as the MD engine with the ff14SBonlysc forcefield and igb8 implicit solvation parameter set with no solvent-accessible surface area (SASA) force component. As flexible, multi-domain proteins, scaffoldins do not resemble the globular proteins for which default simulation parameters are optimized. To identify optimal simulation conditions for scaffoldin simulation, we tested parameter combinations on a model mini-scaffoldin from *Clostridium thermocellum* for which published experimental radius of gyration and inter-cohesin domain distance data exists [184]. The best agreement between simulation and experiment occurred performing temperature replica exchange implicit solvent simulations with zero surface tension (no SASA force).

To measure doc5600 binding site availability of tethered NbE6 scaffoldin molecules, 200ns trajectories from replica exchange molecular dynamics (REMD) simulations were post-processed as follows. A model of the NbE6-doc5600 complex was aligned to each NbE6 domain in each frame of the trajectory. Then, a custom-coded Fortran script looped through the trajectory to check for atom overlaps, which we defined as distances <7 angstroms, between doc5600 atoms and all other atoms in the simulation. Binding site availability was computed as the fraction of frames with no overlap.

Coarse-grained MD parameter optimization and simulation

The all-atom system for coarse-grained parameter optimization comprised a capped 24-mer Gly/Ser linker. A reference trajectory for parameter fitting was generated from a 100ns implicit solvent REMD run of this system, with configurations taken from the 310K replica. All-atom simulations were performed using an in house replica exchange program with

OpenMM as the MD engine. The ff14SBonlysc forcefield and igb8 implicit solvation parameter set were used.

The coarse-grained forcefield was kept very simple – all amino acids were represented by a single site type (1:1 amino acid to CG site mapping) with three interatomic forces – bond length, bond angle, and 1-4 Lennard Jones (LJ). Such simple force field mappings were shown to model a bacterial cellulosome system well [201] and should apply well here since Gly/Ser peptides resemble ideal polymer chains. Bond length and bond angle forces were parametrized as parabolic restraints with a force constant and set point, while the 1-4 LJ was a standard 6-12 potential with length scale σ and energy scale ε. The relative entropy optimized force field parameters were as follows:

Bond length – $k = \frac{586\frac{kJ}{mol}}{nm^2}, x_0 = 0.398\,nm$; Bond angle – $k = \frac{0.187\frac{kJ}{mol}}{Degree^2},\ \theta_0 = 120°$; $1-4\,LJ$ – $\sigma = 0.221\,nm, \varepsilon = 4.60\,kJ/mol.$

To keep the NbE6 domains folded, pairwise restraints between non-CDR sites within each domain were added to the system with $k = 5\frac{kJ}{mol}/nm^2$. Pairwise restraints among folded domains within doc5600 and GH5-doc5600 were also added. These restraints did not include sites mapping to positions with no secondary structure (linkers between GH5 and doc5600 domains or between dockerin domains within doc5600). Since we were interested in simulating TetNb scaffoldins with bound dockerins, attractive potentials were also added between predicted interface residues of the double dockerin and NbE6 with $k = 2\frac{kJ}{mol}/nm^2$.

To collect simulation data, coarse-grained systems were first energy-minimized in OpenMM before production simulations at 320K for 2 microseconds using OpenMM.

Native mass spectrometry

Samples were buffer exchanged in 200 mM ammonium acetate buffer pH 7.1 using Zeba Spin Desalting Columns (7k MWCO, 75 µl, Thermo Fisher). TetNb-36 and a double dockerin-containing GH5 (GH5-Doc) were mixed in equimolar amounts at 10 µM and incubated at room temperature for one hour before buffer exchange. All native MS data was acquired on a Waters Synapt G2s-i ion mobility time-of-flight mass spectrometer. The proteins were loaded into hand-pulled borosilicate glass capillaries (Sutter Instrument). Nanoelectrospray voltage (1.0-1.1 kV) was applied through a Pt wire inserted into the capillary. MassLynx v4.1 (Waters) was used to manually analyze spectra and mass deconvolution was performed using UniDec version 4.3.0 [202].

SDS-PAGE and Native PAGE

Proteins in 1x Laemmli buffer containing 2% SDS and X mM DTT were boiled at 95°C for 10 minutes before being loaded on a 4-15% Tris-glycine gel. Gels were run at 110V for approximately 1.5 hours and stained with Coomassie blue to visualize bands. For native PAGE, samples were prepared in 1x Laemmli buffer with no SDS or DTT and loaded on a 4-15% Tris-glycine gel. Native gels were run with the gel tank on ice at 150V for 1.5 hours and the gels were stained with Coomassie blue to visualize bands.

Plate-based equilibrium binding affinity measurements of soluble dockerins

NbE6-GFP, WT doc5600, c65.4, and c75.5 were produced and purified by IMAC as described above. AF647 fluorophores were conjugated to dockerins using an AF647 NHS ester (Fisher Scientific cat no. A37573). Degrees of conjugation were ~1:1 dye:protein molecule for C65.4 and c75.5 and 2:1 for WT doc5600 as measured by UV-Vis with theoretical extinction coefficients.

Black 96-well plate wells were coated overnight with 100 μL 0.1M sodium carbonate buffer (pH 9.4) of the respective immobilized proteins for each experimental setup at 0.25 μg/mL when NbE6 was immobilized and 1 μg/mL when dockerins were immobilized. Wells were then blocked with 200 μL blocking buffer (PBS + 2% BSA) at room temperature for one hour. After aspirating blocking buffer, wells were washed once with wash buffer (PBS + 0.1% BSA) at the desired pH, pH 7.4 or pH 5.0. Wells were then loaded with 200 μL dockerin or NbE6-GFP in wash buffer at the desired pH at a range of concentrations and incubated on a plate shaker at room temperature for one hour. Wells were then washed 3x with wash buffer at the respective pH and resuspended in 150 μL wash buffer before fluorescence measurements on a Tecan M1000 Infinite plate reader. For NbE6-GFP detection, excitation at 488 nm and emission at 515 nm; for AF647, ex. 651 nm, em. 672 nm.

Isothermal calorimetry of dockerin variants and NbE6

TrxA-tagged C75.5 or doc5600 at 3 μM concentration in PBS at pH 7.4 or pH 5.5 were supplied to the sample cell of a TA NanoITC. The buret syringe was loaded with 30 μM NbE6 dialyzed in the same buffer as the dockerins for respective titrations. Each titration was performed as 20 injections of 5 μL NbE6, with stirring at 250 rpm and the sample cell kept at 25°C. To eliminate signal from the heat of dilution, a blank run at pH 5.5 and pH 7.4 was performed with 30 μM NbE6 injected into the sample cell containing PBS with no dockerin. Peaks in the blank-subtracted raw data were manually integrated and titration curves were fit to an independent, multiple binding site model using the NanoAnalyzer software.

Primers and plasmids used in this work

Table 4.2. Primer sequences used in this work

Name	Sequence	Description
SL61	ACGATTGAAGGTAGATACCC	FWD primer for amplifying doc5600 mutants from pCT-con2 for yeast surface display library sequencing
SL62	ACAGTGGGAACAAAGTCG	REV primer for amplifying doc5600 mutants from pCT-con2 for yeast surface display library sequencing
Fwd Ultramer	agtggtggaggaggctctggtggaggcggtagcggaggcggagggtcgATGAAATGTYRKGCAMMCTCTCWCGGAYACCCATGTTGTAAAGAAGCTMACCCAATAATATTCTACMAMSACSAMSACGGTGATTGGGGTATTGAAAATMACTCCTGGTGTGGAATTATTMAMMACGAAAAACCAGCATGTAGTGAA	Fwd ultramer for His library construction using homologous recombination for pCT-Con2 ligation
Rev Ultramer	ctcgagctattacaagtcctcttcagaaataagcttttgttcggatccAAGCTTATTTAATATTCCACAMYRGTKGTKATTTTCGACACCCCATTCACCACTKTSGTSGKKGTRGTRAACTTGAGTTTCTTTTGAACAACATGGGTAGCCKTGATTAATGATKTKTTCACTACATGCTGGTTTTTCGT	Rev ultramer for His library construction using homologous recombination for pCT-Con2 ligation
SL31	CAAGGCTGAAGACGGTGATTGGGGTATTGAAAATAATTCCTGGTG	Mutate D24A in doc5600 double dockerin sequence (FWD)
SL32	ACCGTCTTCAGCCTTGTAGAATATTATTGGATTAGCTTCTTTACAACATGG	Mutate D24A in doc5600 double dockerin sequence (REV)
SL33	AAGTTTATGCCACTGATGAAAGTGGTGAATGGGGTG	Mutate Y67V in doc5600 double dockerin sequence (FWD)
SL34	ATCAGTGGCATAAACTTGAGTTTCTTTTGAACAACATGGG	Mutate Y67V in doc5600 double dockerin sequence (REV)
SL35	TACACTGCTGAAAGTGGTGAATGGGGTGTCGAAAATAAC	Mutate D69A in doc5600 (FWD)
SL36	CACTTTCAGCAGTGTAATAAACTTGAGTTTCTTTTGAACAACATG	Mutate D69A in doc5600 (REV)

SL37	ATATTCGTCAAGGATGAAGACGGTGATTGGGG	Mutate Y22V in doc5600 (FWD)
SL38	CATCCTTGACGAATATTATTGGATTAGCTTCTTTACAAC	Mutate Y22V in doc5600 (REV)
SL87	CCATGGCTGATATCGGATCCATGAAATGTTGGGCAACCTC	c75.5_FWD for cloning c75.5 variant of doc5600 from pCTcon2 into pET-32a
SL88	TCTTAGGATCCAAGCTTATTTAATATTCC	c75.5_REV for cloning c75.5 variant of doc5600 from pCTcon2 into pET-32a
SL89	AATAAGCTTGGATCCTAAGAATTCGAGCTCCGTCG	pET_c75.5_FWD for cloning c75.5 variant of doc5600 from pCTcon2 into pET-32a
SL90	GAGGTTGCCCAACATTTCATGGATCCGATATCAGCCATGG	pET_c75.5_REV for cloning c75.5 variant of doc5600 from pCTcon2 into pET-32a
SL91	CCATGGCTGATATCGGATCCATGAAATGTTGGGCACACTC	c65.4_FWD for cloning c65.4 variant of doc5600 from pCTcon2 into pET-32a
	Same as c75.5 REV (SL88)	c65.4_REV for cloning c65.4 variant of doc5600 from pCTcon2 into pET-32a
SL92	GAGTGTGCCCAACATTTCATGGATCCGATATCAGCCATGG	pET_c65.4_REV for cloning c65.4 variant of doc5600 from pCTcon2 into pET-32a
	Same as pET_c75.5_FWD (SL89)	pET_c65.4_FWD for cloning c65.4 variant of doc5600 from pCTcon2 into pET-32a
C65.4_H42N	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCACACTCTCTCGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACCACGACGACGACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACAACGAAAAACCAGCATGTAGTGAAAACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTTACTACAACGACGACAGTGGTGAATGGGGTGTCGAAAATAACAACTATTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C65.4_H54N	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCACACTCTCTCGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACCACGACGACGACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAAAACATCATTAATAACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTTACTACAACGACGACAGTGGTGAATGGGGTGTCGAAAATAACAACTATTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C75.5_H42N	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCAACCTCTCACGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACAAAGACGAACACGGTGATTGGGGTATTGAAAATAACTCCTG	

	GTGTGGAATTATTAACAACGAAAAACCAGCATGTAGTGAACACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTCACCACACCGACCACAGTGGTGAATGGGGTGTCGAAAATAACCACCAGTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C75.5_H66Y	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCAACCTCTCACGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACAAAGACGAACACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAACACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTTACCACACCGACCACAGTGGTGAATGGGGTGTCGAAAATAACCACCAGTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C75.5_H80N	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCAACCTCTCACGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACAAAGACGAACACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAACACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTCACCACACCGACCACAGTGGTGAATGGGGTGTCGAAAATAACAACCAGTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C65.4_H42N	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCACACTCTCTCGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACCACGACGACGACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACAACGAAAAACCAGCATGTAGTGAAAACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTTACTACAACGACGACAGTGGTGAATGGGGTGTCGAAAATAACAACTATTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C65.4_H54N	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCACACTCTCTCGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACCACGACGACGACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAAAACATCATTAATAACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTTACTACAACGACGACAGTGGTGAATGGGGTGTCGAAAATAACAACTATTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C75.5_H42N	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCAACCTCTCACGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACAAAGACGAACACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACAACGAAAAACCAGCATGTAGTGAACACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTCACCACACCGACCACAGTGGTGAATGGGGTGTCGAAAATAACCACCAGTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C75.5_H66Y	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCAACCTCTCACGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACAAAGACGAACACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAACACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTTACCACACCGACCACAGTGGTGAATGGGGTGTCGAAAATAACCACCAGTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C65.4_H6T	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCAACCTCTCTCGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACCACGACGACGACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAAAACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTTACTACAACGACGACAGTGGTGAATGGGGTGTCGAAAATAACAACTATTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	

C75.5_H26D	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCAACCTCTCACGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACAAAGACGAAGACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAACACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTCACCACACCGACCACAGTGGTGAATGGGGTGTCGAAAATAACCACCAGTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C65.4_H10Y	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCACACTCTCTCGGATACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACCACGACGACGACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAAAACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTTACTACAACGACGACAGTGGTGAATGGGGTGTCGAAAATAACAACTATTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C75.5_H50K	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCAACCTCTCACGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACAAAGACGAACACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAAAAGATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTCACCACACCGACCACAGTGGTGAATGGGGTGTCGAAAATAACCACCAGTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C65.4_H23K	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCACACTCTCTCGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACAAGGACGACGACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAAAACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTTACTACAACGACGACAGTGGTGAATGGGGTGTCGAAAATAACAACTATTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C75.5_H67Y	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCAACCTCTCACGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACAAAGACGAACACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAACACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTCACTACACCGACCACAGTGGTGAATGGGGTGTCGAAAATAACCACCAGTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C65.4_H54N-H42N	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCACACTCTCTCGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACCACGACGACGACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACAACGAAAAACCAGCATGTAGTGAAAACATCATTAATAACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTTACTACAACGACGACAGTGGTGAATGGGGTGTCGAAAATAACAACTATTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C75.5_H70E	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCAACCTCTCACGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACAAAGACGAACACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAACACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTCACCACACCGACGAAAGTGGTGAATGGGGTGTCGAAAATAACCACCAGTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C65.4_H54N-H10Y	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCACACTCTCTCGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACCACGACGACGACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAAAACATCATTAATAACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTTACTACAACGACGACAGTGGTGAATGGGGT	

	GTCGAAAATAACAACTATTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C75.5_H66Y-H67Y	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCAACCTCTCACGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACAAAGACGAACACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAACACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTTACTACACCGACCACAGTGGTGAATGGGGTGTCGAAAATAACCACCAGTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C65.4_H54N-H23K	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCACACTCTCTCGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACAAGGACGACGACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAAAACATCATTAATAACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTTACTACAACGACGACAGTGGTGAATGGGGTGTCGAAAATAACAACTATTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C75.5_H8L-H26D	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCAACCTCTTTGGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACAAAGACGAAGACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAACACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTCACCACACCGACCACAGTGGTGAATGGGGTGTCGAAAATAACCACCAGTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	
C75.5_H8L	CAATCCGCCCTCACTACAACCGAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGATGAAATGTTGGGCAACCTCTTTGGGACACCCATGTTGTAAAGAAGCTAACCCAATAATATTCTACAAAGACGAACACGGTGATTGGGGTATTGAAAATAACTCCTGGTGTGGAATTATTAACCACGAAAAACCAGCATGTAGTGAACACATCATTAATCACGGCTACCCATGTTGTTCAAAAGAAACTCAAGTTCACCACACCGACCACAGTGGTGAATGGGGTGTCGAAAATAACCACCAGTGTGGAATATTAAATAAGCTTGGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCGAGCTACTCTGGCGTCGATGAGGGA	

Table 4.3. Plasmids used in this study.

Plasmid name	Description
pMES4-NbE6	Production of 6x-His-tagged NbE6
pET-32a-strep-doc5600	Production of 6x-His and Strep-tagged doc5600
pET-32a-strep-doc5600 W28F-W74F	Production of 6x-His and Strep-tagged doc5600 W28F-W74F
pET-32a-strep-doc5600 Y22V	Production of 6x-His and Strep-tagged doc5600 Y22V
pET-32a-strep-doc5600 Y67V	Production of 6x-His and Strep-tagged doc5600 Y67V
pET-32a-strep-doc5600 D69A	Production of 6x-His and Strep-tagged doc5600 D69A
pET-32a-strep-c65.4	Production of 6x-His and Strep-tagged c65.4
pET-32a-strep-c75.5	Production of 6x-His and Strep-tagged c75.5
pMES4-TetNb-36	Production of Tethered Nb with 36mer Gly/Ser linker and 6x-His tag
pMCS068-91-536	Production of GH5-Doc2x with 6x His tag
pCTcon2	Yeast surface display plasmid

pCTcon2-c75.5	Yeast surface display plasmid with doc5600 variant c75.5
pCTcon2-c65.4	Yeast surface display plasmid with doc5600 variant c65.4

Table 4.4. Amino acid sequences of proteins characterized in this study.

Name	Sequence
NbE6	MAQVQLQESGGGLVQAGGSLRLSCAASGRSFSLYRMAWFRQAPGKEREEREFVGAIQSNGARTYYADSVKDRFTISRDNAKNTVYLQMNSLKPEDTAVYYCAADQRPGLGLSRTGWGDFSSWGQGTQVTVSSHHHHHH
WT doc5600	MKCWATSLGYPCCKEANPIIFYKDEDGDWGIENNSWCGIIKNEKPACSEKIINQGYPCCSKETQVYYTDESGEWGVENNNWCGILN
Doc5600 W28F	MKCWATSLGYPCCKEANPIIFYKDEDGD**F**GIENNSWCGIIKNEKPACSEKIINQGYPCCSKETQVYYTDESGE**F**GVENNNWCGILN
Doc5600 Y22V	MKCWATSLGYPCCKEANPIIF**V**KDEDGDWGIENNSWCGIIKNEKPACSEKIINQGYPCCSKETQVYYTDESGEWGVENNNWCGILN
Doc5600 Y67V	MKCWATSLGYPCCKEANPIIFYKDEDGDWGIENNSWCGIIKNEKPACSEKIINQGYPCCSKETQVY**V**TDESGEWGVENNNWCGILN
Doc5600 D69A	MKCWATSLGYPCCKEANPIIFYKDEDGDWGIENNSWCGIIKNEKPACSEKIINQGYPCCSKETQVYYT**A**ESGEWGVENNNWCGILN
Single dockerin	MKQGYKCCSSTNCTIVFTDNDGTWGVENNQWCGISNDCKLAAA
c75.5	MKCWATSHGHPCCKEANPIIFYKDEHGDWGIENNSWCGIINHEKPACSEHIINHGYPCCSKETQVHHTDHSGEWGVENNHQCGILN
c75.5 H8L	MKCWATS**L**GHPCCKEANPIIFYKDEHGDWGIENNSWCGIINHEKPACSEHIINHGYPCCSKETQVHHTDHSGEWGVENNHQCGILN
c75.5 H10Y	MKCWATSHG**Y**PCCKEANPIIFYKDEHGDWGIENNSWCGIINHEKPACSEHIINHGYPCCSKETQVHHTDHSGEWGVENNHQCGILN
c75.5 H26D	MKCWATSHGHPCCKEANPIIFYKDE**D**GDWGIENNSWCGIINHEKPACSEHIINHGYPCCSKETQVHHTDHSGEWGVENNHQCGILN
c75.5 H42N	MKCWATSHGHPCCKEANPIIFYKDEHGDWGIENNSWCGIIN**N**EKPACSEHIINHGYPCCSKETQVHHTDHSGEWGVENNHQCGILN
c75.5 H66Y	MKCWATSHGHPCCKEANPIIFYKDEHGDWGIENNSWCGIINHEKPACSEHIINHGYPCCSKETQV**Y**HTDHSGEWGVENNHQCGILN
c75.5 H67Y	MKCWATSHGHPCCKEANPIIFYKDEHGDWGIENNSWCGIINHEKPACSEHIINHGYPCCSKETQVH**Y**TDHSGEWGVENNHQCGILN
c75.5 H70E	MKCWATSHGHPCCKEANPIIFYKDEHGDWGIENNSWCGIINHEKPACSEHIINHGYPCCSKETQVHHTD**E**SGEWGVENNHQCGILN
c75.5 H80N	MKCWATSHGHPCCKEANPIIFYKDEHGDWGIENNSWCGIINHEKPACSEHIINHGYPCCSKETQVHHTDHSGEWGVENN**N**QCGILN
c65.4	MKCWAHSLGHPCCKEANPIIFYHDDDGDWGIENNSWCGIINHEKPACSENIINHGYPCCSKETQVYYNDDSGEWGVENNNYCGILN
c65.4 H6T	MKCWA**T**SLGHPCCKEANPIIFYHDDDGDWGIENNSWCGIINHEKPACSENIINHGYPCCSKETQVYYNDDSGEWGVENNNYCGILN
c65.4 H10Y	MKCWAHSLG**Y**PCCKEANPIIFYHDDDGDWGIENNSWCGIINHEKPACSENIINHGYPCCSKETQVYYNDDSGEWGVENNNYCGILN
c65.4 H23K	MKCWAHSLGHPCCKEANPIIFY**K**DDDGDWGIENNSWCGIINHEKPACSENIINHGYPCCSKETQVYYNDDSGEWGVENNNYCGILN
c65.4 H42N	MKCWAHSLGHPCCKEANPIIFYHDDDGDWGIENNSWCGIIN**N**EKPACSENIINHGYPCCSKETQVYYNDDSGEWGVENNNYCGILN

c65.4 H54N	MKCWAHSLGHPCCKEANPIIFYHDDDGDWGIENNSWCGIINHEKPACSENIIN**N**GYPCCSKETQVYYNDDSGEWGVENNNYCGILN
c65.4 H42N-H54N	MKCWAHSLGHPCCKEANPIIFYHDDDGDWGIENNSWCGIIN**N**EKPACSENIIN**N**GYPCCSKETQVYYNDDSGEWGVENNNYCGILN
c65.4 H23K-H54N	MKCWAHSLGHPCCKEANPIIFY**K**DDDGDWGIENNSWCGIINHEKPACSENIIN**N**GYPCCSKETQVYYNDDSGEWGVENNNYCGILN
c76.1	MKCWANSLGHPCCKEANPIIFYHDEDGDWGIENHSWCGIIHHEKPACSENIINHGYPCCSKETQVHYTDDSGEWGVENNNYCGILN
c76.2	MKCWATSLGHPCCKEANPIIFYKDEHGDWGIENNSWCGIINNEKPACSEHIINHGYPCCSKETQVYHHDHSGEWGVENNHYCGILN
c76.3	MKCWANSLGHPCCKEAHPIIFYHDDHGDWGIENNSWCGIIQNEKPACSENIINHGYPCCSKETQVHHTDDSGEWGVENNNWCGILN
c75.1	MKCWANSLGHPCCKEANPIIFYKDQDGDWGIENNSWCGIINNEKPACSENIINHGYPCCSKETQVHHTDHSGEWGVENHHHCGILN
c65.1	MKCWANSLGHPCCKEANPIIFYHHEDGDWGIENNSWCGIIKHEKPACSEHIINHGYPCCSKETQVHHPHQSGEWGVENHNWCGILN
c65.6	MKCWANSLGHPCCKEANPIIFYNDEHGDWGIENHSWCGIIKNEKPACSENIINHGYPCCSKETQVHYTDDSGEWGVENNNHCGILN

V. Heterologous expression and characterization of anaerobic fungal enzymes

5.1 Introduction

Drastic drops in the price of DNA sequencing have spurred a dramatic acceleration in the rate at which biologists acquire novel gene sequences, causing an explosion in the numbers of gene annotations across public databases. These novel sequences are of great interest for application-oriented “bioprospecting” – the idea that mining DNA sequences from uncharacterized microbes or microbial communities may yield protein variants with better properties, such as enzymes with higher pH or thermal stability. The continued accumulation of annotated sequence data also presents an interesting opportunity to identify evolutionary relationships among protein variants from different species, perhaps towards inferring more fundamental principles of how sequence evolution imparts proteins with novel or improved functions.

However, a major limitation of sequence-based analysis is that an organism's (or community's) annotated genome (or metagenome) only presents the "genetic potential" of that organism or community and gives us no information about any biochemical properties of the encoded enzymes, a crucial aspect of biocatalysis-oriented bioprospecting in particular. Furthermore, sequence annotation pipelines are very much subject to misannotation, especially when the novel sequences being annotated share little similarity with those used to build the annotation tools [203]. Thus, it is critical that new protein sequences be supplemented with biochemical characterization, both to affirm their annotation and to generate insights into sequence-property relationships governing protein function.

Anaerobic fungi are a perfect example of a class of organisms with high genetic potential for lignocellulose valorization for which we have very little biochemical understanding of how that genetic potential enables rapid biomass degradation *in vivo*. As the above chapters show, there is clear proof that anaerobic fungi excel at hydrolyzing biomass, and a major contribution of this thesis is the most detailed biochemical characterization of fungal cellulosomes to date, which lends significant insight into how these organisms turn recalcitrant biomass into sugars. A major challenge to extending the resolution beyond enzyme complexes to individual enzymes is that they often express very poorly in the model organisms *E. coli* and *S. cerevisiae* [204]. As a result, we still fail to understand exactly what individual fungal cellulosome enzymes do, and characterizing individual cellulosome components in isolation would greatly aid the success of efforts in bioprospecting anaerobic fungal genomes for industrial bioprocessing.

In this chapter, I detail two cases in which we characterize enzymes from two different families, a GH5 endoglucanase and a spore coat CotH, that incorporate into fungal cellulosomes. In the GH5 case, we present the kinetics and structure of an enzyme from a family abundant in fungal cellulosomes, providing a starting point for protein engineering towards improving this enzyme's properties for biocatalytic applications. In the CotH case, we show this class of proteins, predicted to have protein kinase activity, do indeed act as protein kinases, seeding an interesting research direction focused on understanding how these proteins participate in fungal cellulosome biology.

5.2 Results and Discussion

5.2.1 Structure and kinetics of Piromyces finnis CelD

Overall Structure of the CelD Catalytic Domain

The unliganded structure of the wild-type catalytic domain and its inactive E154A mutant in complex with cellotriose were determined by molecular replacement and refined to 2.5 Å and 1.8 Å resolution, respectively (Table 5.1 and Fig. 5.1). The crystals of the apo-form and the complex belonged to orthorhombic $P2_12_12_1$ space group and contained two domain molecules per asymmetric unit. Almost all amino acid residues of the CelD catalytic domain, with the exception of a few side chains and three C-terminal residues, were traceable in the final electron density map. The overall root mean square deviation (RMSD) between apo wild-form and inactive E154A-ligand complex protein models was 0.26 Å for 359/362 Cα pairs, demonstrating high overall similarity and illustrating that substrate binding results in little to no conformational change (Fig. 5.1). The CelD catalytic domain displays strong structural similarity to GH clan A, a group of 28 unique GH families exhibiting a $(\beta/\alpha)_8$–barrel fold in structure, with the highest sequence similarity to proteins from GH5 subfamily 4 (GH5_4), a family of enzymes that predominantly display endoglucanase activity (EC 3.2.1.4) (Fig. 5.1a) [205–207]. In addition to the eight core β/α elements, CelD has another three small helixes located between α_4/β_V, β_V/α_5, and β_{VI}/α_6 secondary structure elements and two short β-strands located on the loop between C-terminal β_{VIII}/α_8 elements (Fig. 5.2).

Table 5.1. Data collection and refinement statistics for crystal structure determination

	Apo wild-type CelD	**E154A CelD-cellotriose complex**
Data collection		
Wavelength (Å)	0.9792	0.9792
Resolution range (Å)	50.00 - 2.45 (2.49 - 2.45)[a]	50.00 – 1.80 (1.83-1.80)
Space group	$P2_12_12_1$	$P2_12_12_1$
Unit cell (*a, b, c*) (Å)	69.726, 81.850, 133.259	69.495, 81.007, 131.922
Total reflections	296,064	421,385
Unique reflections	28,052 (1,348)	69,108 (3,429)
Multiplicity	10.6 (6.4)	6.1 (5.9)
Completeness (%)	99.2 (98.3)	98.9 (99.2)
$\langle I\rangle/\sigma(I)$	14.5 (2.1)	21.2 (1.4)
R_{merge}[b]	0.044 (1.741)	0.1000 (1.578)
Refinement		
Resolution range (Å)	49.36 – 2.46 (2.55 – 2.46)	30.00 – 1.80 (1.82 – 1.80)
R_{work}/R_{free}[c] (%)	18.8(23.5) /23.0(28.2)	17.6(31.2)/20.4(35.3)
Number of non-hydrogen atoms[d]	5918	6391
Protein	5784	5834
Ligands	40	68
Waters	94	489
Clash score	3.08	1.56
Rotamer outlier (%)	1.84	1.48
RMS (bonds, Å)	0.002	0.004
RMS (angles, °)	0.413	0.658
Ramachandran favored (%)	95.36	95.83
Ramachadran allowed (%)	4.5	4.17
Ramachandran outliers (%)	0.14	0.0
Average B-factor (Å^2)	44.3	34.2
Macromolecules	44.4	33.8
Ligands	47.5	34.2
Waters	37.7	38.4
PDB code	8HGX	8HGY

[a] Statistics for the highest-resolution shell are shown in parentheses

[b] $R_{merge} = 100\Sigma(h)\Sigma(i)|I(i)-\langle I\rangle|/\ \Sigma(h)\Sigma(i)I(i)$, where *I*(i) is the ith intensity measurement of reflection h, and $\langle I\rangle$ is the average intensity from multiple observations

[c] $R = \Sigma||Fobs|-|Fcalc||/\ \Sigma|Fobs|$. Where *Fobs* and *Fcalc* are the structure factor amplitudes from the data and the model, respectively. [d] Per asymmetric unit

The sequences most homologous to *P. finnis* CelD that have solved experimental structures were all GH5_4 enzymes from *P. rhizinflata* (PDB: 3AYR, 82% identity), *Ruminococcus champanellensis* (PDB: 6WQP, 43.1% identity), *Acetivibrio cellulolyticus* (PDB: 6MQ4, 39.4% identity), and *Clostridium cellulovorans* (PDB: 3NDY, 40% identity) [208,209] (Fig 5.2). These structures align to *P. finnis* CelD with C_α RMSDs of 0.45 Å, 1.37 Å, 1.44 Å, and 1.82 Å for 3AYR, 6WQP, 6MQ4, and 3NDZ respectively. All enzymes conserve the catalytic glutamic acid residues at positions 154 and 278 characteristic of GH5 enzymes [205] and many of the strictly conserved sites in the multiple sequence alignment (MSA) encode aromatic amino acids, suggesting their role in substrate binding. The primary structural feature differentiating the apo structures is the loop connecting the β_1-strand with the α_1-helix (Fig. 1a and Supplemental Fig. S1). CelD and 3AYR both have 13 residue loops pinned by a disulfide bond between Cys27 and Cys43. 6WQP maintains a long loop of 12 residues, while 6MQ4 and 3NDY contain shorter loops of 9 and 2 residues respectively. Other loops exhibiting structural diversity among the enzymes connect the β_6-strand with the α_6-helix and the β_8-strand with the β_9-strand (Fig. 5.3). The proximity of these loops to the substrate binding site (Fig. 5.1b) suggests their structure and amino acid composition plays an important role in substrate binding specificity.

As only nine GH5 structures from the fungal kingdom have been solved to date, we sought to compare *P. finnis* CelD to other fungal GH5 structures. Based on a sequence alignment of CelD with three fungal GH5 members studied at the structural level (*Piromyces rhizinflata* EglA (PDB 3AYR), a GH5_4, *Trichoderma reesei* EgII (PDB 3QR3), a GH5_5 , and *Thermoascus aurantiacus* EngI (PDB 1GZJ), a GH5_5), EglA from *P. rhizinflata* is unsurprisingly the closest structural homolog from the fungal kingdom to

CelD with sequence identity to CelD of 82% [208,210,211]. Optimal superposition of the CelD catalytic domain structure with corresponding EglA, EgII and EngI homologous domains results in 356, 242 and 254 equivalent Cα atoms with the RMSD values of 0.43 Å, 1.91 Å and 2.22 Å, respectively. Although these enzymes all share the same basic $(\beta/\alpha)_8$–barrel topology, they exhibit significant sequence diversity in the loop areas connecting the major

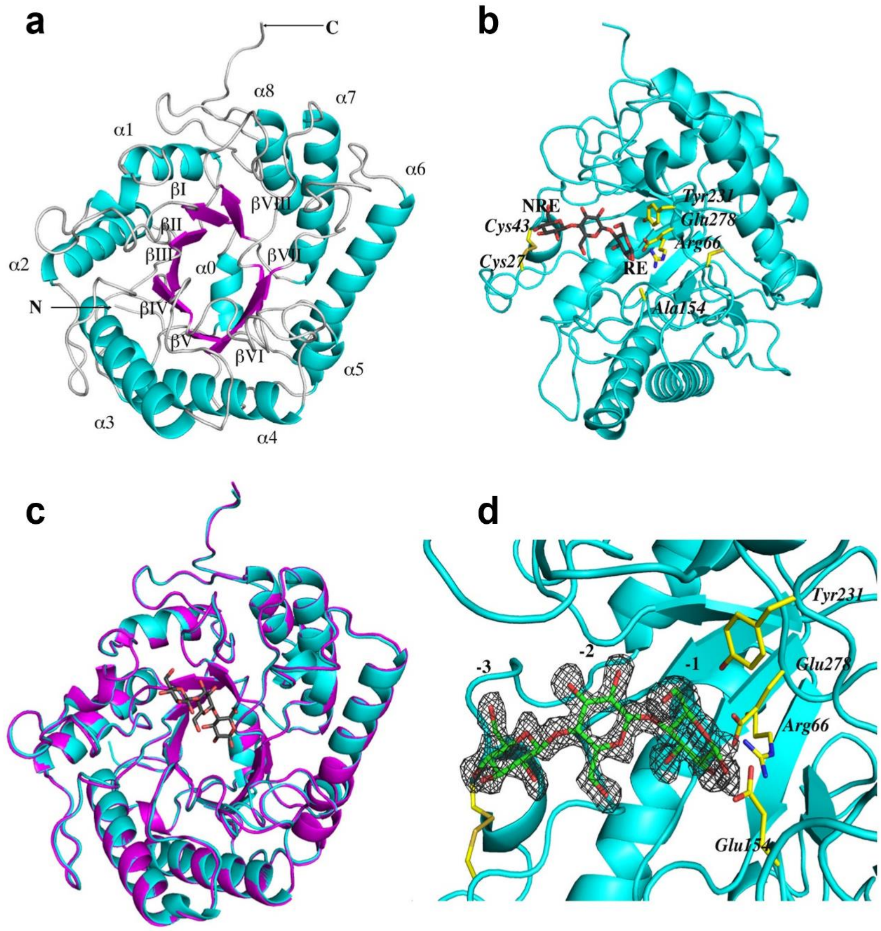


Fig. 5.1 The overall structure of the CelD catalytical domain. (A) Ribbon diagram of the apo-CelD structure with the secondary structural elements indicated. The α-helixes (α1-α8, cyan) flanking the β-strands (βI-βVIII, magenta) are labeled. Arrows indicate the N- and C-termini of the protein. (B) The structure of the E154A CelD variant in complex with the cellotriose (grey sticks). One disulfide bond and the catalytic residues are indicated and are shown as a yellow stick model. The reducing (RE) and non-reducing (NRE) ends of the oligosaccharide are indicated. (C) Superposition of the apo wild-type (magenta) and the E154A-ligand (cyan) crystal structures. The catalytic domains are superimposed by aligning the Cα atoms and are presented as ribbon diagrams. The cellotriose ligand is shown as a grey stick. (D) The substrate binding area in the ligand-bound complex. The protein moiety is presented as a cyan ribbon. The cellotriose molecule (green) bound to the -3, -2, and -1 glucose-binding subsites and the residues of the active site (yellow) are shown as sticks. The electron density map (grey mash) around the bound ligand is countered at the 1.4σ level.

structural elements. There are several extra residue insertions observed for CelD and EglA in the loops between β_I/α_1, β_{IV}/α_4, α_5/β_{VI}, and β_{VIII}/α_8 compared to very compact loop structures of EgII and EngI. As previously mentioned for the N-terminal loop, these loops could contribute differently to substrate binding and enzyme specificity as well as to thermal stability of the enzyme. It is also interesting to note that the *T. reesei* EgII cellulase contains 8 cysteine residues, which form four disulfide bonds and one of these bridges (Cys222 –

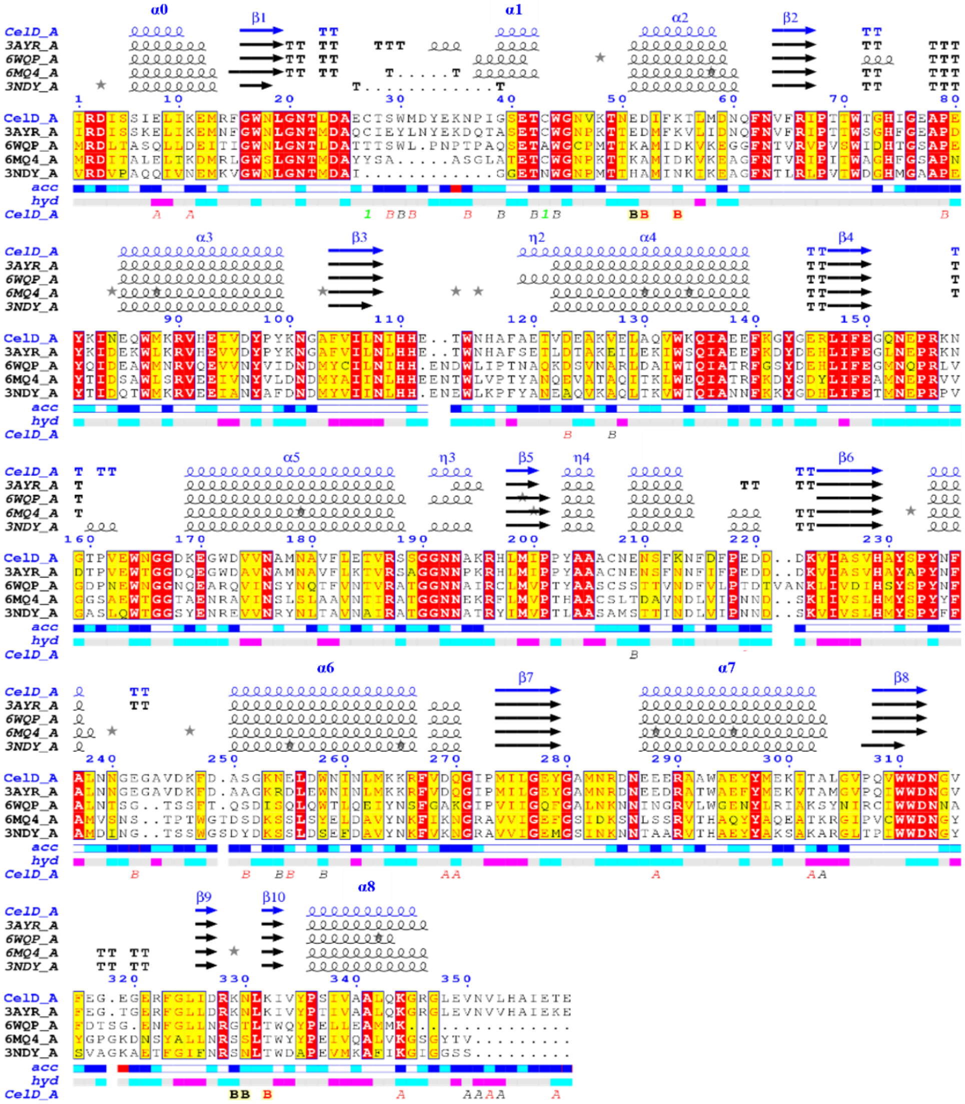

α0
β1
α1
α2
β2
CelD_A IRDISSIELLIKEMRFGWNLGNTLDAECTSWMDYEKNPIGSETCWGNVKTNEDIFKTLMDNQFNVFRIPTTWTGHIGEAPE
3AYR_A IRDISSKELLIKEMNFGWNLGNTMDAQCIEYLNYEKDQTASETCWGNPKTTEDMFKVLIDNQFNVFRIPTTWSGHFGEAPD
6WQP_A MRDLTASQLLDEITIGWNLGNTLDATTTSWL.PNPTPAQSETAWGCPMTTKAMIDKVKEGGFNTVRVPVSWIDHTGSAPE
6MQ4_A MRDITALELTKDMRLGWSLGNTMDAYYSA.....ASGLATETCWGNPKTTKAMIDKVKEAGFNTVRIPITWAGHFGSAPN
3NDY_A VRDVPAQQIVNEMKVGWNLGNTMDAI............GGETNWGNPMTTHAMINKIKEAGFNTLRLPVTWDGHMGAAPE
acc
hyd
α3
β3
η2
α4
β4
CelD_A YKINEQWMKRVHEIVDYPYKNGAFVILNIHHE..TWNHAFAETVDEAKVELAQVWKQIAEEFKGYGERLIFEGQNEPRKN
3AYR_A YKIDEKWLKRVHEVVDYPYKNGAFVILNLHHE..TWNHAFSETLDTAKEILEKIWSQIAEEFKDYDEHLIFEGLNEPRKN
6WQP_A YQIDEAWMNRVQEVVNYVIDNDMYCILNIHH.ENDWLIPTNAQKDSVNARLDAIWTQIATRFGSYDEHLIFEGMNQPRLV
6MQ4_A YTIDSAWLSRVEEIVNYVLDNDMYAIINLHHEENTWLVPTYANQEVATAQITKLWEQIATRFKDYSDYLIFEAMNEPRVV
3NDY_A YTIDQTWMKRVEEIANYAFDNDMYVIINLHH.ENEWLKPFYANEAQVKAQLTKVWTQIANNFKKYGDHLIFETMNEPRPV
α5
η3
β5
η4
β6
CelD_A GTPVEWNGGDKEGWDVVNAMNAVFLETVRSSGGNNAKRHLMIPPYAAACNENSFKNFDFPEDD..DKVIASVHAYSPYNF
3AYR_A DTPVEWTGGDQEGWDAVNAMNAVFLKTVRSAGGNNPKRHLMIPPYAAACNENSFNNFIFPEDD..DKVIASVHAYAPYNF
6WQP_A GDPNEWNGGNQEARQVINSYNQTFVNTVRATGGNNAIRCLMVPTYAASCSSTTVNDFVLPTDTVANKLIVDIHSYSPYNF
6MQ4_A GGSAEWTGGTAENRAVINSLSLAAVNTIRATGGNNEKRFLMVPTHAACSLTDAVNDLVIPNND..SKIIVSLHMYSPYYF
3NDY_A GASLQWTGGSYENREVVNRYNLTAVNAIRATGGNNATRYIMVPTLAASAMSTTINDLVIPNND..SKVIVSLHMYSPYFF
α6
β7
α7
β8
CelD_A ALNNGEGAVDKFD.ASGKNELDWNINLMKKRFVDQGIPMILGEYGAMNRDNEEERAAWAEYYMEKITALGVPQVWWDNGV
3AYR_A ALNNGEGAVDKFD.AAGKRDLEWNINLMKKRFVDQGIPMILGEYGAMNRDNEEDRATWAEFYMEKVTAMGVPQIWWDNGV
6WQP_A ALNTSG..TSSFT.QSDISQLQWTLQEIYNSFGAKGIPVIIGQFGALNKNNINGRVLWGENYLRIAKSYNIRCIWWDNNA
6MQ4_A AMVSNS..TPTWGTDSDKSSLSYELDAVYNKFIKNGRAVVIGEFGSIDKSNLSSRVTHAQYYAQEATKRGIPVCWWDNGY
3NDY_A AMDING..TSSWGSDYDKSSLDSEFDAVYNKFVKNGRAVVIGEMGSINKNNTAARVTHAEYYAKSAKARGLTPIWWDNGY
β9
β10
α8
CelD_A FEG.EGERFGLIDRKNLKIVYPSIVAALQKGRGLEVNVLHAIETE
3AYR_A FEG.TGERFGLLDRKNLKIVYPTIVAALQKGRGLEVNVVHAIEKE
6WQP_A FDTSG.ENFGLLNRGTLTWQYPELLEAMMK...............
6MQ4_A YGPGKDNSYALLNRSSLTWYYPEIVQALVKGSGYTV.........
3NDY_A SVAGKAETFGIFNRSNLTWDAPEVMKAFIKGIGGSS.........

Fig. 5.2. Sequence alignment of homologous GH5_4 enzymes with known structures. Strictly conserved residues are shown in red block and chemically similar residues shown in yellow block. “acc” represents the solvent accessibility and “hyd” the hydropathy of the sequence. 6WQP contains two mutations from the wild-type enzyme sequence (Uniprot Accession no. D4LAX7), E154Q and E278Q in the multiple sequence alignment. Figure generated with Endscript (Robert and Gouet 2014).

Cys249), pinning the α_6/β_{VI} loop to the N-terminus of the α_7–helix, corresponds to that observed in *Thermoascus aurantiacus* EngI (Fig. 5.4). *T. aurantiacus* EngI is a hyperthermophilic enzyme with a T_m of about 81 °C, meanwhile the reported T_m value for *T. reesei* EgII is 69.5 °C [212]. CelD and EglA also exhibit one disulfide bridge located in the N-terminal loop, but there is no thermal stability data yet available for these enzymes.

Structure of the CelD – cellotriose complex

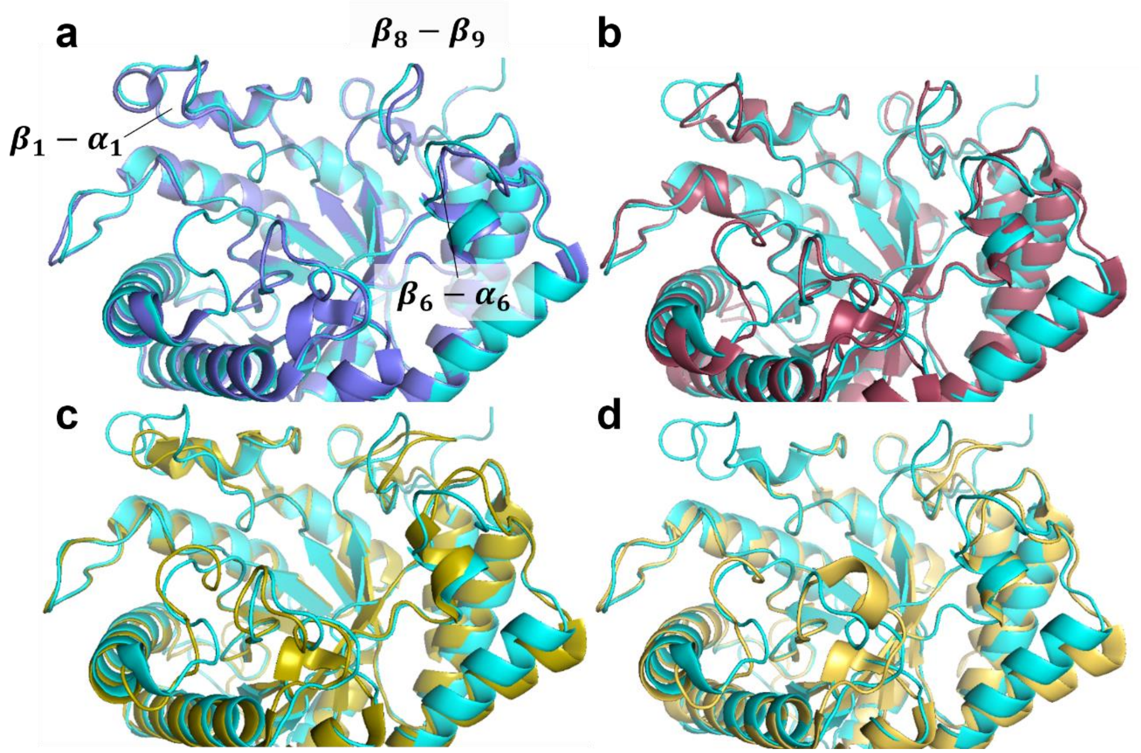


Fig. 5.3. Structural alignment of CelD (cyan) to 3AYR (a), 6WQP (b), 6MQ4 (c), and 3NDZ (d). The loops exhibiting the greatest structural diversity are annotated and CelD is shown in the same orientation in each panel.

Attempts to crystallize the CelD active catalytic domain with cellobiose or cellotriose substrates were unsuccessful using both soaking and co-crystallization approaches. We additionally attempted to crystallize CelD with its natural dockerin domains, which would have represented the first full structure of a fungal dockerin-fused enzyme but failed to get high quality crystals. After introducing the inactivating E154A mutation to the active site of a single CelD catalytic domain, we were able to obtain high quality crystals of the E154A-cellotriose complex. As described above, we did not observe substantial conformational changes between the apo-enzyme structure and the mutant structure with the bound ligand. Thus, the catalytic residues in the apo- CelD structure are likely to be in a catalytically competent position and superposition of the active enzyme structure and the mutant structure with the bound ligand provides sound structural information about the substrate binding and mechanism of the enzyme catalysis. Like other GH5 enzymes, CelD contains two invariant catalytic glutamate residues, the acid/base Glu154 and the nucleophile Glu278. Superposition of two structures solved in this work confirms that the carboxylate OE1 and OE2 oxygens of the catalytic acid/base glutamate point toward the O1 atom of the -1 glucopyranose unit at 1.5 Å and 1.7 Å, respectively, and the carboxylate oxygen of the complementary nucleophile glutamate forms a hydrogen bond with the anomeric C1 carbon of the same saccharide unit at 3.1 Å (Fig. 5.5a). Meanwhile, the nucleophile Glu278 is sandwiched between two conserved Arg66 and Tyr231 residues which form hydrogen bonds to the carboxylate oxygens and appear to serve a supportive role to stabilize this Glu residue throughout catalysis as observed for other GH5 enzymes (Fig. 5.5c) [208,213].

We found 8 ordered water molecules making multiple hydrogen bonds with both the protein moiety and the cellotriose molecule in the ligand-bound structure; two of them are located near the anomeric carbon of the -1 saccharide moiety (Fig. 5.5a). One of these water

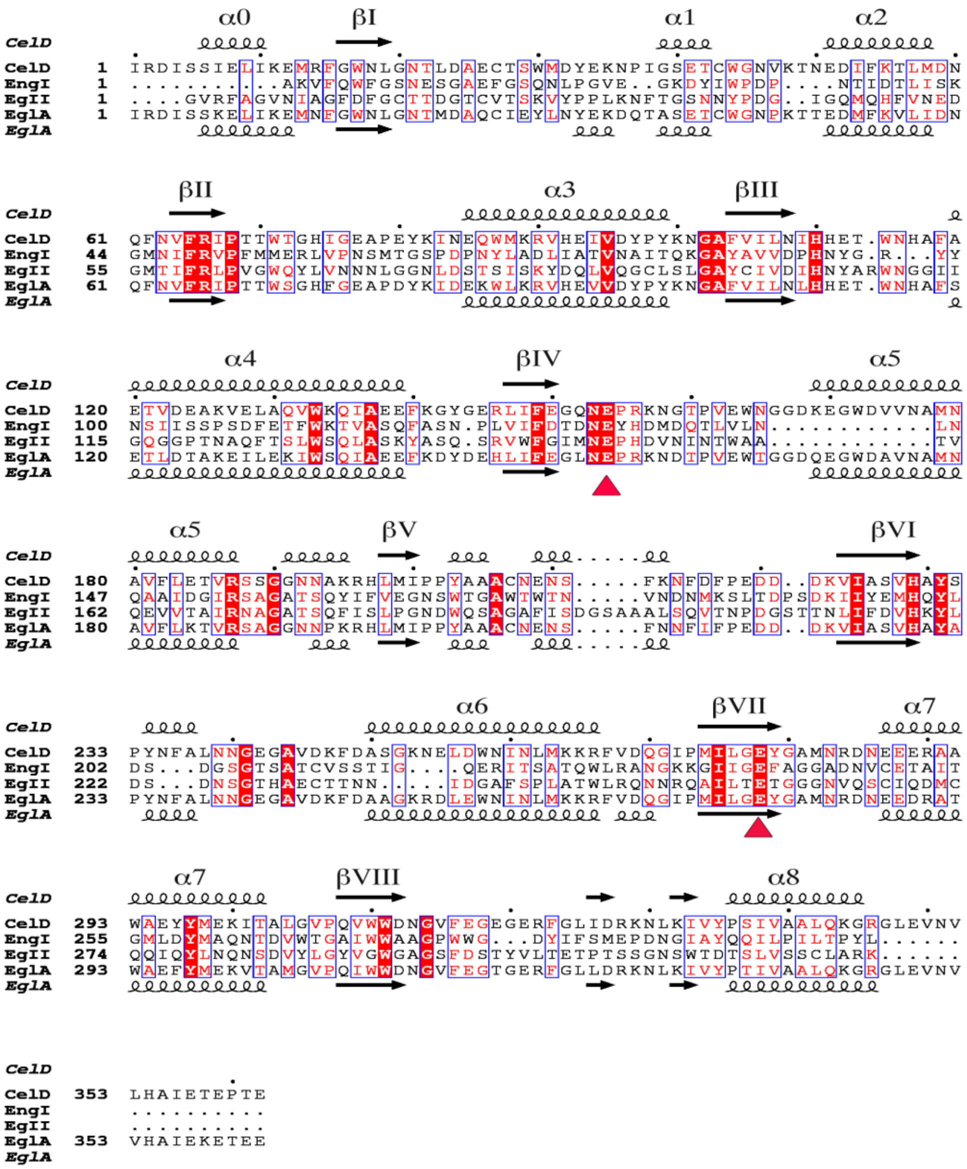

α0
βI
α1
α2
CelD
CelD 1 IRDISSIELIKEMRFGWNLGNTLDAECTSWMDYEKNPIGSETCWGNVKTNEDIFKTLMDN
EngI 1 ...........AKVFQWFGSNESGAEFGSQNLPGVE..GKDYIWPDP....NTIDTLISK
EgII 1 ....GVRFAGVNIAGFDFGCTTDGTCVTSKVYPPLKNFTGSNNYPDG..IGQMQHFVNED
EglA 1 IRDISSKELIKEMNFGWNLGNTMDAQCIEYLNYEKDQTASETCWGNPKTTEDMFKVLIDN
EglA
βII
α3
βIII
CelD
CelD 61 QFNVFRIPTTWTGHIGEAPEYKINEQWMKRVHEIVDYPYKNGAFVILNIHHET.WNHAFA
EngI 44 GMNIFRVPFMMERLVPNSMTGSPDPNYLADLIATVNAITQKGAYAVVDPHNYG.R...YY
EgII 55 GMTIFRLPVGWQYLVNNNLGGNLDSTSISKYDQLVQGCLSLGAYCIVDIHNYARWNGGII
EglA 61 QFNVFRIPTTWSGHFGEAPDYKIDEKWLKRVHEVVDYPYKNGAFVILNLHHET.WNHAFS
EglA
α4
βIV
α5
CelD
CelD 120 ETVDEAKVELAQVWKQIAEEFKGYGERLIFEGQNEPRKNGTPVEWNGGDKEGWDVVNAMN
EngI 100 NSIISSPSDFETFWKTVASQFASN.PLVIFDTDNEYHDMDQTLVLN............LN
EgII 115 GQGGPTNAQFTSLWSQLASKYASQ.SRVWFGIMNEPHDVNINTWAA............TV
EglA 120 ETLDTAKEILEKIWSQIAEEFKDYDEHLIFEGLNEPRKNDTPVEWTGGDQEGWDAVNAMN
EglA
α5
βV
βVI
CelD
CelD 180 AVFLETVRSSGGNNAKRHLMIPPYAAACNENS.....FKNFDFPEDD..DKVIASVHAYS
EngI 147 QAAIDGIRSAGATSQYIFVEGNSWTGAWTWTN.....VNDNMKSLTDPSDKIIYEMHQYL
EgII 162 QEVVTAIRNAGATSQFISLPGNDWQSAGAFISDGSAAALSQVTNPDGSTTNLIFDVHKYL
EglA 180 AVFLKTVRSAGGNNPKRHLMIPPYAAACNENS.....FNNFIFPEDD..DKVIASVHAYA
EglA
α6
βVII
α7
CelD
CelD 233 PYNFALNNGEGAVDKFDASGKNELDWNINLMKKRFVDQGIPMILGEYGAMNRDNEEERAA
EngI 202 DS...DGSGTSATCVSSTIG....QERITSATQWLRANGKKGIIGEFAGGADNVCETAIT
EgII 222 DS...DNSGTHAECTTNN.....IDGAFSPLATWLRQNNRQAILTETGGGNVQSCIQDMC
EglA 233 PYNFALNNGEGAVDKFDAAGKRDLEWNINLMKKRFVDQGIPMILGEYGAMNRDNEEDRAT
EglA
α7
βVIII
α8
CelD
CelD 293 WAEYYMEKITALGVPQVWWDNGVFEGEGERFGLIDRKNLKIVYPSIVAALQKGRGLEVNV
EngI 255 GMLDYMAQNTDVWTGAIWWAAGPWWG...DYIFSMEPDNGIAYQQILPILTPYL......
EgII 274 QQIQYLNQNSDVYLGYVGWGAGSFDSTYVLTETPTSSGNSWTDTSLVSSCLARK......
EglA 293 WAEFYMEKVTAMGVPQIWWDNGVFEGTGERFGLLDRKNLKIVYPTIVAALQKGRGLEVNV
EglA
CelD
CelD 353 LHAIETEPTE
EngI ....
EgII ....
EglA 353 VHAIEKETEE
EglA

Fig. 5.4. Sequence alignment of fungal cellulases from *Piromyces finnis* (CelD), *Thermoascus aurantiacus* (EngI), *Trichoderma reesei* (EgII), and *Piromyces rhizinflata* (EglA). Strictly conserved residues are shown in red block, and chemically similar residues in red text. The residue numbering is shown on the left for each catalytic domain. Dashed lines indicate deletions. The acid/base and nucleophile residues are indicated below by red triangles. The secondary structure elements of CelD and EglA are shown above and below the alignment, respectively. The sequence alignment revealed close homology between CelD and EglA cellulases (82% amino acid residue identity). The figure was generated with ESPript (http://espript.ibcp.fr).

molecules may participate in the catalytic mechanism of the CelD as a nucleophilic attack on the glycosyl-enzyme intermediate [214].

Holo CelD E154A-cellotriose structure reveals key substrate recognition sites

Structural comparison of the CelD cellulase with other known GH5_4 family enzymes suggests that a cleft containing the enzyme active site may provide a favorable platform for binding oligomers up to seven sugar units within the -3 to +4 subsites. This cleft presents a flat platform for interacting with negative sugar subsites and a U-shaped groove that appears

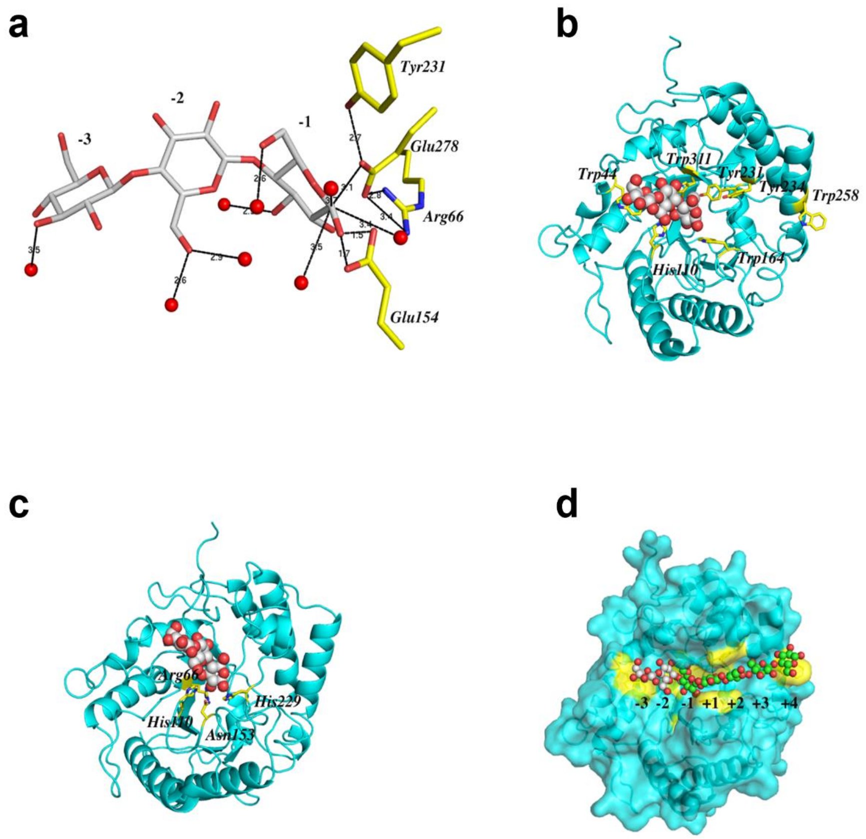


Fig. 5.5 Substrate binding area of the CelD cellulase. (A) Zoomed in view of the active site with bound cellotriose molecule (grey stick) surrounded by 8 water molecules (red spheres) and several catalytic residues (yellow stick). The hydrogen bonds are shown as black dotted lines and location of the substrate-binding sites (the -3. -2, and -1) are labeled. (B) Ribbon representation of the CelD-ligand complex (cyan). The positions of the conserved aromatic residues involved in substrate binding are shown in yellow stick, the cellotriose ligand is shown as grey spheres. (C) The same as in (B) except the positions of the conserved and spatially conserved residues associated with cellulolytic activity are shown in yellow stick. (D) Total view of the protein surface (in the same color as in B and C) showing the wide CelD active side cleft with the modeled hepta-oligosaccharide substrate is presented. The hepta-olygosaccharide is shown as spheres. The modeled saccharide units from the -1 subsite to the +4 subsite shown as green spheres and experimentally observed carbohydrate units at the -3 and -2 subsites shown as grey spheres. The scissile glycoside bond is between the -1 and +1 sites. The positions of conserved aromatic residues served to mediate carbohydrate binding in the encounter complex are indicated in yellow color.

to orient the substrate for catalysis and provide interaction sites for positive sugar subsites. The CelD active site groove is lined with aromatic residues, Trp44, Trp164, Tyr231, Tyr234, Trp258, and Trp311, and all these positions except Trp258 are strictly conserved among the GH5_4 enzymes we analyzed (Fig. 5.2). Trp164 and Tyr234 are proximal to the +2 and +3 subsites and Trp44 and Tyr231 are close to the -3 and -1 binding sites, respectively (Fig. 5.5b,d). Trp258 is positioned to interact with a linear polysaccharide chain at the +4 position.

The overall shape of the CelD polysaccharide binding site appears optimal for strictly linear polysaccharides, but an indent in the enzyme surface into which the C6 atom of the -2 backbone glucose points (Fig. 5.5c) suggests this enzyme may accommodate a substrate like xyloglucan, which contains a glucose backbone with branched xylose and galactose sugars. Structural and biochemical analysis of several GH5_4 enzymes suggests some structural and

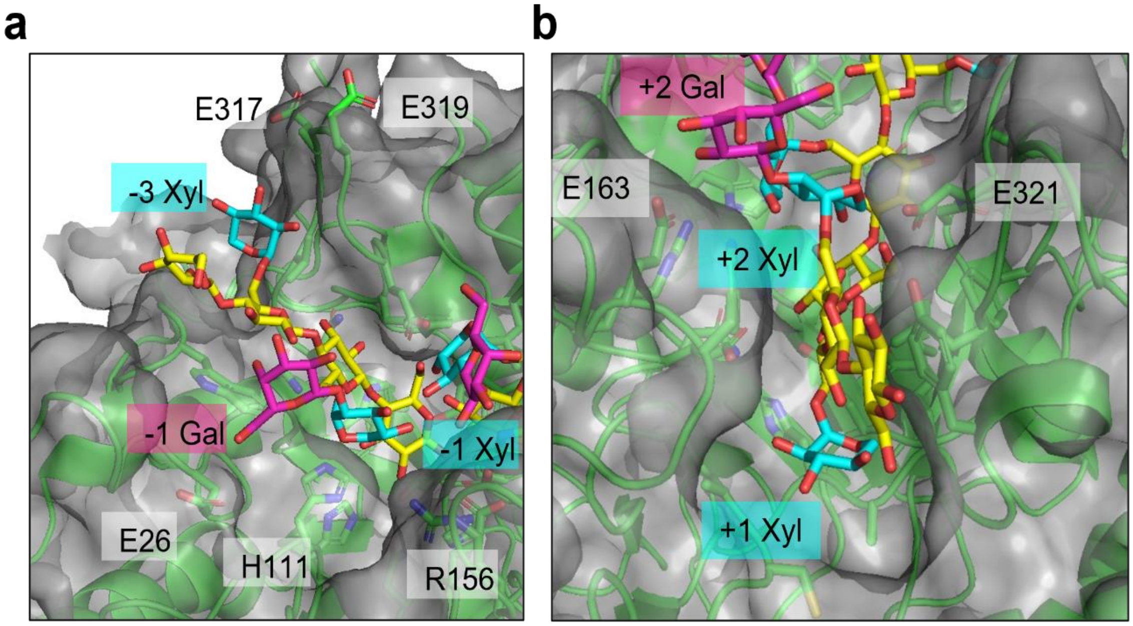


Fig. 5.6 Potential contacts between xyloglucan branched sugars and CelD modeled by alignment with 4W88 and 2JEQ with annotated CelD residues interacting with negative substrate subsites (a) and positive substrate subsites (b). CelD is shown both as a gray surface and green cartoon representation to visualize both atomic contacts and the fit of xyloglucan to the active site.

sequence signatures indicative of enzyme activity on branched polysaccharide substrates include aromatic and polar side chains within the loops between β_3 and α_4, β_4 and α_5, and β_8 and β_9 [209]. Aligning the CelD structure to the xyloglucan oligosaccharide-bound structures 2JEQ [215] and 4W88 [216] suggests CelD can spatially accommodate xyloglucan, with potential favorable hydrogen bonding interactions with the -3 xylose branch at E317 or E319, the -2 xylose and galactose involving H111, R156, and E26, the +2 galactose at E163, and the +2 xylose at E321 (Fig. 5.6). However, CelD does not appear to have any aromatic side chains poised to interact with branched sugars that would suggest this enzyme is specific for branched polysaccharides.

Characterization CelD substrate specificity and enzyme kinetics

We tested the *P. finnis* CelD catalytic domain against several soluble and insoluble substrates to determine the catalytic domain's specificity for cellulose vs hemicellulose polysaccharides, its endoglucanase vs β-glucosidase activity, and its preference for β-1,4 vs other α linkages. As indicated by structural analysis, CelD hydrolyzes the linear cellulose analog carboxymethylcellulose (CMC) and the branched polysaccharides β-D-glucan (mixed linkage glucan or MLG) and xyloglucan, but displays very poor activity against insoluble, phosphoric acid swollen cellulose (PASC) (Table 5.2), supporting CelD's characterization as a broad-spectrum endoglucanase.

Table 5.2. Specific activity of the CelD catalytic domain against several substrates. Activities are reported as the mean ± standard deviation in units of $\boldsymbol{\mu mol\ glucose\ equivalent\ min.^{-1}}$ (U) per μmol enzyme. CMC: carboxymethylcellulose, PASC: phosphoric acid swollen cellulose.

	CMC	**β-D-glucan**	**Xyloglucan**	**PASC**
Specific activity $(\frac{U}{\mu mol\ enzyme})$	117.8 ± 5.8	249.4 ± 10.8	191.4 ± 12.8	0.11 ± 0.14

The enzyme showed no activity against xylan or arabinogalactan nor β-glucosidase activity (Data not shown). Relative positions of key pyranose binding residue W44, which interacts with the substrate at the -3 position, and the catalytic E154 position corroborates the lack of β-glucosidase activity (Fig. 5.5a-b).

Table 5.3. Kinetic parameters of P. finnis CelD for CMC hydrolysis in comparison to other GH5 family members

Enzyme	**Kcat (s^{-1})**	**Km (g/L CMC)**	**Kcat/Km (L/g/s)**	**Source**
Piromyces finnis CelD catalytic domain	6.0 ± 0.58	7.6 ± 2.1	0.8 ± 0.2	This work
Trichoderma reesei EglII	1331 ± 37	0.84 ± 0.14	1584	[217]
Thermotoga maritima Cel12A	791	2.1	402	[218]
Penicillium verruculosum EG2	152	22.8 ± 0.2	6.6 ± 0.1	[219]
Thermoanaerobacter tengcongensis MB4 Cel5A	1.94 ± 0.01	1.4 ± 0.1	1.4 ± 0.1	[220]
Aspergillus fumigatus Egl2010	6	2.0 ± 0.6	3.0	[221]
Aspergillus nidulans Egl2010	4	29 ± 8.8	0.1	[221]
Fusarium graminearum Egl2010	14	13 ± 4.8	1.1	[221]
Aureoblasidium pullulans SEQ15654	29	10 ± 2.9	2.9	[221]
T. reesei Eg2/Cel5a	4	2.6 ± 0.7	1.5	[221]
Gloeophyllum trabeum SEQ630	6	13 ± 3.9	0.5	[221]
Sporotrichum thermophile SEQ13822	6	3.3 ± 0.6	1.8	[221]
Clostridium thermocellulum EngD	30.1	6.5	4.6	[221]
Martelella mediterranea Cel5D	3.5	8.8 ± 0.1	0.4	[222]

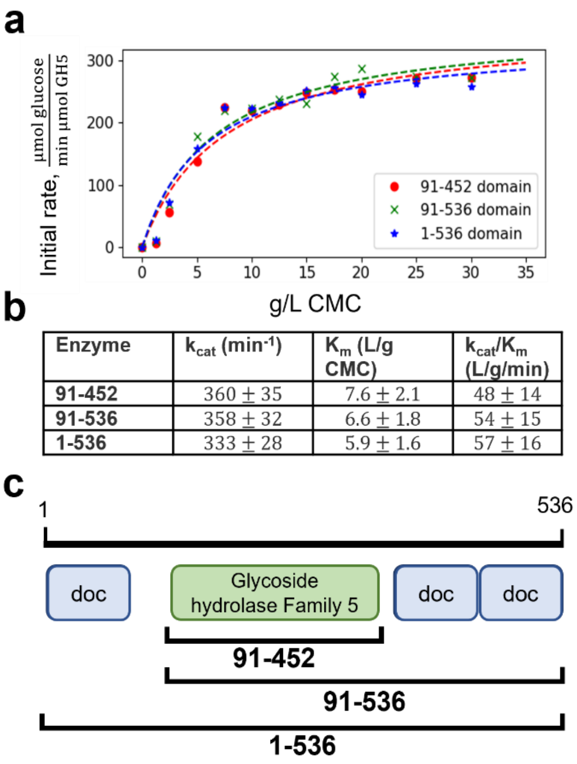
We measured kinetic rate parameters for the CelD catalytic domain acting on the soluble cellulase substrate carboxymethylcellulose (CMC) at 39 °C and pH 5.5, the physiological temperature for anaerobic fungi and acidic pH typical of endoglucanase enzymes. Our measured k_{cat} and K_m for *P. finnis* CelD are comparable to those of other fungal endoglucanases but well below those of thermophilic bacterial cellulose degraders like *Thermotoga maritima* and *Clostridium thermocellum* (Table 5.3).

CelD cellulase kinetics are unperturbed by the addition of N- and C-terminal dockerins

A key question we sought to answer was whether CelD's natural C-terminal dockerin domains conferred any catalytic benefit and whether this enzyme could tolerate non-natural dockerin domains fusions, such as one on its N-terminus. True modularity in the construction of catalytic domain – dockerin domain chimeras is highly desirable in building synthetic enzyme systems for applications like lignocellulose valorization. Our results show that CelD's intrinsic kinetics are unchanged when dockerin domains are fused to the N- or C-terminus of this protein, indicating the natural CelD protein is highly modular and suggesting that the CelD catalytic domain can accommodate other fusion partners (Fig. 5.7).

The effect of dockerin domains on the activity of gut fungal enzymes has previously been evaluated in only a few cases with conflicting results. While no change in activity was observed upon removal of the native C-terminal dockerin from a *Piromyces* mannanase at 39°C [58], Huang and co-authors found removal of the native C-terminal double dockerin from *Neocallimastix frontalis* Xyn11A and Xyn11B to increase specific xylanase activity at all temperatures (39 - 70°C) [223]. We have also found the addition of the C-terminal double dockerin from *Piromyces finnis* CelD to *Thermotoga maritima* enzymes Cel5A and XynA to cause insignificant changes to specific enzyme activity at 80°C [7].

This lack of consistency suggests the effect of fungal dockerin fusions on catalytic domain activity is context dependent, at least when evaluating enzymes recombinantly



Enzyme	k_{cat} (min^{-1})	K_m (L/g CMC)	k_{cat}/K_m (L/g/min)
91-452	360 ± 35	7.6 ± 2.1	48 ± 14
91-536	358 ± 32	6.6 ± 1.8	54 ± 15
1-536	333 ± 28	5.9 ± 1.6	57 ± 16

Fig. 5.7 Addition of N- and C-terminal fungal dockerin domains does not affect enzyme kinetics. Kinetic parameters for hydrolysis of carboxymethylcellulose (CMC) were fit from initial rate data for the GH5 catalytic domain alone (91-452) as well as the GH5 with two C-terminal dockerins (91-536) and the GH5 with N- and C-terminal dockerins (1-536). Dockerins are abbreviated "doc" for short. Initial rates for each enzyme at each substrate concentration were taken from time course measurements of released reducing sugar vs time, as quantified by the DNS assay. Fit parameter uncertainties are reported ± one standard deviation from non-linear least squares fit.

produced in *E. coli*. Unfortunately, our attempts to crystallize a construct containing both the catalytic and dockerin domains were unsuccessful, and a structure of a complete, dockerin-containing enzyme from an anaerobic fungus with which to more definitively address these questions remains unsolved. Negative effects of dockerin domains on enzymatic activity have previously been tied to a reduction in protein thermostability and melting temperature [223]. However, it is difficult to decouple potential intrinsic instability of the dockerin domain from the possibility that these domains, which are known to possess several disulfide bonds [59,224], are misfolded when produced recombinantly in *E. coli*. More efficient disulfide bond formation was shown to have a dramatic impact on the measured enzymatic activity of a non-dockerin containing *Neocallimastix patriciarum* xylanase, which the authors evaluated by producing the same enzyme in *E. coli* and *Pichia pastoris* [225]. Ongoing work investigating dockerin-containing enzymes from their native system will build on these previous results to address this outstanding question of how the dockerin domain contributes to enzyme activity and stability.

Anaerobic fungi deploy an array of CAZymes that act in solution and as members of multi-enzyme cellulosomes to rapidly hydrolyze lignocellulose. However, very few enzymes in the vast CAZyme repertoire encoded by anaerobic fungal genomes have been functionally characterized, and as a result, we have little biochemical understanding of how anaerobic fungi excel at degrading biomass, which presents a challenge towards converting anaerobic fungal enzyme systems into useful biotechnologies. By characterizing the atomic resolution structure and kinetic properties of the *P. finnis* CelD GH5 endoglucanase, we provide additional insight towards gaining biochemical understanding of anaerobic fungal enzyme systems. The kinetic data indicate the domains of CelD are highly modular and can likely be augmented to functionalize this GH5 enzyme with other domains, while the structure presents a platform for rational engineering of this enzyme for higher thermostability or activity criteria.

5.2.2 Heterologous expression and functional characterization of spore coat CotH proteins from Piromyces finnis

Many dockerin proteins contain spore coat CotH domains

The functional diversity of fungal cellulosomes has been inferred by analyzing the dockerin-containing genes annotated in gut fungal genomes, finding that, of the 5783 dockerin-containing genes, *1058 (18%) have CotH domains* (mycocosm.jgi.doe.gov). This proportion is second only to CAZymes among dockerin-containing genes, which comprise 27%. The fact that CotH domains comprise such a significant fraction of dockerin-containing genes suggests they are important to some aspect of cellulosome biology. Our initial hypothesis was that, as protein kinases, perhaps CotH domains phosphorylate dockerin domains to post-translationally mediate cellulosome assembly or disassembly. Other hypotheses suggest CotH participate in cellulosome protein trafficking and secretion

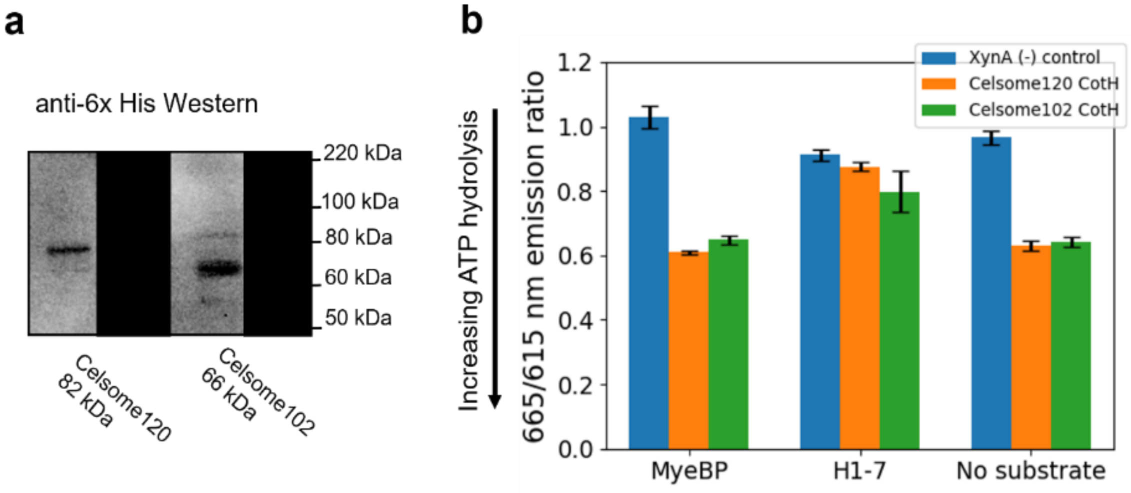


Figure 5.7. P. finnis CotH proteins possess ATP hydrolysis activity. (a) Western blot showing expression of the two P. finnis CotH proteins, Celsome102 and Celsome120 at their expected size. (b) Production of ADP during in vitro kinase activity assays as measured by the Adapta kinase activity assay (Fisher Scientific). Mg^{2+} is the divalent cation supplied in these experiments. Both Celsome102 and Celsome120 autophosphorylate in the absence of substrate and potentially phosphorylate MyeBP peptide but appear not to phosphorylate the PKA substrate H1-7. XynA, a xylanase with no kinase activity, was used as a negative control. Error bars represent the SEM of technical triplicates.

or localization to the cell wall, more in line with their supported function in *Bacillus* bacteria [226].

Structural bioinformatics suggests anaerobic fungal CotH domains are kinases

Towards elucidating the role of CotH proteins in fungal cellulosome biology, we first considered whether CotH genes from anaerobic fungal genomes could act as protein kinases. Using *Piromyces finnis* as a representative anaerobic fungus, I performed a multiple sequence alignment of all annotated dockerin-containing CotH protein sequences (129 sequences total) to identify potential conserved sites that are key to ATP binding, a necessary event in protein kinase function. The structure of *B. subtilis* CotH, proven to be a protein kinase [226], served as a reference, highlighting a PWDXD motif that binds ATP and Mg^{2+} ions to hydrolyze ATP (Figure 5.6). Indeed, this motif was highly conserved among the *P. finnis* CotH domains, suggesting they may also act as protein kinases.

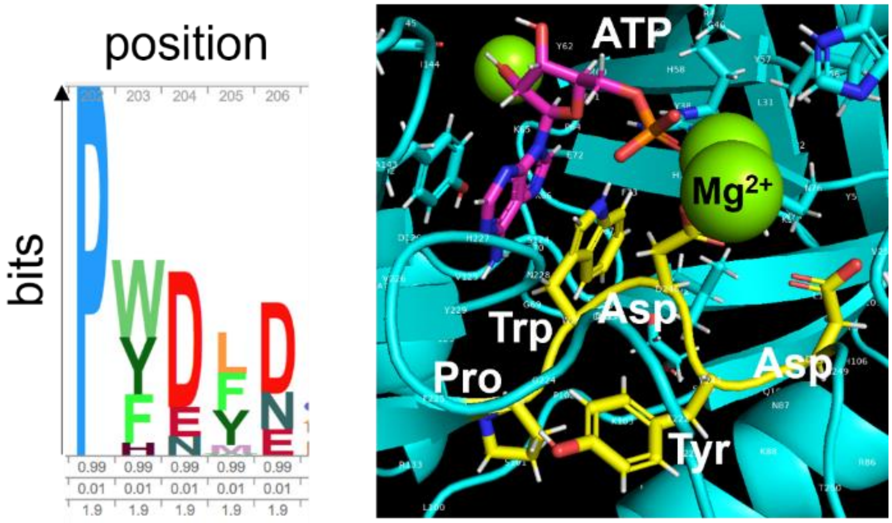


Figure 5.6. P. finnis CotH genes conserve key sites critical to protein kinase activity. (Left) Logo plot from multiple sequence alignment of 129 CotH domain sequences from P. finnis showing the conserved PWDXD motif characteristic of kinases. (Right) The PWDXD motif in B. subtilis CotH (PDB ID 5JDA) is directly involved in binding ATP and Mg^{2+} ions to mediate ATP hydrolysis.

To show that *P. finnis* CotH domains could also hydrolyze ATP, I produced and purified two CotH dockerin-containing proteins in *E. coli* and tested their ability to hydrolyze ATP using a kinase activity assay kit that detects the generation of ADP. The two CotH proteins expressed reasonably well as C-terminal 6x-His tagged proteins in *E. coli* (Figure 5.7a). Though sometimes protein kinases autophosphorylate, a substrate may be required to catalyze ATP hydrolysis. With no *a priori* knowledge of *P. finnis* kinase substrate specificity, we used myelin basic protein (MyeBP), a generic protein kinase substrate, and the protein kinase A substrate histone H1 phosphorylation site (H1-7), as the bacterial CotH structurally conserves key active site residues from canonical protein kinase A [226].

The *P. finnis* CotH proteins hydrolyze ATP in the absence of substrate or in the presence of MyeBP, but not in the presence of H1-7 (Fig. 5.7b). The change in ATP hydrolysis when changing the substrate suggests that substrate phosphorylation and not just ATP hydrolysis are being performed. This should be confirmed by testing catalytically inactive variants of Celsome102 and Celsome120 and/or using conventional γ-^{32}P ATP to directly detect phosphate addition to kinase substrates.

5.3 Conclusions

While the genomes of under-studied organisms like anaerobic fungi may hold a treasure trove of useful protein sequences that exceed current biotechnologies, these uncharacterized gene products must be biochemically verified to be useful for industrial applications. The GH5 and CotH proteins studied here represent two case studies from which we gain biochemical insight into the function of anaerobic fungi's primary lignocellulolytic machine, the cellulosome. In particular, the verification that anaerobic fungal CotH proteins are likely protein kinases raises interesting questions about the role they play in cellulosome biology. Future work leveraging cell-free expression systems and *in vitro* biochemical assays will help elucidate the substrates that CotH kinases target. As they become available, bioimaging techniques that label CotH proteins may also prove useful in illuminating the role that CotH proteins play in anaerobic fungal biology. More broadly, continued enzymatic characterization individual cellulosomal CAZymes like the GH5 studied here may enable the rational construction of anaerobic fungal enzyme systems with predictable biochemistry, which would be of great interest to the waste biomass valorization community.

5.4 Materials and Methods

Enzyme Cloning, Expression, and Purification

The cDNA fragment corresponding to the cellulase catalytic domain CelD (91-452 residues, GenBank accession number: ORX48147.1) was cloned from *Piromyces finnis* isolated from horse feces [42,82]. The PCR amplified construct was cloned into pMCSG68 expression vector with an N-terminal His-tag of MHHHHHHSSGVDLWSHPQFEKGTENLYFQSNA [227]. The E154A point mutation was introduced by using the QuickChange kit according to the manufacturer's protocol using the following primers: 5'-GGTCAAAACGCACCAAGAAAGAACGGTACTCCAGTTGA-3' as a forward primer and 5'-CTTTCTTGGTGCGTTTTGACCTTCGAAGATTAAACGTTCA-3' as a reverse primer. Both wild-type and mutated amino acid sequences were verified by nucleotide sequencing of the cloned constructs.

Genetic constructs containing CotH-dockerin genes were synthesized and subcloned into pET-28a by the Joint Genome Institute. We received strain cryostocks with the sequence-verified plasmids in *E. coli* DH5a and BL21(DE3).

Recombinant catalytic domains of wild-type CelD and the E154A mutant were expressed as a soluble protein in *E.coli* BL21-gold (DE3) cells by induction with 0.5 mM isopropyl β-D-1-thiogalactopyranoside at 18 ^{0}C for 12 hours. The protein sample was isolated from culture media using a nickel-nitrolotriacetic acid (Ni-NTA) column (Thermo Fisher Scientific). Protein for crystallization first underwent proteolytic cleavage by a TEV protease at 4 °C for 12 h. followed by a two-step purification procedure including a filtration through the Ni-NTA column (5 ml) to remove uncleaved product and His-tagged TEV-protease and a size exclusion chromatography on a Superdex 200 column (GE Healthcare) in 15 mM Tris-HCl buffer, pH 7.5, supplemented with 150 mM NaCl.

For recombinant CotH-dockerin protein production, *E. coli* BL21(DE3) cells containing the pET-28a Celsome 102 or pET-28a Celsome 120 plasmid were grown at 37°C with shaking from a starting OD_{600} of 0.05 to an OD_{600} of 0.5. Expression was induced by addition of IPTG to a final concentration of 0.4 mM, after which cultures were transferred to a 30°C incubator with shaking and incubated overnight. Cells were resuspended in PBS + 10mM imidazole buffer and lysed by bead beating, 10 cycles of 30 seconds. Cell debris was pelleted by centrifugation at 14,000x g for 10 minutes and the 6x-His tagged CotH proteins were purified from the clarified lysate by batch IMAC purification following the manufacturer's instructions. The presence of the CotH protein throughout the purification process was observed by SDS-PAGE and Western blot with anti-6x His-tagged antibody.

Complex Formation, Crystallization and Structure Determination

The best crystals of apo wild-type of the catalytic domain were obtained at 16 ^{o}C from sitting drops containing 0.4 μL of the protein sample at concentration of 21 mg/ml and 0.4 μL of reservoir solution consisting of 0.1 M Tris-HCl buffer, pH 8.7, 0.5 M $LiCl_2$ and 28% PEG 6,000.

The complex between the E154A mutant protein and cellotriose was formed by adding the ligand stock solution (Sigma-Aldrich) to the protein solution at 10 mM final concentration. The complex crystals were produced using a similar crystallization approach to that described above except the reservoir solution contained 0.1 M sodium acetate, 1.2 M $LiCl_2$ and 24% PEG 6,000 and the protein concentration was 16 mg/ml. Several cycles of microseeding under similar crystallization conditions were carried out to obtain crystals suitable for x-ray analysis for both structures.

For data collection, crystals were harvested with 20 % (v/v) ethylene glycol in the reservoir solution. Diffraction data were collected from a single flash-frozen crystal on the SBC-CAT BM beamline (APS, Argonne National Laboratory). Data were indexed and processed with HKL-3000 [228].

The structure of the apo-form of the catalytic domain was solved by molecular replacement using the PHASER program from the CCP4 software suite, with the structure of the catalytic domain of EglA GH5 endoglucanase from *Piromyces rhizinflata* (PDB code 3AYR) as a search model [229,230]. The refined model of the apo-form structure was used as a search model to solve the structure of the complex between the E154A mutant and cellotriose. The final refined models were obtained by carrying out several cycles consisting of manual model building using COOT, followed by structure refinement with Phaser from the CCP4 software suite. Coordinates have been deposited in the Protein Data Bank (PDB code ….).

Enzymatic Activity Assays

Cellulolytic activity of *P. finnis* CelD on CMC, beechwood xylan, MLG, xyloglucan, phosphoric acid swollen cellulose (PASC), and arabinogalactan was assessed using the dinitrosalicylic acid (DNS) reducing sugar assay essentially as described elsewhere [231].

CMC was purchased from Sigma Aldrich (St. Louis, MO, USA), xyloglucan, MLG, and arabinogalactan were purchased from Fisher Scientific (Waltham, MA, USA), and beechwood xylan was purchased from Megazyme (Bray, Ireland). PASC was prepared as described previously [232]. For specific activity measurements on CMC, MLG, PASC, and xyloglucan, 200 µL reactions in 0.1M sodium acetate pH 5.5 containing CelD at 0.70 µM and substrate at 1% w/v final concentration were statically incubated at 39°C. The same mixture substituting enzyme for acetate buffer was used as a negative control. Three 60 µL samples were taken after 1, 2, and 24 hours of incubation time and their reducing sugar composition was measured using the DNS assay [231]. Briefly, 100 µL of DNS was added to each reaction sample and the mixture incubated at 95°C for 5 minutes. 100 µL of this mixture was added to 100 µL of water and the absorbance at 540 nm measured using a Tecan Infinite® M1000 plate reader (Tecan Group, Mannedorf, Switzerland). A_{540nm} was converted to g/L glucose equivalents with a standard curve of glucose in 0.1M sodium acetate buffer pH 5.5. Specific activities were calculated using protein concentrations measured by A_{280nm} with appropriate parameters and reducing sugar concentrations measured by DNS assay using standard curves of glucose. Absorbance measurements were blank subtracted by a negative control of substrate without enzyme.

For specific activity measurements on xylan and arabinogalactan, 30 µL of protein in 0.1M sodium acetate pH 5.5 (0.1 mg total protein) was added to 30 µL of 2% (w/v) freshly prepared, unautoclaved polysaccharide solution in 0.1 M sodium acetate (pH 5.5). Reactions were performed at 39°C unless otherwise stated and in triplicate, with reaction times of 45 minutes for xylan and 14 hours for arabinogalactan.

Kinetic parameters for CelD on CMC were obtained by adding 5 µL enzyme in 0.1M sodium acetate pH 5.5 to 195 µL of pre-warmed CMC substrate at concentrations of 0 – 30 g/L CMC to a final enzyme concentration of 0.05 µM. Enzyme-substrate mixtures were incubated at 39°C with shaking at 188 RPM. Three 60 µL sample were taken after 5 – 35 minutes of incubation time and their reducing sugar composition was measured as described above. Initial rates were extracted by linear regression of the A_{540nm} vs time curve and converted to the appropriate units using a glucose standard curve after subtraction of A_{540nm} signal from a substrate only control. Initial rate vs substrate concentration data were then fit to a Michaelis-Menten model by nonlinear regression using the Scipy Python package to determine k_{cat} and K_m parameters for each GH5 variant.

β-Glucosidase and β-galactosidase activities were assessed by adding 30 µL of protein (0.1 mg total protein) to 970 µL of 5 mM 4-nitrophenyl β-D-glucopyranoside (pNPG) or 4-nitrophenyl β-D-galactopyranoside (*p*NPGal) in 50 mM sodium phosphate (dibasic) buffer (pH 7.0) with 2% (w/v) bovine serum albumin. Absorbance at 405 nm as measured by a Tecan Infinite® M1000 plate reader tracked reaction progression over 24 hours.

CotH kinase activity was assays using an Adapta Universal Kinase Activity Assay kit (ThermoFisher Scientific). For each *in vitro* reaction, purified CotH was supplied at 25 µg/mL with 0.75 mg/mL substrate (if applicable) and 0.35 mM ATP in 1x kinase buffer A. Reactions were performed in triplicate at 10µL total volume in a 384-well microplate with incubation at 30°C for 1 hour. MyeBP (sequence NKRPSNRSKYL) was procured from VWR (part no. 89143-546) and H1-7 (sequence RRKASGP) was procured from Enzo Life Sciences (part no. 89160-930).

VI. Conclusions

6.1 Summary and Perspectives

6.1.1 Development and application of bioimaging tools for studying cellulosomes in native anaerobic fungal cultures

As the major means of generating carbon for primary metabolism, cellulosomes are crucial pieces of anaerobic fungal biology that enable cellular growth and reproduction. Besides clear evidence that cellulosomes are efficient lignocellulose degrading machines, we understand very little about the processes of cellulosome production and deployment, especially in the context of the multi-staged life cycle through which anaerobic fungi proliferate. Additionally, annotated dockerin containing genes in anaerobic fungal genomes point to many cellulosome-incorporating proteins with functions unrelated to biomass degradation, raising questions about the scope of cellulosome function in anaerobic fungal biology. Where anaerobic fungal cells localize cellulosomes, and which life stages produce them are key questions to address towards gaining a better understanding of how anaerobic fungi degrade biomass and how cellulosomes participate in anaerobic fungal biology.

In Chapter 2, we use immunofluorescence microscopy and newly developed, cellulosome-specific antibodies to measure cellulosome localization across stages of the fungal life cycle, in several *Neocallimastigomycota* species, when anaerobic fungi are grown on cellulosic and non-cellulosic substrates. We uncover for the first time that cellulosome proteins are produced by anaerobic fungi at all stages of the fungal life cycle when grown on cellulosic substrates, and that cellulosome localization shifts from the spherical zoospore surface to the biomass-interfacing rhizoids as cells mature. We also observed non-cellulosic, soluble sugar substrates to induce cellulosome protein deployment only in zoospores and not mature thalli or zoosporangia. These insights provide clear, interesting observations through which to interpret future investigations of zoospore biology and anaerobic fungal cellular development, which are important processes to grasp as anaerobic fungi are domesticated for useful means.

6.1.2 Native cellulosome purification and structural/functional characterization

Anaerobic fungi are attractive as sources for innovative bioprocessing technologies because of their demonstrated, natural excellence as crude biomass degraders. Fungal cellulosomes account for the great majority of hydrolysis activity in anaerobic fungal supernatants, but the enzyme content and biochemical details explaining how fungal cellulosomes so efficiently degrade biomass are lacking. Furthermore, the mechanism by which fungal cellulosomes assemble is still unsolved – a major barrier to reconstituting heterologous fungal cellulosomes for detailed investigation into cellulosome function.

The key challenge in elucidating fungal cellulosome structure, composition, and biochemistry has been cellulosome purification; cellulosome complexes have not been purified at sufficient purity or yield to enable structural or functional study. In Chapter 3, we introduce a cellulosome purification method capable of producing, at high yield, cellulosome complexes for structural analysis by cryo-EM and kinetic analysis by nanostructure initiator mass spectrometry (NIMS). We use this method to investigate how *N. californiae* alters the composition of its cellulosome when grown on substrates with different physical and chemical characteristics, and also measure how these cellulosome complexes with different compositions perform in hydrolyzing cellulose, hemicellulose, and lignocellulose.

While there is still ample room for method optimization, the method presented in this work is likely to serve as a workhorse towards solving the first fungal cellulosome structure. The insights into cellulosome composition suggest the collection of enzyme activities fungal cellulosomes employ to efficiently degrade lignocellulose. In combination with cellulosome hydrolysis kinetics data presented here, these cellulosome composition measurements additionally provide a path towards rationally optimizing cellulosome content for degrading particular substrates, towards developing fungal cellulosomes for industrial bioprocessing. The clear observation of purified fungal cellulosome complexes by cryo-EM lays to rest any controversy over whether anaerobic fungi produce cellulosomes like bacteria do and represents a small but critical step towards elucidating the minimal protein-protein interaction mediating fungal cellulosome assembly.

6.1.3 Synthetic protein complex engineering with fungal cellulosome parts

The modular assembly framework by which fungal cellulosomes assemble is an attractive platform for constructing synthetic protein complexes for synthetic biology applications. However, the unknown identity of dockerin's binding partner, cohesin, has made it impossible to build protein complexes from fungal cellulosome parts. In Chapter 4, we develop a synthetic fungal dockerin binding partner, a double dockerin binding nanobody, with which to build protein complexes. Using a simulation-guided approach, we design artificial "scaffoldin" proteins comprising multiple nanobody domains that form multimeric complexes with dockerin-containing proteins. Towards engineering protein complexes with stimuli-responsive assembly behavior, we engineer dockerin-nanobody pairs with pH-dependent binding behavior, showcasing the potential of the dockerin-nanobody system as a toolkit for building designer, stimuli-responsive protein complexes.

These novel parts provide, for the first time, a framework for reconstituting fungal cellulosomes *in vitro*, which enables direct interrogation of cellulosome composition-activity relationships and the construction of designer fungal cellulosomes. The suite of stimuli-responsive assembly parts additionally presents a platform for constructing protein complexes that post-translationally “shuffle” their composition in response to environmental changes, a novel approach to reversible, post-translational control over multi-protein system function. Furthermore, the combined computational and experimental approach we employ to engineer dockerins with stimuli-responsive nanobody binding behavior provides a general strategy for future efforts seeking to engineer fungal cellulosome parts with other properties, such as binding sensitivity to other environmental factors, affinity enhancements, or binding specificity.

6.1.4 Characterization of heterologously produced anaerobic fungal proteins

Very few proteins from anaerobic fungi are experimentally characterized, which poses a challenge to gaining a mechanistic understanding of anaerobic fungal biology. For cellulosomes specifically, many cellulosomal genes are annotated with functions unrelated to biomass degradation, and sequence-based annotation is notoriously poor at categorizing specific enzyme activity (e.g. endoglucanase vs endoxylanase activity) in CAZymes. Thus, without functional data to supplement genome annotations, understanding what biological functions cellulosomes participate in, how cellulosomes hydrolyze lignocellulose, and which cellulosomal enzymes account for which enzyme activities is very difficult. In Chapter 5, we biochemically characterize gene products for two important families of cellulosomal enzymes, proteins annotated as CotH spore coat protein kinases and a Glycoside Hydrolase family 5 (GH5) cellulase. CotH spore coat kinases are an unusual component of fungal cellulosomes with no known biological function, but we show two anaerobic fungal CotH proteins hydrolyze ATP and likely act as protein kinases in a small step towards understanding their function. We present a structure and enzyme kinetic data for a GH5, one of the most prominent enzyme families in fungal cellulosomes, from *Piromyces finnis*, showing it possesses endoglucanase activity on par with many other fungal endoglucanases. While small, the advancements in our understanding of fungal cellulosome function gained from this work are tangible and provide precedent for continued characterization of anaerobic fungal proteins in heterologous systems.

6.2 Future Directions

While the results and methodologies presented here represent important progress towards characterizing and engineering fungal cellulosomes, there is much work to be done to fully realize the biotechnological potential of fungal cellulosomes. As an extension to Chapter 2, the application of complementary microscopy techniques, along with other cellulosome labeling imaging probes, will go a long way in elucidating the processes of cellulosome production, secretion, and deployment *in vivo*. Much of the anaerobic fungal cell biology knowledge we have comes from electron microscopy in the 1980's. Significantly improved electron microscope technology will likely enable the visualization within cells of the organelles critical to protein production and secretion and can perhaps lend unforeseen insight into basic biological processes like rhizoid development that are key to fungal growth. Such techniques enabled significant advancements in understanding hyphae growth-coupled protein secretion processes in filamentous fungi widely relevant to industrial enzyme production [97]. Imaging tools compatible with live cell imaging, such as the nanobody developed in Chapter 4, also present an attractive path by which to learn more about anaerobic fungal cell biology. Dockerin-binding nanobodies functionalized with quantum dots can be delivered into live cells via electroporation, enabling the tracking of cellulosome proteins within live cells [233]. Nanobodies developed against unexpected cellulosome members, like CotH spore coat kinases, could help elucidate the function of these proteins in cellulosome biology when used with live cell imaging.

Chapter 3 presents in depth characterization of cellulosomes from *Neocallimastix californiae*, but not from any other anaerobic fungi because first attempts to purify cellulosomes from other anaerobic fungal genera using the same method did not produce clear high MW complexes. This method should be tested again on *Anaeromyces robustus* and a *Piromyces* fungus to confirm these results, especially with Native PAGE, a higher resolution size separation technique than SEC. Additionally, while the presented method worked well enough, it is long and labor intensive because of the affinity digestion component. Since cellulosome complexes are purified from the fungal supernatant, it would be prudent to explore using a very high MW (500 kDa or more MW cutoff) filtration device to simply concentrate cellulosomes from growth supernatant and apply the filtrate to an SEC column. It should be noted that this is what the bacterial cellulosome field has evolved towards, so I would expect it to work [51], but it is possible a solid cellulose support could be required to initiate fungal cellulosome assembly, meaning supernatant filtration without any adsorption step would not work.

Chapter 4 presents a synthetic dockerin binding partner with which to build synthetic fungal cellulosomes and demonstrates how these parts can be engineered to make designer protein complexes with stimuli-responsive composition. Resulting from this work is a set of parts - a tethered nanobody “scaffoldin” capable of forming trimers with two dockerin proteins, and several dockerins with pH-sensitive binding activity to the nanobody. However, a full demonstration of a synthetic complex with predictable, pH-dependent composition, remains to be done. Designer protein complexes with stimuli-dependent composition are a novel concept and a first priority should be building a theoretical framework for predicting how these systems will behave as a function of protein concentration, pH, and binding thermodynamic parameters. A likely complication could be avidity, where one dockerin can bind each of multiple nanobody sites within the scaffoldin, conflicting with the monomeric stimuli-responsive binding phenotype. Applying a similar yeast surface display-based approach to engineer dockerin-nanobody pairs with high specificities will be important for producing protein complexes with the predicted stimuli-responsive phenotype. Furthermore, binding pairs that are not cross-reactive will also be more generally useful in constructing protein complexes that are easily modeled and for controlling protein organization within complexes rather than relying on random assembly. Perhaps in parallel with the above work, functional demonstrations of synthetic mini-cellulosomes with the dockerin-nanobody system would also be valuable in proving the utility of this system.

6.3 Overall Conclusions

In all, this work contributes to the ongoing efforts towards realizing the biotechnological potential of anaerobic fungi and their cellulosome machinery. Microscopy-based techniques provided insights into cellulosome production and deployment as processes in the anaerobic fungal cell life cycle. A developed cellulosome purification methodology and characterization of key cellulosomal enzymes enabled detailed composition information and structural and biochemical data on how native cellulosomes from the anaerobic fungus

Neocallimastix californiae efficiently hydrolyze lignocellulose. A synthetic dockerin binding partner provided the first platform by which to build protein complexes with fungal cellulosome enzymes, and engineered variants of these parts with stimuli-responsive assembly behavior provide a novel framework for controlling protein complex function via environmentally controlled composition "shuffling." Though much remains to be done to translate fungal cellulosomes into biotechnologies, their potential in a range of applications in bioprocessing and beyond is clear and will likely only become more apparent as native fungal cellulosomes are better understood and as engineered fungal cellulosomes are explored.

References

1. Lillington, S. P. *et al.* Cellulosome Localization Patterns Vary across Life Stages of Anaerobic Fungi. *mBio* **12**, e00832-21 (2021).
2. Biggs, B. W. *et al.* Enabling commercial success of industrial biotechnology. *Science* **374**, 1563–1565 (2021).
3. Aden, A. *et al. Lignocellulosic Biomass to Ethanol Process Design and Economics Utilizing Co-current Dilute Acid Prehydrolysis and Enzymatic Hydrolysis for Corn Stover*. (2002).
4. Lillington, S. P., Leggieri, P. A., Heom, K. A. & O'Malley, M. A. Nature's recyclers: anaerobic microbial communities drive crude biomass deconstruction. *Curr. Opin. Biotechnol.* **62**, 38–47 (2020).
5. Haitjema, C. H., Solomon, K. V., Henske, J. K., Theodorou, M. K. & O'Malley, M. A. Anaerobic gut fungi: Advances in isolation, culture, and cellulolytic enzyme discovery for biofuel production. *Biotechnol. Bioeng.* **111**, 1471–1482 (2014).
6. Fontes, C. M. G. A. & Gilbert, H. J. Cellulosomes: Highly Efficient Nanomachines Designed to Deconstruct Plant Cell Wall Complex Carbohydrates. *Annu. Rev. Biochem.* **79**, 655–681 (2010).
7. Gilmore, S. P., Lillington, S. P., Haitjema, C. H., de Groot, R. & O'Malley, M. A. Designing chimeric enzymes inspired by fungal cellulosomes. *Synth. Syst. Biotechnol.* **5**, 23–32 (2020).
8. Wheeldon, I. *et al.* Substrate channelling as an approach to cascade reactions. *Nat. Chem.* **8**, 299–309 (2016).
9. Lee, H., DeLoache, W. C. & Dueber, J. E. Spatial organization of enzymes for metabolic engineering. *Metab. Eng.* **14**, 242–251 (2012).
10. B. Quin, M., K. Wallin, K., Zhang, G. & Schmidt-Dannert, C. Spatial organization of multi-enzyme biocatalytic cascades. *Org. Biomol. Chem.* **15**, 4260–4271 (2017).

11. Levasseur, A. *et al.* Design and Production in Aspergillus niger of a Chimeric Protein Associating a Fungal Feruloyl Esterase and a Clostridial Dockerin Domain. *Appl. Environ. Microbiol.* **70**, 6984–6991 (2004).
12. Moraïs, S. *et al.* Deconstruction of Lignocellulose into Soluble Sugars by Native and Designer Cellulosomes. *mBio* **3**, e00508-12 (2012).
13. Fenchel, T. Microbial Behavior in a Heterogeneous World. vol. 296 1068–1071 (2002).
14. Bian, X.-Y. *et al.* Insights into the Anaerobic Biodegradation Pathway of n-Alkanes in Oil Reservoirs by Detection of Signature Metabolites. *Sci. Rep.* **5**, (2015).
15. Ren, Y. *et al.* A comprehensive review on food waste anaerobic digestion: Research updates and tendencies. *Bioresour. Technol.* **247**, 1069–1076 (2018).
16. Peng, X. N., Gilmore, S. P. & O'Malley, M. A. Microbial communities for bioprocessing: lessons learned from nature. *Curr. Opin. Chem. Eng.* **14**, 103–109 (2016).
17. Seshadri, R. *et al.* Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nat. Biotechnol.* **36**, (2018).
18. Podolsky, I. A. *et al.* Harnessing Nature's Anaerobes for Biotechnology and Bioprocessing. *Annu. Rev. Chem. Biomol. Eng.* **10:null**, (2019).
19. Agu, C. V., Ujor, V. & Ezeji, T. C. Metabolic engineering of Clostridium beijerinckii to improve glycerol metabolism and furfural tolerance. *Biotechnol. Biofuels* **12**, (2019).
20. Charubin, K., Bennett, R. K., Fast, A. G. & Papoutsakis, E. T. Engineering Clostridium organisms as microbial cell-factories: challenges & opportunities. *Metab. Eng.* **50**, 173–191 (2018).
21. Edwards, J. E. *et al.* PCR and Omics Based Techniques to Study the Diversity, Ecology and Biology of Anaerobic Fungi: Insights, Challenges and Opportunities. *Front. Microbiol.* **8**, (2017).
22. Hess, M. *et al.* Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. *Science* **331**, 463–467 (2011).
23. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* **9**, (2018).
24. Roussel, E. G. *et al.* Extending the Sub-Sea-Floor Biosphere. vol. 320 1046–1046 (2008).
25. Ulloa, O., Canfield, D. E., DeLong, E. F., Letelier, R. M. & Stewart, F. J. Microbial oceanography of anoxic oxygen minimum zones. vol. 109 15996–16003 (2012).
26. Arrigo, K. R. Marine microorganisms and global nutrient cycles. *Nature* **437**, 349–355 (2005).
27. Lin, S. Y. Accessibility of cellulose: A critical review. *Fibre Sci. Technol.* **5**, 303–314 (1972).
28. Ralph, J. & Helm, R. Lignin/Hydroxycinnamic Acid/Polysaccharide Complexes: Synthetic Models for Regiochemical Characterization. in *Forage Cell Wall Structure and Digestibility* (1993).
29. Pollegioni, L., Tonin, F. & Rosini, E. Lignin-degrading enzymes. *FEBS J.* **282**, 1190–1213 (2015).
30. Chen, H. Chemical Composition and Structure of Natural Lignocellulose. in *Biotechnology of Lignocellulose: Theory and Practice* 25–71 (Springer Netherlands, 2014). doi:10.1007/978-94-007-6898-7_2.

31. Iiyama, K., Lam, T. B. T. & Stone, B. A. Covalent Cross-Links in the Cell Wall. *Plant Physiol.* **104**, 315–320 (1994).
32. Zhu, N. *et al.* Metagenomic and metaproteomic analyses of a corn stover-adapted microbial consortium EMSD5 reveal its taxonomic and enzymatic basis for degrading lignocellulose. *Biotechnol. Biofuels* **9**, 1–23 (2016).
33. D, L. *et al.* The metagenome of an anaerobic microbial community decomposing poplar wood chips. *PLoS ONE* 7 (2012).
34. Comtet-marre, S. *et al.* Metatranscriptomics Reveals the Active Bacterial and Eukaryotic Fibrolytic Communities in the Rumen of Dairy Cow Fed a Mixed Diet. Frontiers in Microbiology. *Front. Microbiol.* 8 (2017).
35. Cantarel, B. I. *et al.* The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Res.* **37**, 233–238 (2008).
36. Lombard, V., Ramulu, H. G., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, 490–495 (2014).
37. Speda, J., Jonsson, B. H., Carlsson, U. & Karlsson, M. Metaproteomics-guided selection of targeted enzymes for bioprospecting of mixed microbial communities. *Biotechnol. Biofuels* **10**, 1–17 (2017).
38. Helbert, W. *et al.* Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space. *Proc. Natl. Acad. Sci.* **116**, (2019).
39. Alessi, A. M. *et al.* Defining functional diversity for lignocellulose degradation in a microbial community using multi-omics studies. *Biotechnol. Biofuels* **11**, 1–16 (2018).
40. Solden, L. M. *et al.* Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. *Nat. Microbiol.* **3**, 1274–1284 (2018).
41. Krauss, J., Zverlov, S., VV, & W.H. In Vitro Reconstitution of the Complete Clostridium thermocellum Cellulosome and Synergistic Activity on Crystalline Cellulose. *Appl. Environ. Microbiol.* **78**, 4301–4307 (2012).
42. Solomon, K. V. *et al.* Early-branching gut fungi possess a large, comprehensive array of biomass-degrading enzymes. *Science* **351**, 1192–1195 (2016).
43. Brunecky, R. *et al.* Revealing Nature's Cellulase Diversity: The Digestion Mechanism of Caldicellulosiruptor bescii CelA. *Science* **342**, 1513–1516 (2013).
44. DeBoy, R. T. *et al.* Insights into Plant Cell Wall Degradation from the Genome Sequence of the Soil Bacterium Cellvibrio japonicus. *J. Bacteriol.* **190**, 5455–5463 (2008).
45. Lamed, R., Setter, E. & Bayer, E. A. Characterization of a cellulose-binding, cellulase-containing complex in clostridium thermocellum. *J. Bacteriol.* **156**, 828–836 (1983).
46. Carvalho, A. L. *et al.* Cellulosome assembly revealed by the crystal structure of the cohesin–dockerin complex. *Proc. Natl. Acad. Sci.* **100**, 13809–13814 (2003).
47. Artzi, L., Bayer, E. A. & Moraïs, S. Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. *Nat. Rev. Microbiol.* **15**, 83–95 (2017).
48. Carvalho, A. L. *et al.* Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 13809–13814 (2003).
49. 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.* **103**, 305–310 (2006).

50. Xu, Q. *et al.* Cell Biology: Dramatic performance of Clostridium thermocellum explained by its wide range of cellulase modalities. *Sci. Adv.* **2**, (2016).
51. Artzi, L., Morag, E., Barak, Y., Lamed, R. & Bayer, E. A. Clostridium clariflavum: Key cellulosome players are revealed by proteomic analysis. *mBio* **6**, 1–12 (2015).
52. Ichikawa, S. *et al.* Cellulosomes localise on the surface of membrane vesicles from the cellulolytic bacterium Clostridium thermocellum. *FEMS Microbiol. Lett.* **366**, 1–9 (2019).
53. Fierobe, H. P. *et al.* Action of designer cellulosomes on homogeneous Versus complex substrates: Controlled incorporation of three distinct enzymes into a defined trifunctional scaffoldin. *J. Biol. Chem.* **280**, 16325–16334 (2005).
54. Moraïs, S. *et al.* Assembly of xylanases into designer cellulosomes promotes efficient hydrolysis of the xylan component of a natural recalcitrant cellulosic substrate. *mBio* **2**, 1–11 (2011).
55. 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**, (2009).
56. Wilson, C. A. & Wood, T. M. Studies on the cellulase of the rumen anaerobic fungus Neocallimastix frontalis, with special reference to the capacity of the enzyme to degrade crystalline cellulose. *Enzyme Microb. Technol.* **14**, 258–264 (1992).
57. Ali, B. R. S. *et al.* Cellulases and hemicellulases of the anaerobic fungus Piromyces constitute a multiprotein cellulose-binding complex and are encoded by multigene families. *FEMS Microbiol. Lett.* **125**, 15–21 (1995).
58. Fanutti, C., Ponyi, T., Black, G. W., Hazlewood, G. P. & Gilbert, H. J. The Conserved Noncatalytic 40-Residue Sequence in Cellulases and Hemicellulases from Anaerobic Fungi Functions as a Protein Docking Domain. *J. Biol. Chem.* **270**, 29314–29322 (1995).
59. Raghothama, S. *et al.* Characterization of a cellulosome dockerin domain from the anaerobic fungus Piromyces equi. *Nat. Struct. Biol.* **8**, 775–778 (2001).
60. Fillingham, I. J., Kroon, P. A., Williamson, G., Gilbert, H. J. & Hazlewood, G. P. A modular cinnamoyl ester hydrolase from the anaerobic fungus Piromyces equi acts synergistically with xylanase and is part of a multiprotein cellulose-binding cellulase–hemicellulase complex. *Biochem J* **224**, 215–224 (1999).
61. Dalrymple, B. P. *et al.* Three Neocallimastix patriciarum esterases associated with the degradation of complex polysaccharides are members of a new family of hydrolases. *Microbiology* **143**, 2605–2614 (1997).
62. Dijkerman, R., Op Den Camp, H. J. M., Van Der Drift, C. & Vogels, G. D. The role of the cellulolytic high molecular mass (HMM) complex of the anaerobic fungus Piromyces sp. strain E2 in the hydrolysis of microcrystalline cellulose. *Arch. Microbiol.* **167**, 137–142 (1997).
63. Gordon, G. L. R. & Phillips, M. W. The role of anaerobic gut fungi in ruminants. *Nutr. Res. Rev.* **11**, 133–168 (1998).
64. Ford, C. W., Elliott, R. & Maynard, P. J. The effect of chlorite delignification on digestibility of some grass forages and on intake and rumen microbial activity in sheep fed barley straw. *J. Agric. Sci.* **108**, 129–136 (1987).

65. Gordon, G. L. R. & Phillips, M. W. Removal of anaerobic fungi from the rumen of sheep by chemical treatment and the effect on feed consumption and in vivo fibre digestion. *Lett. Appl. Microbiol.* **17**, 220–223 (1993).
66. Orpin, C. G. Studies on the Rumen Flagellate Neocallimastix frontalis. *Microbiology* **91**, 249–262 (1975).
67. Laundon, D., Chrismas, N., Wheeler, G. & Cunliffe, M. Chytrid rhizoid morphogenesis resembles hyphal development in multicellular fungi and is adaptive to resource availability. *Proc. R. Soc. B Biol. Sci.* **287**, 20200433 (2020).
68. Gold, J. J., Heath, I. B. & Bauchop, T. Ultrastructural description of a new chytrid genus of caecum anaerobe, Caecomyces equi gen. nov., sp. nov., assigned to the Neocallimasticaceae. *Biosystems* **21**, 403–415 (1988).
69. Trinci, A. P. J. *et al.* Anaerobic fungi in herbivorous animals. *Mycol. Res.* **98**, 129–152 (1994).
70. Miles, E. W., Rhee, S. & Davies, D. R. The molecular basis of substrate channeling. *J. Biol. Chem.* **274**, 12193–12196 (1999).
71. Wu, N., Tsuji, S. Y., Cane, D. E. & Khosla, C. Assessing the balance between protein-protein interactions and enzyme-substrate interactions in the channeling of intermediates between polyketide synthase modules. *J. Am. Chem. Soc.* **123**, 6465–6474 (2001).
72. Shoham, Y., Lamed, R. & Bayer, E. A. The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides. *Trends Microbiol.* **7**, 275–281 (1999).
73. Elion, E. A. The Ste5p scaffold. *J. Cell Sci.* **114**, 3967–3978 (2001).
74. Bhattacharyya, R. P., Reményi, A., Yeh, B. J. & Lim, W. A. Domains, motifs, and scaffolds: The role of modular interactions in the evolution and wiring of cell signaling circuits. *Annu. Rev. Biochem.* **75**, 655–680 (2006).
75. Dueber, J. E. *et al.* Synthetic Protein Scaffolds Provide Modular Control over Metabolic Flux. *Nat. Biotechnol.* **27**, 5–8 (2009).
76. You, C. & Zhang, Y.-H. P. Self-Assembly of Synthetic Metabolons through Synthetic Protein Scaffolds: One-Step Purification, Co-immobilization, and Substrate Channeling. *ACS Synth. Biol.* **2**, 102–110 (2013).
77. You, C., Myung, S. & Zhang, Y. H. P. Facilitated substrate channeling in a self-assembled trifunctional enzyme complex. *Angew. Chem. - Int. Ed.* **51**, 8787–8790 (2012).
78. Tsai, S. L., DaSilva, N. A. & Chen, W. Functional Display of Complex Cellulosomes on the Yeast Surface via Adaptive Assembly. *ACS Synth. Biol.* **2**, 14–21 (2013).
79. Borne, R., Bayer, E. A., Pagès, S., Perret, S. & Fierobe, H.-P. Unraveling enzyme discrimination during cellulosome assembly independent of cohesin-dockerin affinity. *FEBS J.* **280**, 5764–5779 (2013).
80. Vazana, Y. *et al.* A synthetic biology approach for evaluating the functional contribution of designer cellulosome components to deconstruction of cellulosic substrates. *Biotechnol. Biofuels* **6**, 182 (2013).
81. Gilmore, S. P., Henske, J. K. & O’Malley, M. A. Driving biomass breakdown through engineered cellulosomes. *Bioengineered* **6**, 204–208 (2015).
82. Haitjema, C. H. *et al.* A parts list for fungal cellulosomes revealed by comparative genomics. *Nat. Microbiol.* **2**, 1–8 (2017).

83. Haitjema, C. H. *et al.* A parts list for fungal cellulosomes revealed by comparative genomics. *Nat. Microbiol.* **2**, 17087 (2017).
84. Brown, J. L. *et al.* Co-cultivation of the anaerobic fungus Caecomyces churrovis with Methanobacterium bryantii enhances transcription of carbohydrate binding modules, dockerins, and pyruvate formate lyases on specific substrates. *Biotechnol. Biofuels* **14**, 1–16 (2021).
85. Wilken, S. E. *et al.* Experimentally validated reconstruction and analysis of a genome-scale metabolic model of an anaerobic Neocallimastigomycota fungus. *mSystems* **6**, 1–22 (2021).
86. Gilmore, S. P., Lillington, S. P., Haitjema, C. H., de Groot, R. & O'Malley, M. A. Designing chimeric enzymes inspired by fungal cellulosomes. *Synth. Syst. Biotechnol.* **5**, 23–32 (2020).
87. Orpin, C. G. Invasion of Plant Tissue in the Rumen by the Flagellate Neocallimastix frontalis. *Microbiology* **98**, 423–430 (1977).
88. Bauchop, T. Rumen Anaerobic Fungi of Cattle and Sheep. *Appl. Environ. Microbiol.* **38**, 148–158 (1979).
89. Liggenstoffer, Youssef, N. H., Couger, M. B. & Elshahed, M. S. Phylogenetic diversity and community structure of anaerobic gut fungi (phylum Neocallimastigomycota) in ruminant and non-ruminant herbivores. *ISME J.* **4**, 1225–1235 (2010).
90. Hanafy, R. A. *et al.* Seven new Neocallimastigomycota genera from wild, zoo-housed, and domesticated herbivores greatly expand the taxonomic diversity of the phylum. *Mycologia* **112**, 1212–1239 (2020).
91. Henske, J. K. *et al.* Transcriptomic characterization of Caecomyces churrovis: a novel, non-rhizoid-forming lignocellulolytic anaerobic fungus. *Biotechnol. Biofuels* **10**, 305 (2017).
92. Youssef, N. H. *et al.* The Genome of the Anaerobic Fungus Orpinomyces sp. Strain C1A Reveals the Unique Evolutionary History of a Remarkable Plant Biomass Degrader. *Appl. Environ. Microbiol.* **79**, 4620–4634 (2013).
93. Wilken, St. E. *et al.* Experimentally Validated Reconstruction and Analysis of a Genome-Scale Metabolic Model of an Anaerobic Neocallimastigomycota Fungus. *mSystems* **6**, e00002-21 (2021).
94. Seppälä, S., Wilken, St. E., Knop, D., Solomon, K. V. & O'Malley, M. A. The importance of sourcing enzymes from non-conventional fungi for metabolic engineering and biomass breakdown. *Metab. Eng.* **44**, 45–59 (2017).
95. Bayer, E. A. & Lamed, R. Ultrastructure of the cell surface cellulosome of Clostridium thermocellum and its interaction with cellulose. *J. Bacteriol.* **167**, 828–836 (1986).
96. Ljungdahl, L. G. *et al.* Cellulosomes of anaerobic fungi. in *Cellulosome* (Nova Science Publishers Inc, 2006).
97. Wösten, H. A. B., Moukha, S. M., Sietsma, J. H. & Wessels, J. G. H. Localization of growth and secretion of proteins in Aspergillus niger. *Microbiology* **137**, 2017–2023 (1991).
98. Calkins, S. S. *et al.* Development of an RNA interference (RNAi) gene knockdown protocol in the anaerobic gut fungus Pecoramyces ruminantium strain C1A. *PeerJ* **6**, e4276 (2018).

99. Gruninger, R. J. *et al.* Anaerobic fungi (phylum Neocallimastigomycota): advances in understanding their taxonomy, life cycle, ecology, role and biotechnological potential. *FEMS Microbiol. Ecol.* **90**, 1–17 (2014).

100. Wilson, C. A. & Wood, T. M. Studies on the cellulase of the rumen anaerobic fungus Neocallimastix frontalis, with special reference to the capacity of the enzyme to degrade crystalline cellulose. *Enzyme Microb. Technol.* **14**, 258–264 (1992).

101. Dijkerman, R., Vervuren, M. B., Op Den Camp, H. J. & van der Drift, C. Adsorption characteristics of cellulolytic enzymes from the anaerobic fungus Piromyces sp. strain E2 on microcrystalline cellulose. *Appl. Environ. Microbiol.* **62**, 20–25 (1996).

102. Dijkerman, R., Bhansing, D. C. P., Op den Camp, H. J. M., van der Drift, C. & Vogels, G. D. Degradation of structural polysaccharides by the plant cell-wall degrading enzyme system from anaerobic fungi: An application study. *Enzyme Microb. Technol.* **21**, 130–136 (1997).

103. Dijkerman, R., Op den Camp, H. J. M., Van der Drift, C. & Vogels, G. D. The role of the cellulolytic high molecular mass (HMM) complex of the anaerobic fungus Piromyces sp. strain E2 in the hydrolysis of microcrystalline cellulose. *Arch. Microbiol.* **167**, 137–142 (1997).

104. Ho, Y. W., Abdullah, N. & Jalaludin, S. Penetrating Structures of Anaerobic Rumen Fungi in Cattle and Swamp Buffalo. *Microbiology* **134**, 177–181 (1988).

105. Baral, N. R. *et al.* Approaches for More Efficient Biological Conversion of Lignocellulosic Feedstocks to Biofuels and Bioproducts. *ACS Sustain. Chem. Eng.* **7**, 9062–9079 (2019).

106. Steenbakkers, P. J. M. *et al.* The Major Component of the Cellulosomes of Anaerobic Fungi from the Genus Piromyces is a Family 48 Glycoside Hydrolase. *DNA Seq.* **13**, 313–320 (2002).

107. Lowe, S. E., Griffith, G. G., Milne, A., Theodorou, M. K. & Trinci, A. P. J. The Life Cycle and Growth Kinetics of an Anaerobic Rumen Fungus. *Microbiology* **133**, 1815–1827 (1987).

108. Lowe, S. E., Theodorou, M. K. & Trinci, A. P. Cellulases and xylanase of an anaerobic rumen fungus grown on wheat straw, wheat straw holocellulose, cellulose, and xylan. *Appl. Environ. Microbiol.* **53**, 1216–1223 (1987).

109. Williams, A. G. & Orpin, C. G. Polysaccharide-degrading enzymes formed by three species of anaerobic rumen fungi grown on a range of carbohydrate substrates. *Can. J. Microbiol.* **33**, 418–426 (1987).

110. Lamed, R., Morag (Morgenstern), E., Mor-Yosef, O. & Bayer, E. A. Cellulosome-like entities inBacteroides cellulosolvens. *Curr. Microbiol.* **22**, 27–33 (1991).

111. Bayer, E. A., Kenig, R. & Lamed, R. Adherence of Clostridium thermocellum to cellulose. *J. Bacteriol.* **156**, 818–827 (1983).

112. Salama-Alber, O. *et al.* Atypical Cohesin-Dockerin Complex Responsible for Cell Surface Attachment of Cellulosomal Components: BINDING FIDELITY, PROMISCUITY, AND STRUCTURAL BUTTRESSES *. *J. Biol. Chem.* **288**, 16827–16838 (2013).

113. Henske, J. K., Gilmore, S. P., Haitjema, C. H., Solomon, K. V. & O’Malley, M. A. Biomass-degrading enzymes are catabolite repressed in anaerobic gut fungi. *AIChE J.* **64**, 4263–4270 (2018).

114. Rosenblum, E. B., Stajich, J. E., Maddox, N. & Eisen, M. B. Global gene expression profiles for life stages of the deadly amphibian pathogen Batrachochytrium dendrobatidis. *Proc. Natl. Acad. Sci.* **105**, 17034–17039 (2008).

115. Dee, J. M., Mollicone, M., Longcore, J. E., Roberson, R. W. & Berbee, M. L. Cytology and molecular phylogenetics of Monoblepharidomycetes provide evidence for multiple independent origins of the hyphal habit in the Fungi. *Mycologia* **107**, 710–728 (2015).

116. Torralba, S., Raudaskoski, M., Pedregosa, A. M. & Laborda, F. Effect of cytochalasin A on apical growth, actin cytoskeleton organization and enzyme secretion in Aspergillus nidulans. *Microbiology* **144**, 45–53 (1998).

117. Theodorou, M. K., Brookman, J. & Trinci, A. P. J. Anaerobic fungi. in *Methods in gut microbial ecology for ruminants* 55–66 (Springer, 2005).

118. Aden, A. & Foust, T. Technoeconomic analysis of the dilute sulfuric acid and enzymatic hydrolysis process for the conversion of corn stover to ethanol. *Cellulose* **16**, 535–545 (2009).

119. Aden, A. *et al. Lignocellulosic Biomass to Ethanol Process Design and Economics Utilizing Co-current Dilute Acid Prehydrolysis and Enzymatic Hydrolysis for Corn Stover*. *National Renewable Energy Laboratory Report* vol. TP-510-324 (2002).

120. Lillington, S. P. *et al.* Cellulosome localization patterns vary across life stages of anaerobic fungi. *mBio* **12**, (2021).

121. Uversky, V. N. & Kataeva, I. A. *Cellulosome*. (Nova Science Publishers Inc, 2006).

122. Gilbert, H. J. *et al.* Characterization of a cellulosome dockerin domain from the anaerobic fungus Piromyces equi. *Nat. Struct. Biol.* **8**, 775–778 (2001).

123. Brown, J. L. *et al.* Co-cultivation of the anaerobic fungus Caecomyces churrovis with Methanobacterium bryantii enhances transcription of carbohydrate binding modules, dockerins, and pyruvate formate lyases on specific substrates. *Biotechnol. Biofuels* **14**, 234 (2021).

124. Lankiewicz, T. S., Lillington, S. P. & O'Malley, M. A. Enzyme Discovery in Anaerobic Fungi (Neocallimastigomycetes) Enables Lignocellulosic Biorefinery Innovation. *Microbiol. Mol. Biol. Rev.* **86**, e00041-22.

125. Meyer, A. S., Rosgaard, L. & Sørensen, H. R. The minimal enzyme cocktail concept for biomass processing. *J. Cereal Sci.* **50**, 337–344 (2009).

126. Morag (Morgenstern), E., Bayer, E. A. & Lamed, R. Affinity digestion for the near-total recovery of purified cellulosome from Clostridium thermocellum. *Enzyme Microb. Technol.* **14**, 289–292 (1992).

127. Kunath, B. J., Bremges, A., Weimann, A., McHardy, A. C. & Pope, P. B. Metagenomics and CAZyme Discovery. in *Protein-Carbohydrate Interactions: Methods and Protocols* vol. 1588 255–277 (Springer, 2017).

128. Solomon, K. V *et al.* Early-branching gut fungi possess a large, comprehensive array of biomass-degrading enzymes. *Science* **351**, 1192–5 (2016).

129. Li, C. *et al.* Comparison of dilute acid and ionic liquid pretreatment of switchgrass: Biomass recalcitrance, delignification and enzymatic saccharification. *Bioresour. Technol.* **101**, 4900–4906 (2010).

130. Zybailov, B. *et al.* Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J. Proteome Res.* **5**, 2339–2347 (2006).

131. Steenbakkers, P. J. M. *et al.* The major component of the cellulosomes of anaerobic fungi from the genus piromyces is a family 48 glycoside hydrolase. *DNA Seq.* **13**, 313–320 (2002).

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

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

134. Zverlov, V. V., Kellermann, J. & Schwarz, W. H. Functional subgenomics of Clostridium thermocellum cellulosomal genes: Identification of the major catalytic components in the extracellular complex and detection of three new enzymes. *Proteomics* **5**, 3646–3653 (2005).

135. Lamed, R., Kenig, R., Setter, E. & Bayer, E. A. Major characteristics of the cellulolytic system of Clostridium thermocellum coincide with those of the purified cellulosome. *Enzyme Microb. Technol.* **7**, 37–41 (1985).

136. Cosgrove, D. J. Loosening of plant cell walls by expansins. *Nature* **407**, 321–326 (2000).

137. Artzi, L., Morag, E., Shamshoum, M. & Bayer, E. A. Cellulosomal expansin: functionality and incorporation into the complex. *Biotechnol. Biofuels* **9**, 61 (2016).

138. 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).

139. Moreira, L. R. S. & Filho, E. X. F. An overview of mannan structure and mannan-degrading enzyme systems. *Appl. Microbiol. Biotechnol.* **79**, 165–178 (2008).

140. Zverlov, V. V., Velikodvorskaya, G. A. & Schwarz, W. H. Two new cellulosome components encoded downstream of celI in the genome of Clostridium thermocellum: the non-processive endoglucanase CelN and the possibly structural protein CseP. *Microbiology* **149**, 515–524 (2003).

141. Northen, T. R. *et al.* A nanostructure-initiator mass spectrometry-based enzyme activity assay. *Proc. Natl. Acad. Sci.* **105**, 3678–3683 (2008).

142. Ju, X., Bowden, M., Engelhard, M. & Zhang, X. Investigating commercial cellulase performances toward specific biomass recalcitrance factors using reference substrates. *Appl. Microbiol. Biotechnol.* **98**, 4409–4420 (2014).

143. Hespell, R. B., O'Bryan, P. J., Moniruzzaman, M. & Bothast, R. J. Hydrolysis by commercial enzyme mixtures of AFEX-treated corn fiber and isolated xylans. *Appl. Biochem. Biotechnol.* **62**, 87–97 (1997).

144. García-Alvarez, B. *et al.* Molecular Architecture and Structural Transitions of a Clostridium thermocellum Mini-Cellulosome. *J. Mol. Biol.* **407**, 571–580 (2011).

145. Theodorou, M. K., Davies, D. R., Nielsen, B. B., Lawrence, M. I. & Trinci, A. P. J. Determination of growth of anaerobic fungi on soluble and cellulosic substrates using a pressure transducer. *Microbiology* **141**, 671–678 (1995).

146. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).

147. Deng, K. *et al.* Rapid kinetic characterization of glycosyl hydrolases based on oxime derivatization and nanostructure-initiator mass spectrometry (NIMS). *ACS Chem. Biol.* **9**, 1470–9 (2014).

148. Northen, T. R. *et al.* Clathrate nanostructures for mass spectrometry. *Nature* **449**, 1033–1036 (2007).

149. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).

150. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

151. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).

152. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods* **10**, 584–590 (2013).

153. Mastronarde, D. N. & Held, S. R. Automated tilt series alignment and tomographic reconstruction in IMOD. *J. Struct. Biol.* **197**, 102–113 (2017).

154. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

155. Fierobe, H. P. *et al.* Action of designer cellulosomes on homogeneous Versus complex substrates: Controlled incorporation of three distinct enzymes into a defined trifunctional scaffoldin. *J. Biol. Chem.* **280**, 16325–16334 (2005).

156. Vazana, Y., Moraïs, S., Barak, Y., Lamed, R. & Bayer, E. A. Chapter twenty-three - Designer Cellulosomes for Enhanced Hydrolysis of Cellulosic Substrates. in *Methods in Enzymology* (ed. Gilbert, H. J.) vol. 510 429–452 (Academic Press, 2012).

157. Dueber, J. E. *et al.* Synthetic protein scaffolds provide modular control over metabolic flux. *Nat. Biotechnol.* **27**, 5–8 (2009).

158. Wilner, O. I., Shimron, S., Weizmann, Y., Wang, Z.-G. & Willner, I. Self-Assembly of Enzymes on DNA Scaffolds: En Route to Biocatalytic Cascades and the Synthesis of Metallic Nanowires. *Nano Lett.* **9**, 2040–2043 (2009).

159. Park, S. H., Zarrinpar, A. & Lim, W. A. Rewiring MAP kinase pathways using alternative scaffold assembly mechanisms. *Science* **299**, 1061–1064 (2003).

160. Tompa, P., Batke, J., Ovadi, J., Welch, G. R. & Srere, P. A. Quantitation of the interaction between citrate synthase and malate dehydrogenase. *J. Biol. Chem.* **262**, 6089–6092 (1987).

161. An, S., Kumar, R., Sheets, E. D. & Benkovic, S. J. Reversible compartmentalization of de novo purine biosynthetic complexes in living cells. *Science* **320**, 103–106 (2008).

162. An, S., Kyoung, M., Allen, J. J., Shokat, K. M. & Benkovic, S. J. Dynamic regulation of a metabolic multi-enzyme complex by protein kinase CK2. *J. Biol. Chem.* **285**, 11093–11099 (2010).

163. Zhao, E. M. *et al.* Light-based control of metabolic flux through assembly of synthetic organelles. *Nat. Chem. Biol.* **15**, 589–597 (2019).

164. Boyken, S. E. *et al.* De novo design of tunable, pH-driven conformational changes. *Science* **364**, 658–664 (2019).

165. Gordley, R. M. *et al.* Engineering dynamical control of cell fate switching using synthetic phospho-regulons. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 13528–13533 (2016).

166. Ngo, T. A., Nakata, E., Saimura, M., Kodaki, T. & Morii, T. A protein adaptor to locate a functional protein dimer on molecular switchboard. *Methods* **67**, 142–150 (2014).

167. Price, J. V., Chen, L., Whitaker, W. B., Papoutsakis, E. & Chen, W. Scaffoldless engineered enzyme assembly for enhanced methanol utilization. *Proc. Natl. Acad. Sci.* **113**, 12691–12696 (2016).
168. Berckman, E. A. & Chen, W. Self-assembling protein nanocages for modular enzyme assembly by orthogonal bioconjugation. *Biotechnol. Prog.* **37**, (2021).
169. Padilla, J. E., Colovos, C. & Yeates, T. O. Nanohedra: Using symmetry to design self assembling protein cages, layers, crystals, and filaments. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 2217–2221 (2001).
170. Tsai, S. L., DaSilva, N. A. & Chen, W. Functional display of complex cellulosomes on the yeast surface via adaptive assembly. *ACS Synth. Biol.* **2**, 14–21 (2013).
171. Ellis, G. A. *et al.* Artificial Multienzyme Scaffolds: Pursuing in Vitro Substrate Channeling with an Overview of Current Progress. *ACS Catal.* **9**, 10812–10869 (2019).
172. Hyun, J., Lee, W. K., Nath, N., Chilkoti, A. & Zauscher, S. Capture and release of proteins on the nanoscale by stimuli-responsive elastin-like polypeptide 'switches'. *J. Am. Chem. Soc.* **126**, 7330–7335 (2004).
173. Guntas, G. *et al.* Engineering an improved light-induced dimer (iLID) for controlling the localization and activity of signaling proteins. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 112–117 (2015).
174. Mitkas, A. A., Valverde, M. & Chen, W. Dynamic modulation of enzyme activity by synthetic CRISPR–Cas6 endonucleases. *Nat. Chem. Biol.* **18**, 492–500 (2022).
175. Gilmore, S. P., Henske, J. K. & O'Malley, M. A. Driving biomass breakdown through engineered cellulosomes. *Bioengineered* **6**, 204–208 (2015).
176. Haitjema, C. H. *et al.* A parts list for fungal cellulosomes revealed by comparative genomics. *Nat. Microbiol.* **2**, 1–8 (2017).
177. Schönichen, A., Webb, B. A., Jacobson, M. P. & Barber, D. L. Considering protonation as a posttranslational modification regulating protein structure and function. *Annu. Rev. Biophys.* **42**, 289–314 (2013).
178. Muyldermans, S. *et al.* Camelid immunoglobulins and nanobody technology. *Vet. Immunol. Immunopathol.* **128**, 178–183 (2009).
179. Nagy, T. *et al.* Characterization of a Double Dockerin from the Cellulosome of the Anaerobic Fungus Piromyces equi. *J. Mol. Biol.* **373**, 612–622 (2007).
180. Kozakov, D., Brenke, R., Comeau, S. R. & Vajda, S. PIPER: An FFT-Based Protein Docking Program with Pairwise Potentials. *Proteins Struct. Funct. Bioinforma.* **65**, 392–406 (2006).
181. Brenke, R. *et al.* Application of asymmetric statistical potentials to antibody-protein docking. *Bioinformatics* **28**, 2608–2614 (2012).
182. Kozakov, D. *et al.* The ClusPro web server for protein-protein docking. *Nat. Protoc.* **12**, 255–278 (2017).
183. Cao, H., Wang, J., He, L., Qi, Y. & Zhang, J. Z. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. *J. Chem. Inf. Model.* **59**, 1508–1514 (2019).
184. Molinier, A.-L. *et al.* Synergy, structure and conformational flexibility of hybrid cellulosomes displaying various inter-cohesins linkers. *J. Mol. Biol.* **405**, 143–157 (2011).
185. Bayer, E. A., Lamed, R., White, B. A. & Flint, H. J. From cellulosomes to cellulosomics. *Chem. Rec.* **8**, (2008).

186. Shell, M. S. Coarse-graining with the relative entropy. in *Advances in Chemical Physics* vol. 161 (John Wiley & Sons, 2016).
187. Schröter, C. *et al.* A generic approach to engineer antibody pH-switches using combinatorial histidine scanning libraries and yeast display. *mAbs* **7**, 138–151 (2015).
188. Murtaugh, M. L., Fanning, S. W., Sharma, T. M., Terry, A. M. & Horn, J. R. A combinatorial histidine scanning library approach to engineer highly pH-dependent protein switches. *Protein Sci.* **20**, 1619–1631 (2011).
189. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.* **15**, 553–557 (1997).
190. Peng, X., Swift, C. L., Theodorou, M. K. & Malley, M. A. O. Chapter 5 Methods for Genomic Characterization and Maintenance of Anaerobic Fungi. **1775**, 53–67.
191. Chow, K. M. *et al.* Immunization of alpacas (Lama pacos) with protein antigens and production of antigen-specific single domain antibodies. *J. Vis. Exp.* **2019**, 1–7 (2019).
192. Chao, G. *et al.* Isolating and engineering human antibodies using yeast surface display. *Nat. Protoc.* **1**, 755–768 (2006).
193. Van Deventer, J. A. & Wittrup, K. D. *Yeast surface display for antibody isolation: Library construction, library screening, and affinity maturation*. *Methods in Molecular Biology* vol. 1131 (2014).
194. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
195. Ozkan, S. B., Wu, G. A., Chodera, J. D. & Dill, K. A. Protein folding by zipping and assembly. *Proc. Natl. Acad. Sci.* **104**, 11987–11992 (2007).
196. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, 1–17 (2017).
197. Case, D. A. *et al.* AMBER 2016. Preprint at (2016).
198. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
199. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
200. Liu, H. & Naismith, J. H. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnol.* **8**, 91 (2008).
201. Bomble, Y. J. *et al.* Modeling the self-assembly of the cellulosome enzyme complex. *J. Biol. Chem.* **286**, 5614–5623 (2011).
202. Marty, M. T. *et al.* Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles. *Anal. Chem.* **87**, 4370–4376 (2015).
203. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLOS Comput. Biol.* **5**, e1000605 (2009).
204. O'Malley, M. A., Theodorou, M. K. & Kaiser, C. A. Evaluating expression and catalytic activity of anaerobic fungal fibrolytic enzymes native topiromyces sp E2 inSaccharomyces cerevisiae. *Environ. Prog. Sustain. Energy* **31**, 37–46 (2012).
205. Jenkins, J., Lo Leggio, L., Harris, G. & Pickersgill, R. B-glucosidase, B-galactosidase, family A cellulases, family F xylanases, and two barley glycanases form a superfamily of enzymes with 8-fold B/a architecture and with two conserved glutamates

near the carboxy-terminal ends of B-strands four and seven. *FEBS Lett.* **362**, 281–285 (1995).

206. Pickersgill, R., Harris, G., Leggio, L. L., Mayans, O. & Jenkins, J. Superfamilies: the 4/7 superfamily of βα-barrel glycosidases and the right-handed parallel β-helix superfamily. *Biochem. Soc. Trans.* **26**, 190–197 (1998).

207. Drula, E. *et al.* The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.* **50**, D571–D577 (2022).

208. Tseng, C.-W. *et al.* Substrate binding of a GH5 endoglucanase from the ruminal fungus Piromyces rhizinflata. *Acta Crystallogr. Sect. F* **67**, 1189–1194 (2011).

209. Glasgow, E. M. *et al.* A structural and kinetic survey of GH5_4 endoglucanases reveals determinants of broad substrate specificity and opportunities for biomass hydrolysis. *J. Biol. Chem.* **295**, 17752–17769 (2020).

210. Lo Leggio, L. & Larsen, S. The 1.62 Å structure of Thermoascus aurantiacus endoglucanase: completing the structural picture of subfamilies in glycoside hydrolase family 5. *FEBS Lett.* **523**, 103–108 (2002).

211. Lee, T. M., Farrow, M. F., Arnold, F. H. & Mayo, S. L. A structural study of Hyprocrea jecorina Cel5A. *Protein Sci.* **20**, 1935–1940 (2011).

212. Lee, T. M., Farrow, M. F., Arnold, F. H. & Mayo, S. L. Biochemical characterization and mode of action of a thermostable endoglucanase purified from Thermoascus aurantiacus. *Arch. Biochem. Biophys.* **404**, 243–253 (2002).

213. Dominguez, R. *et al.* A common protein fold and similar active site in two distinct families of B-glycanases. **2**, (1995).

214. Davies, G. J. *et al.* Snapshots along an Enzymatic Reaction Coordinate: Analysis of a Retaining β-Glycoside Hydrolase. *Biochemistry* **37**, 11707–11713 (1998).

215. Gloster, T. M. *et al.* Characterization and Three-dimensional Structures of Two Distinct Bacterial Xyloglucanases from Families GH5 and GH12. *J. Biol. Chem.* **282**, 19177–19189 (2007).

216. dos Santos, C. R., Cordeiro, R. L., Wong, D. W. S. & Murakami, M. T. Structural Basis for Xyloglucan Specificity and α-d-Xylp(1 → 6)-d-Glcp Recognition at the −1 Subsite within the GH5 Family. *Biochemistry* **54**, 1930–1942 (2015).

217. Qin, Y., Wei, X., Song, X. & Qu, Y. Engineering endoglucanase II from Trichoderma reesei to improve the catalytic efficiency at a higher pH optimum. *J. Biotechnol.* **135**, 190–195 (2008).

218. Cheng, Y.-S. *et al.* Enhanced activity of Thermotoga maritima cellulase 12A by mutating a unique surface loop. *Appl. Microbiol. Biotechnol.* **95**, 661–669 (2012).

219. Merzlov, D. A. *et al.* Properties of enzyme preparations and homogeneous enzymes — Endoglucanases EG2 Penicillium verruculosum and LAM Myceliophthora thermophila. *Biochem. Mosc.* **80**, 473–482 (2015).

220. Liang, C. *et al.* Cloning and characterization of a thermostable and halo-tolerant endoglucanase from Thermoanaerobacter tengcongensis MB4. *Appl. Microbiol. Biotechnol.* **89**, 315–326 (2011).

221. Ren, H. N. Cloning, Expression and Characterization of Novel Fungal Endoglucanases. (Concordia University, 2009).

222. Dong, J., Hong, Y., Shao, Z. & Liu, Z. Molecular cloning, purification, and characterization of a novel, acidic, pH-stable endoglucanase from Martelella mediterranea. *J. Microbiol.* **48**, 393–398 (2010).

223. Huang, Y.-H., Huang, C.-T. & Hseu, R.-S. Effects of dockerin domains on Neocallimastix frontalis xylanases. *FEMS Microbiol. Lett.* **243**, 455–460 (2005).

224. Nagy, T. *et al.* Characterization of a Double Dockerin from the Cellulosome of the Anaerobic Fungus Piromyces equi. *J. Mol. Biol.* **373**, 612–622 (2007).

225. Cheng, Y.-S. *et al.* Structural Analysis of a Glycoside Hydrolase Family 11 Xylanase from Neocallimastix patriciarum: Insights into the molecular basis of a thermophilic enzyme. *J. Biol. Chem.* **289**, 11020–11028 (2014).

226. Nguyen, K. B. *et al.* Phosphorylation of spore coat proteins by a family of atypical protein kinases. *Proc. Natl. Acad. Sci.* **113**, E3482–E3491 (2016).

227. Kim, Y. *et al.* High-throughput protein purification and quality assessment for crystallization. *Methods* **55**, 12–28 (2011).

228. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).

229. McCoy, A. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

230. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallografica* **D67**, 235–242 (2011).

231. King, B. C., Donnelly, M. K., Bergstrom, G. C., Walker, L. P. & Gibson, D. M. An optimized microplate assay system for quantitative evaluation of plant cell wall-degrading enzyme activity of fungal culture extracts. *Biotechnol. Bioeng.* **102**, 1033–1044 (2009).

232. Morag (Morgenstern), E., Bayer, E. A. & Lamed, R. Affinity digestion for the near-total recovery of purified cellulosome from Clostridium thermocellum. *Enzyme Microb. Technol.* **14**, 289–292 (1992).

233. Katrukha, E. A. *et al.* Probing cytoskeletal modulation of passive and active intracellular dynamics using nanobody-functionalized quantum dots. *Nat. Commun.* **8**, 14772 (2017).

Appendix

Proteomics sequence coverage data from Chapter III

Table A1. Sequence coverage of database proteins identified in fungal cellulosome samples. Mean and max coverage are computed from all nine proteomics samples spanning CB, FP, and RCG cellulosomes.

Mycocosm protein ID	Protein families	Mean coverage	Max coverage
703870	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	58.9%	63.0%
216413	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	56.4%	61.9%
216560	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	58.0%	61.4%
374999	Glycoside Hydrolase Family 3 / Non-Catalytic Module Family DOC2	47.8%	57.2%
386237	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	52.7%	55.6%
386240	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	52.7%	55.6%
697093	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	45.1%	55.3%
700294	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	46.2%	52.6%
388381	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	46.5%	47.8%
704850	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	36.1%	47.4%
49410	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	37.8%	45.8%
705896	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	41.8%	44.3%
699671	Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	26.5%	43.2%
460356	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	40.9%	43.0%
699432	Glycoside Hydrolase Family 3 / Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 6 protein	41.6%	42.3%
706242	Non-Catalytic Module Family DOC2	19.8%	41.5%
410842	Glycoside Hydrolase Family 9 / Non-Catalytic Module Family DOC2	39.2%	41.5%
702410	Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	26.6%	41.1%

379692	Glycoside Hydrolase Family 124 / Non-Catalytic Module Family DOC2	14.9%	41.1%
703763	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	33.7%	41.0%
704451	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	34.0%	40.9%
382173	Glycoside Hydrolase Family 9 / Non-Catalytic Module Family DOC2	35.4%	40.6%
706678	Glycoside Hydrolase Family 124 / Non-Catalytic Module Family DOC2	15.9%	40.6%
375792	Glycoside Hydrolase Family 74 / Non-Catalytic Module Family DOC2	31.8%	40.0%
456409	Non-Catalytic Module Family DOC2	27.6%	39.7%
698650	Glycoside Hydrolase Family 43 / Carbohydrate-Binding Module Family 6 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	16.4%	38.5%
709463	Glycoside Hydrolase Family 10 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	27.9%	38.3%
388629	Glycoside Hydrolase Family 9 / Non-Catalytic Module Family DOC2	35.2%	38.2%
696953	Glycoside Hydrolase Family 9 / Non-Catalytic Module Family DOC2	30.7%	38.0%
414424	Carbohydrate Esterase Family 1 / Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 11 protein	21.0%	37.9%
462323	Glycoside Hydrolase Family 10 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	28.9%	37.2%
705894	Carbohydrate Esterase Family 6 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	18.2%	35.7%
522750	Glycoside Hydrolase Family 9 / Non-Catalytic Module Family DOC2	33.9%	35.4%
700260	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	30.4%	35.1%
706034	Glycoside Hydrolase Family 39 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	19.8%	34.5%
704136	Glycoside Hydrolase Family 43 / Carbohydrate-Binding Module Family 6 / Non-Catalytic Module Family DOC2	12.3%	34.2%
705689	Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	21.7%	34.1%
384271	Non-Catalytic Module Family DOC2 / Distantly related to plant expansins	28.0%	33.5%

703601	Glycoside Hydrolase Family 43 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	11.6%	33.5%
705447	Glycoside Hydrolase Family 74 / Non-Catalytic Module Family DOC2	31.4%	33.4%
707987	Glycoside Hydrolase Family 11 / Glycoside Hydrolase Family 10 / Non-Catalytic Module Family DOC2	16.2%	33.1%
200796	Carbohydrate-Binding Module Family 35 / Glycoside Hydrolase Family 26 / Non-Catalytic Module Family DOC2	25.3%	32.5%
385984	Glycoside Hydrolase Family 3 / Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 6 protein	22.3%	32.3%
662725	Carbohydrate-Binding Module Family 35 / Glycoside Hydrolase Family 26 / Non-Catalytic Module Family DOC2	24.7%	32.3%
439315	Carbohydrate-Binding Module Family 35 / Glycoside Hydrolase Family 26 / Non-Catalytic Module Family DOC2	25.2%	32.2%
382578	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 29 protein	29.0%	31.9%
389999	Glycoside Hydrolase Family 9 / Non-Catalytic Module Family DOC2	25.0%	31.8%
697256	Glycoside Hydrolase Family 10 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	15.9%	31.6%
700584	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 6 protein	17.2%	31.4%
704137	Glycoside Hydrolase Family 43 / Carbohydrate-Binding Module Family 6 / Non-Catalytic Module Family DOC2	11.4%	31.1%
706357	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 30 protein	17.8%	30.8%
393138	Glycoside Hydrolase Family 9 / Non-Catalytic Module Family DOC2	29.4%	30.7%
448341	Non-Catalytic Module Family DOC2 / Distantly related to plant expansins	27.1%	30.6%
671922	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 13 / Carbohydrate Esterase Family 1 protein	18.2%	30.4%
703809	Glycoside Hydrolase Family 10 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	15.6%	30.0%
699681	Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	15.9%	29.8%

668729	Glycoside Hydrolase Family 9 / Non-Catalytic Module Family DOC2	25.5%	29.8%
265432	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 6 protein	22.3%	29.5%
702792	Glycoside Hydrolase Family 43 / Carbohydrate-Binding Module Family 6 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	14.5%	29.3%
504126	Non-Catalytic Module Family DOC2	11.3%	28.8%
390877	Glycoside Hydrolase Family 9 / Non-Catalytic Module Family DOC2	25.9%	28.6%
702162	Glycoside Hydrolase Family 10 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	19.5%	28.5%
447807	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 5 protein	19.2%	28.5%
701588	Non-Catalytic Module Family DOC2	16.0%	28.4%
381575	Non-Catalytic Module Family DOC2 / Distantly related to plant expansins	24.4%	28.3%
92306	Glycoside Hydrolase Family 5 / Non-Catalytic Module Family DOC2	25.5%	28.0%
678698	Glycoside Hydrolase Family 5 / Carbohydrate-Binding Module Family 10 / Non-Catalytic Module Family DOC2	21.0%	27.9%
519770	Glycoside Hydrolase Family 5 / Carbohydrate-Binding Module Family 10 / Non-Catalytic Module Family DOC2	20.6%	27.8%
702459	Non-Catalytic Module Family DOC2 / Distantly related to plant expansins	19.1%	27.6%
704578	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 15 protein	24.2%	27.6%
150710	Carbohydrate Esterase Family 6 / Non-Catalytic Module Family DOC2	11.2%	27.6%
707399	Non-Catalytic Module Family DOC2 / Distantly related to plant expansins	23.9%	27.4%
407760	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 6 protein	26.0%	26.9%
518420	Glycoside Hydrolase Family 2 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	9.2%	26.7%
703294	Glycoside Hydrolase Family 5 / Non-Catalytic Module Family DOC2	15.7%	26.4%
463299	Glycoside Hydrolase Family 11 / Non-Catalytic Module Family DOC2	20.3%	26.4%
447808	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 5 protein	18.6%	26.3%
409744	Carbohydrate Esterase Family 6 / Non-Catalytic Module Family DOC2	10.8%	26.2%

701753	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 5 protein	19.9%	26.2%
385462	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 13 / Glycoside Hydrolase Family 95 protein	10.1%	26.1%
677808	Glycoside Hydrolase Family 2 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	9.0%	26.1%
503204	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 13 / Carbohydrate Esterase Family 1 protein	9.8%	25.6%
391852	Glycoside Hydrolase Family 5 / Carbohydrate-Binding Module Family 1 / Non-Catalytic Module Family DOC2	14.6%	25.5%
699126	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 6 protein	24.3%	25.3%
699455	Glycoside Hydrolase Family 74 / Non-Catalytic Module Family DOC2	16.6%	24.9%
458000	Distantly related to plant expansins / Carbohydrate-Binding Module Family 63 / Non-Catalytic Module Family DOC2	17.5%	24.8%
677809	Glycoside Hydrolase Family 2 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	8.5%	24.8%
706039	Glycoside Hydrolase Family 39 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	16.6%	24.7%
392363	Non-Catalytic Module Family DOC2	14.1%	24.6%
661836	Carbohydrate Esterase Family 1 / Non-Catalytic Module Family DOC2	11.6%	24.5%
462502	Glycoside Hydrolase Family 9 / Non-Catalytic Module Family DOC2	11.3%	24.2%
677813	Putative scaffoldin	8.1%	24.1%
503798	Glycoside Hydrolase Family 45 / Non-Catalytic Module Family DOC2	19.8%	24.0%
390444	Glycoside Hydrolase Family 5 / Non-Catalytic Module Family DOC2	19.9%	24.0%
453694	Glycoside Hydrolase Family 5 / Non-Catalytic Module Family DOC2	19.8%	23.7%
517403	Distantly related to plant expansins / Carbohydrate-Binding Module Family 63 / Non-Catalytic Module Family DOC2	19.9%	23.5%
698941	Glycoside Hydrolase Family 5 / Non-Catalytic Module Family DOC2	19.2%	23.5%
707328	Glycoside Hydrolase Family 5 / Non-Catalytic Module Family DOC2	19.2%	23.5%

388331	Carbohydrate-Binding Module Family 35 / Glycoside Hydrolase Family 26 / Non-Catalytic Module Family DOC2	19.3%	23.4%
642334	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 6 protein	21.0%	23.2%
701901	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 45 protein	15.0%	22.7%
451288	Non-Catalytic Module Family DOC2	9.0%	22.7%
233793	Putative scaffoldin	17.5%	22.6%
676442	Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	12.2%	22.2%
514294	Glycoside Hydrolase Family 5 / Non-Catalytic Module Family DOC2	19.5%	22.0%
425970	Carbohydrate-Binding Module Family 35 / Glycoside Hydrolase Family 26 / Non-Catalytic Module Family DOC2	16.6%	21.8%
393333	Carbohydrate-Binding Module Family 10 / Glycoside Hydrolase Family 5 / Non-Catalytic Module Family DOC2	14.8%	21.7%
675650	Glycoside Hydrolase Family 8 / Non-Catalytic Module Family DOC2	7.2%	21.6%
394704	Non-Catalytic Module Family DOC2	7.2%	21.6%
703027	Glycoside Hydrolase Family 39 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	8.2%	21.5%
392645	Carbohydrate-Binding Module Family 35 / Glycoside Hydrolase Family 26 / Non-Catalytic Module Family DOC2	16.3%	21.4%
703642	Carbohydrate Esterase Family 6 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	15.9%	21.4%
100603	Glycoside Hydrolase Family 45 / Non-Catalytic Module Family DOC2	16.1%	21.3%
700158	Non-Catalytic Module Family DOC2	7.7%	21.3%
677706	Non-Catalytic Module Family DOC2	7.2%	21.0%
662688	Non-Catalytic Module Family DOC2 / Distantly related to plant expansins	16.1%	20.7%
383322	Non-Catalytic Module Family DOC2	14.1%	20.7%
706661	Glycoside Hydrolase Family 10 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	14.0%	20.4%
460043	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 1 / Glycoside Hydrolase Family 45 protein	14.2%	20.3%
92305	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 5 protein	18.9%	20.2%

465072	Non-Catalytic Module Family DOC2 / Distantly related to plant expansins	17.9%	20.1%
388844	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 13 / Carbohydrate Esterase Family 1 protein	12.8%	20.1%
387693	Non-Catalytic Module Family DOC2 / Distantly related to plant expansins	15.0%	20.1%
700562	Glycoside Hydrolase Family 5 / Non-Catalytic Module Family DOC2	12.4%	20.0%
673436	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 45 protein	14.3%	19.9%
74392	Carbohydrate-Binding Module Family 52 / Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 16 protein	13.3%	19.9%
456027	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 13 / Carbohydrate Esterase Family 1 protein	7.5%	19.8%
386646	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 5 protein	14.4%	19.8%
704225	Non-Catalytic Module Family DOC2	15.3%	19.6%
398700	Non-Catalytic Module Family DOC2	11.8%	19.5%
136312	Glycoside Hydrolase Family 5 / Non-Catalytic Module Family DOC2	18.8%	19.4%
709512	Carbohydrate Esterase Family 6 / Non-Catalytic Module Family DOC2	8.8%	19.3%
668330	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 6 protein	17.6%	19.2%
702396	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 45 protein	15.2%	18.9%
508324	Glycoside Hydrolase Family 43 / Carbohydrate-Binding Module Family 6 / Non-Catalytic Module Family DOC2	8.0%	18.8%
705046	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 30 protein	10.5%	18.8%
101653	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	13.1%	18.6%
709143	No annotation	11.5%	18.5%
700014	Glycoside Hydrolase Family 11 / Glycoside Hydrolase Family 10 / Non-Catalytic Module Family DOC2	14.4%	18.5%
392147	Glycoside Hydrolase Family 5 / Non-Catalytic Module Family DOC2	8.9%	18.3%
706376	Glycoside Hydrolase Family 39 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	14.9%	18.3%
392576	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 43 protein	7.6%	18.3%

673330	Putative scaffoldin	12.3%	18.2%
677811	Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	6.0%	18.1%
699138	Putative scaffoldin	6.0%	18.1%
702500	Glycoside Hydrolase Family 124 / Non-Catalytic Module Family DOC2	6.3%	17.9%
700182	Non-Catalytic Module Family DOC2	5.8%	17.5%
669750	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 95 protein	5.8%	17.5%
375710	Carbohydrate Esterase Family 6 / Non-Catalytic Module Family DOC2	11.9%	17.5%
708206	Glycoside Hydrolase Family 2 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	6.0%	17.3%
381145	Glycoside Hydrolase Family 10 / Carbohydrate-Binding Module Family 29 / Non-Catalytic Module Family DOC2	12.2%	17.2%
511708	Putative scaffoldin	11.8%	17.2%
705678	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 9 protein	5.9%	16.9%
705908	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	6.5%	16.7%
708349	Glycoside Hydrolase Family 5 / Non-Catalytic Module Family DOC2	6.3%	16.6%
699925	Carbohydrate Esterase Family 3 / Non-Catalytic Module Family DOC2	5.5%	16.4%
71876	Putative scaffoldin	5.5%	16.4%
425520	Glycoside Hydrolase Family 11 / Non-Catalytic Module Family DOC2	11.0%	16.4%
391416	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 1 / Glycoside Hydrolase Family 45 protein	10.8%	15.7%
706219	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	6.7%	15.3%
703965	Glycoside Hydrolase Family 11 / Non-Catalytic Module Family DOC2	11.6%	15.0%
705926	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 10 protein	14.0%	14.7%
518456	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 1 / Glycoside Hydrolase Family 45 protein	11.5%	14.7%
709739	Glycoside Hydrolase Family 2 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	4.9%	14.5%
697808	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	10.7%	14.4%

446039	Glycoside Hydrolase Family 5 / Carbohydrate-Binding Module Family 1 / Non-Catalytic Module Family DOC2	10.7%	14.2%
697835	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 6 protein	8.6%	14.0%
698413	Glycoside Hydrolase Family 43 / Carbohydrate-Binding Module Family 6 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	6.4%	13.9%
666037	Glycoside Hydrolase Family 45 / Non-Catalytic Module Family DOC2	10.7%	13.6%
511696	Putative scaffoldin	8.9%	13.6%
390947	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 5 protein	7.5%	13.6%
505309	Putative scaffoldin	8.2%	13.5%
702940	Non-Catalytic Module Family DOC2	7.6%	13.4%
702029	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	9.9%	13.1%
375638	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 26 protein	9.1%	12.6%
435585	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 1 protein	9.5%	12.5%
456134	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 45 protein	11.5%	12.5%
386889	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 45 protein	8.8%	12.4%
704810	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 45 protein	8.7%	12.4%
403028	Glycoside Hydrolase Family 74 / Non-Catalytic Module Family DOC2	9.4%	12.3%
674772	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 5 protein	7.8%	12.1%
703877	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	5.5%	11.7%
704453	Glycoside Hydrolase Family 53 / Non-Catalytic Module Family DOC2	5.8%	11.7%
701791	Glycoside Hydrolase Family 5 / Non-Catalytic Module Family DOC2	7.1%	11.7%
707026	Non-Catalytic Module Family DOC2	6.4%	11.0%
697018	Distantly related to plant expansins / Carbohydrate-Binding Module Family 63 / Non-Catalytic Module Family DOC2	7.1%	11.0%
386676	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 16 protein	10.0%	11.0%
416680	Glycoside Hydrolase Family 5 / Non-Catalytic Module Family DOC2	5.3%	11.0%

393040	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	4.7%	11.0%
673939	Putative scaffoldin	3.7%	10.7%
368822	Non-Catalytic Module Family DOC2	8.4%	10.6%
700737	Non-Catalytic Module Family DOC2	4.8%	10.5%
462277	Non-Catalytic Module Family DOC2	6.6%	10.5%
705285	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 45 protein	7.1%	10.4%
706190	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	4.2%	10.4%
385667	Non-Catalytic Module Family DOC2	5.4%	10.2%
149630	Putative scaffoldin	6.1%	10.1%
706220	Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	9.0%	10.0%
667808	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 1 protein	3.3%	9.9%
709049	Putative scaffoldin	3.3%	9.9%
389496	Non-Catalytic Module Family DOC2	6.8%	9.9%
462738	Non-Catalytic Module Family DOC2	7.1%	9.8%
405488	Non-Catalytic Module Family DOC2	8.2%	9.7%
450949	Non-Catalytic Module Family DOC2	5.1%	9.6%
703071	Glycoside Hydrolase Family 43 / Carbohydrate-Binding Module Family 22 / Non-Catalytic Module Family DOC2	6.6%	9.3%
392622	Non-Catalytic Module Family DOC2	4.6%	9.1%
673626	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 16 protein	6.8%	9.0%
389665	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 9 protein	3.3%	8.7%
383447	Non-Catalytic Module Family DOC2	6.2%	8.6%
697530	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 26 protein	4.8%	8.6%
638514	Glycoside Hydrolase Family 39 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	4.7%	8.2%
699855	Non-Catalytic Module Family DOC2	3.3%	8.2%
377479	Non-Catalytic Module Family DOC2	3.2%	8.1%
672438	Putative scaffoldin	4.6%	8.1%
627791	Glycoside Hydrolase Family 39 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	4.7%	8.0%
428788	Non-Catalytic Module Family DOC2	3.8%	7.9%
385759	Non-Catalytic Module Family DOC2	3.2%	7.7%
389498	Non-Catalytic Module Family DOC2	4.5%	7.7%

432482	Non-Catalytic Module Family DOC2	4.6%	7.6%
366334	Non-Catalytic Module Family DOC2	2.5%	7.5%
677176	Putative scaffoldin	2.6%	7.4%
706493	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	4.8%	7.3%
698246	Carbohydrate Esterase Family 6 / Carbohydrate-Binding Module Family 13 / Non-Catalytic Module Family DOC2	3.5%	6.9%
672034	Non-Catalytic Module Family DOC2	2.2%	6.6%
673388	Putative scaffoldin	2.2%	6.6%
708408	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 10 protein	3.9%	6.5%
644786	Carbohydrate Esterase Family 6 / Non-Catalytic Module Family DOC2	4.1%	6.4%
646397	Non-Catalytic Module Family DOC2	3.5%	6.3%
704715	Non-Catalytic Module Family DOC2	3.5%	6.3%
141915	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 10 protein	2.1%	6.2%
664792	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 26 protein	3.2%	6.2%
21290	Non-Catalytic Module Family DOC2	3.3%	5.9%
369784	Non-Catalytic Module Family DOC2	2.0%	5.9%
673656	Non-Catalytic Module Family DOC2	2.3%	5.8%
149016	No annotation	2.3%	5.7%
703212	Non-Catalytic Module Family DOC2	2.0%	5.4%
230365	Glycoside Hydrolase Family 9 / Carbohydrate-Binding Module Family 13 / Glycoside Hydrolase Family 43 / Non-Catalytic Module Family DOC2	3.1%	5.4%
708536	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 29 / Carbohydrate-Binding Module Family 1 protein	4.5%	5.4%
386877	Non-Catalytic Module Family DOC2	3.1%	5.3%
674871	Non-Catalytic Module Family DOC2	1.8%	5.3%
456326	Polysaccharide Lyase Family 4 / Non-Catalytic Module Family DOC2	1.7%	5.1%
140696	Putative scaffoldin	3.5%	5.0%
678574	Putative scaffoldin	2.8%	5.0%
463642	Glycoside Hydrolase Family 53 / Non-Catalytic Module Family DOC2	3.1%	4.9%
669604	Non-Catalytic Module Family DOC2	2.2%	4.9%
669624	Non-Catalytic Module Family DOC2	3.1%	4.8%
456468	No annotation	2.9%	4.6%
151284	Putative scaffoldin	3.1%	4.6%
699417	Non-Catalytic Module Family DOC2	1.5%	4.4%
512314	Putative scaffoldin	1.5%	4.4%

700366	Non-Catalytic Module Family DOC2	2.0%	4.3%
452532	Non-Catalytic Module Family DOC2	2.7%	4.3%
505259	Putative scaffoldin	2.5%	4.2%
431319	Non-Catalytic Module Family DOC2	1.6%	4.2%
702582	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 1 protein	2.4%	4.1%
708729	Non-Catalytic Module Family DOC2	1.4%	3.8%
458433	Carbohydrate Esterase Family 16 / Non-Catalytic Module Family DOC2	1.2%	3.7%
663410	Non-Catalytic Module Family DOC2	1.2%	3.7%
389504	Non-Catalytic Module Family DOC2	1.8%	3.7%
106854	Non-Catalytic Module Family DOC2	1.2%	3.6%
460913	Non-Catalytic Module Family DOC2	1.6%	3.6%
360664	Non-Catalytic Module Family DOC2	3.5%	3.5%
365664	Non-Catalytic Module Family DOC2	1.2%	3.5%
384989	Non-Catalytic Module Family DOC2	1.4%	3.4%
709446	Non-Catalytic Module Family DOC2	1.6%	3.4%
454775	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 10 protein	1.1%	3.3%
452530	Non-Catalytic Module Family DOC2	1.1%	3.3%
706453	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	1.1%	3.3%
82127	Non-Catalytic Module Family DOC2	1.1%	3.2%
436430	Non-Catalytic Module Family DOC2	1.1%	3.2%
383375	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 1 protein	1.3%	3.1%
382908	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	1.0%	3.1%
374168	Non-Catalytic Module Family DOC2	1.0%	3.1%
700300	Putative scaffoldin	1.0%	3.1%
705262	Non-Catalytic Module Family DOC2	1.5%	3.0%
675862	Non-Catalytic Module Family DOC2	1.6%	2.9%
460692	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 10 protein	1.6%	2.9%
699950	Non-Catalytic Module Family DOC2	1.3%	2.9%
370499	Non-Catalytic Module Family DOC2	1.0%	2.9%
664085	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 10 protein	2.2%	2.8%
371646	No annotation	0.9%	2.8%
701280	Glycoside Hydrolase Family 5 / Carbohydrate-Binding Module Family 10 protein	1.3%	2.7%
513728	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	0.9%	2.7%
517270	Putative scaffoldin	0.9%	2.7%

375096	Glycoside Hydrolase Family 5 / Carbohydrate-Binding Module Family 10 protein	1.7%	2.7%
676437	Non-Catalytic Module Family DOC2	0.9%	2.6%
709532	Non-Catalytic Module Family DOC2	0.9%	2.6%
462091	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 43 protein	2.4%	2.5%
673563	Putative scaffoldin	0.8%	2.5%
390323	Non-Catalytic Module Family DOC2	1.6%	2.5%
704226	Non-Catalytic Module Family DOC2	1.1%	2.4%
670098	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	0.8%	2.4%
673152	Putative scaffoldin	0.8%	2.4%
625725	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 45 protein	0.8%	2.3%
705498	Non-Catalytic Module Family DOC2	0.8%	2.3%
433303	Non-Catalytic Module Family DOC2	0.8%	2.3%
463830	Non-Catalytic Module Family DOC2	0.7%	2.2%
449056	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 10 protein	1.7%	2.2%
203180	Putative scaffoldin	0.8%	2.1%
705733	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 10 protein	1.8%	2.1%
664791	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 26 protein	1.3%	2.0%
450165	No annotation	0.9%	2.0%
505319	Non-Catalytic Module Family DOC2	1.1%	1.9%
625708	Non-Catalytic Module Family DOC2	1.4%	1.9%
698631	Non-Catalytic Module Family DOC2	0.6%	1.8%
699528	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 43 protein	0.6%	1.8%
671216	Non-Catalytic Module Family DOC2	1.6%	1.8%
673021	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 45 protein	0.6%	1.7%
458251	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 30 protein	0.8%	1.6%
461921	Glycoside Hydrolase Family 5 / Non-Catalytic Module Family DOC2	0.7%	1.5%
150427	Non-Catalytic Module Family DOC2	0.8%	1.5%
505077	Non-Catalytic Module Family DOC2	0.9%	1.4%
702584	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 1 protein	0.7%	1.4%
667490	Putative scaffoldin	0.8%	1.4%
464557	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 1 protein	0.7%	1.3%
455299	Non-Catalytic Module Family DOC2	0.4%	1.2%

661461	Non-Catalytic Module Family DOC2	0.9%	1.2%
668892	Putative scaffoldin	0.5%	1.2%
433293	Non-Catalytic Module Family DOC2	0.4%	1.1%
452527	Non-Catalytic Module Family DOC2	0.4%	1.1%
668261	Putative scaffoldin	0.6%	1.1%
703386	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	0.7%	1.1%
702205	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 45 protein	0.4%	1.1%
666197	Non-Catalytic Module Family DOC2	0.3%	1.0%
700145	No annotation	0.3%	1.0%
438200	Non-Catalytic Module Family DOC2	0.3%	1.0%
518968	Putative scaffoldin	0.6%	1.0%
456457	Non-Catalytic Module Family DOC2	0.5%	0.9%
448538	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 1 protein	0.6%	0.9%
385801	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 1 protein	0.7%	0.9%
667474	Putative scaffoldin	0.3%	0.9%
700286	Putative scaffoldin	0.3%	0.9%
391851	Non-Catalytic Module Family DOC2 / Carbohydrate-Binding Module Family 1 protein	0.4%	0.9%
511980	Putative scaffoldin	0.5%	0.9%
707763	Non-Catalytic Module Family DOC2	0.3%	0.8%
697225	Carbohydrate Esterase Family 15 / Non-Catalytic Module Family DOC2	0.5%	0.8%
513668	Putative scaffoldin	0.7%	0.8%
509540	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 26 protein	0.4%	0.8%
706654	Carbohydrate Esterase Family 15 / Non-Catalytic Module Family DOC2	0.5%	0.8%
250285	Putative scaffoldin	0.4%	0.8%
390705	Glycoside Hydrolase Family 48 / Non-Catalytic Module Family DOC2	0.5%	0.8%
705268	Carbohydrate Esterase Family 15 / Non-Catalytic Module Family DOC2	0.3%	0.8%
676642	Non-Catalytic Module Family DOC2	0.4%	0.8%
699677	Non-Catalytic Module Family DOC2 / Carbohydrate Esterase Family 1 protein	0.3%	0.8%
383415	Glycoside Hydrolase Family 114 / Non-Catalytic Module Family DOC2	0.3%	0.8%
678545	Non-Catalytic Module Family DOC2	0.4%	0.8%
677307	Non-Catalytic Module Family DOC2	0.4%	0.7%
456489	Glycoside Hydrolase Family 114 / Non-Catalytic Module Family DOC2	0.2%	0.7%

516589	Non-Catalytic Module Family DOC2	0.2%	0.7%
465112	Carbohydrate-Binding Module Family 18 / Non-Catalytic Module Family DOC2	0.2%	0.6%
677091	Putative scaffoldin	0.3%	0.6%
678784	Non-Catalytic Module Family DOC2	0.3%	0.6%
663240	Non-Catalytic Module Family DOC2	0.4%	0.6%
516594	Non-Catalytic Module Family DOC2	0.4%	0.6%
512813	Putative scaffoldin	0.3%	0.6%
703768	Non-Catalytic Module Family DOC2 / Glycoside Hydrolase Family 97 protein	0.2%	0.6%
449973	Non-Catalytic Module Family DOC2	0.5%	0.6%
250107	Non-Catalytic Module Family DOC2	0.2%	0.5%
676636	Non-Catalytic Module Family DOC2	0.3%	0.5%
677665	Putative scaffoldin	0.3%	0.4%
516642	Putative scaffoldin	0.3%	0.4%
151004	Putative scaffoldin	0.2%	0.4%
512289	Putative scaffoldin	0.3%	0.4%
507240	Putative scaffoldin	0.2%	0.4%
679711	Putative scaffoldin	0.2%	0.4%
371464	Non-Catalytic Module Family DOC2	0.2%	0.4%
219071	Non-Catalytic Module Family DOC2	0.1%	0.4%
668006	Putative scaffoldin	0.3%	0.3%
705106	Putative scaffoldin	0.2%	0.3%
676679	Putative scaffoldin	0.3%	0.3%
676635	Non-Catalytic Module Family DOC2	0.2%	0.3%
700555	Putative scaffoldin	0.3%	0.3%
511110	Putative scaffoldin	0.2%	0.3%
645975	Putative scaffoldin	0.1%	0.3%
632984	Putative scaffoldin	0.1%	0.3%
644130	Putative scaffoldin	0.2%	0.3%
511722	Putative scaffoldin	0.2%	0.2%
509185	Putative scaffoldin	0.1%	0.2%
365833	Non-Catalytic Module Family DOC2	0.1%	0.2%

Table A2. Mean NSAF-computed mole fractions of enzyme families within cellulosome preparations from *N. californiae* grown on cellobiose, filter paper, and reed canary grass. Note Hemicellulases include all enzyme families listed below “No functional annotation.” Data are reported as the mean ± standard deviation of N=3 biological replicates.

Family	**Cellobiose**	**Filter paper**	**Reed canary grass**
GH48	0.478 ± 0.038	0.392 ± 0.040	0.374 ± 0.027
Expansin	0.051 ± 0.005	0.115 ± 0.025	0.083 ± 0.009

GH45	0.036 ± 0.007	0.069 ± 0.013	0.064 ± 0.016
GH6	0.038 ± 0.017	0.052 ± 0.028	0.026 ± 0.005
GH9	0.061 ± 0.003	0.061 ± 0.008	0.056 ± 0.005
GH5	0.091 ± 0.022	0.090 ± 0.033	0.064 ± 0.007
GH2/GH3	0.013 ± 0.002	0.012 ± 0.006	0.015 ± 0.003
GH3+GH6	0.019 ± 0.001	0.022 ± 0.011	0.014 ± 0.005
GH26	0.088 ± 0.017	0.028 ± 0.006	0.014 ± 0.003
Hemicellulases	0.046 ± 0.004	0.058 ± 0.007	0.091 ± 0.006
CE1/CE6/CE15	0.028 ± 0.004	0.020 ± 0.004	0.060 ± 0.005
CotH kinase/phosphatase	0.015 ± 0.000	0.017 ± 0.004	0.022 ± 0.004
scaffoldin	0.007 ± 0.001	0.012 ± 0.008	0.009 ± 0.003
No functional annotation	0.037 ± 0.005	0.049 ± 0.007	0.069 ± 0.008
GH10/GH11	0.025 ± 0.003	0.035 ± 0.006	0.037 ± 0.004
GH43	0.003 ± 0.001	0.004 ± 0.001	0.024 ± 0.003
GH16	0.007 ± 0.002	0.005 ± 0.002	0.003 ± 0.002
GH30	0.003 ± 0.001	0.001 ± 0.001	0.005 ± 0.001
GH39	0.005 ± 0.001	0.005 ± 0.003	0.011 ± 0.002
GH53	0.001 ± 0.001	0.000 ± 0.000	0.001 ± 0.000
GH74	0.010 ± 0.002	0.009 ± 0.002	0.011 ± 0.002

Native mass spectrometry spectra for nanobody-dockerin work

Native mass spectrometry (MS) is a useful technique for detecting the presence of protein complexes in protein mixtures and is especially useful for measuring protein complex stoichiometry.

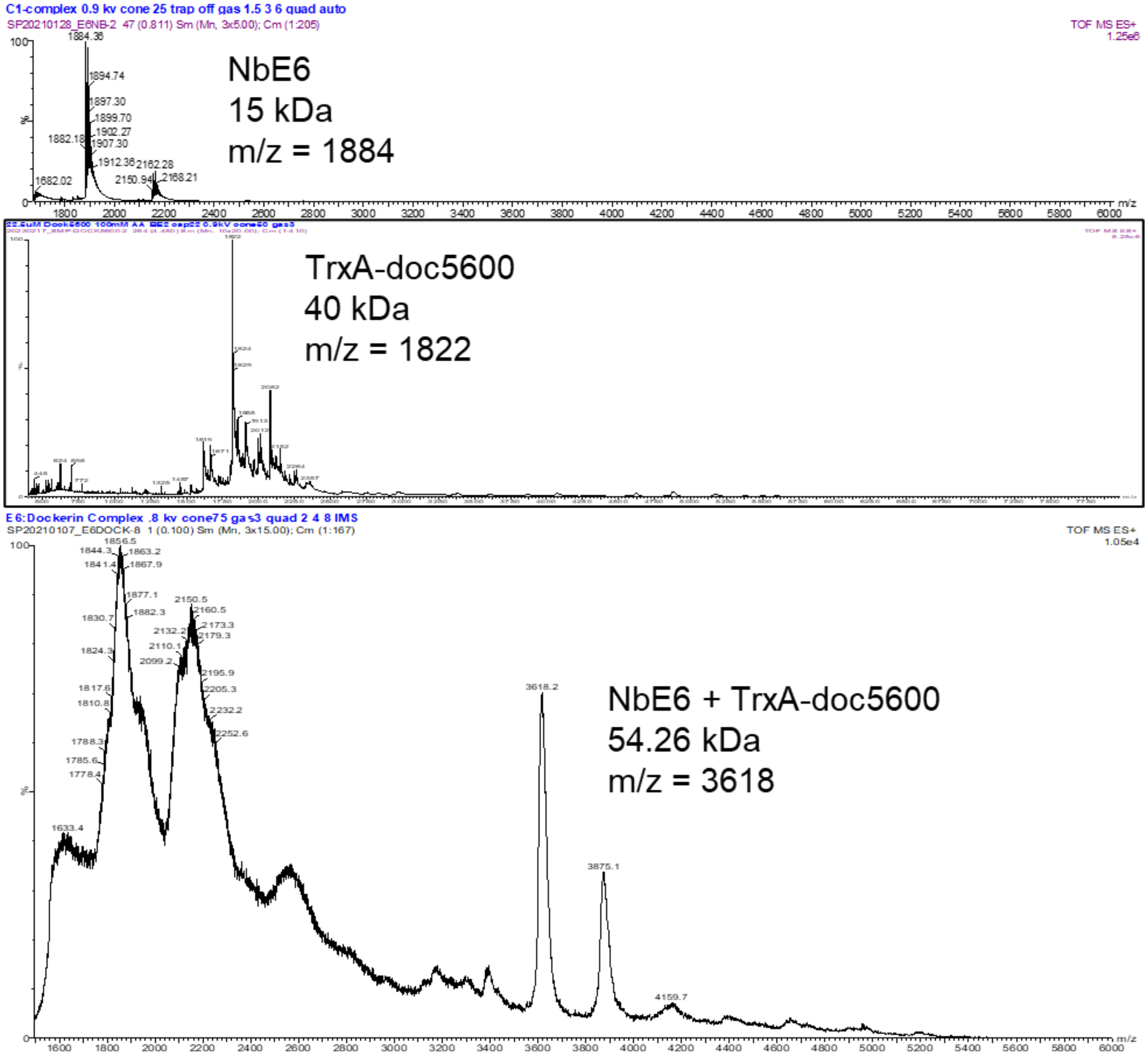


Figure A1. Native MS spectra show 1:1 complex containing NbE6 and doc5600. (Top) Spectrum of NbE6 only. (Middle) Spectrum of TrxA-doc5600 only. (Bottom) Spectrum of a mixture of NbE6 and TrxA-doc5600.

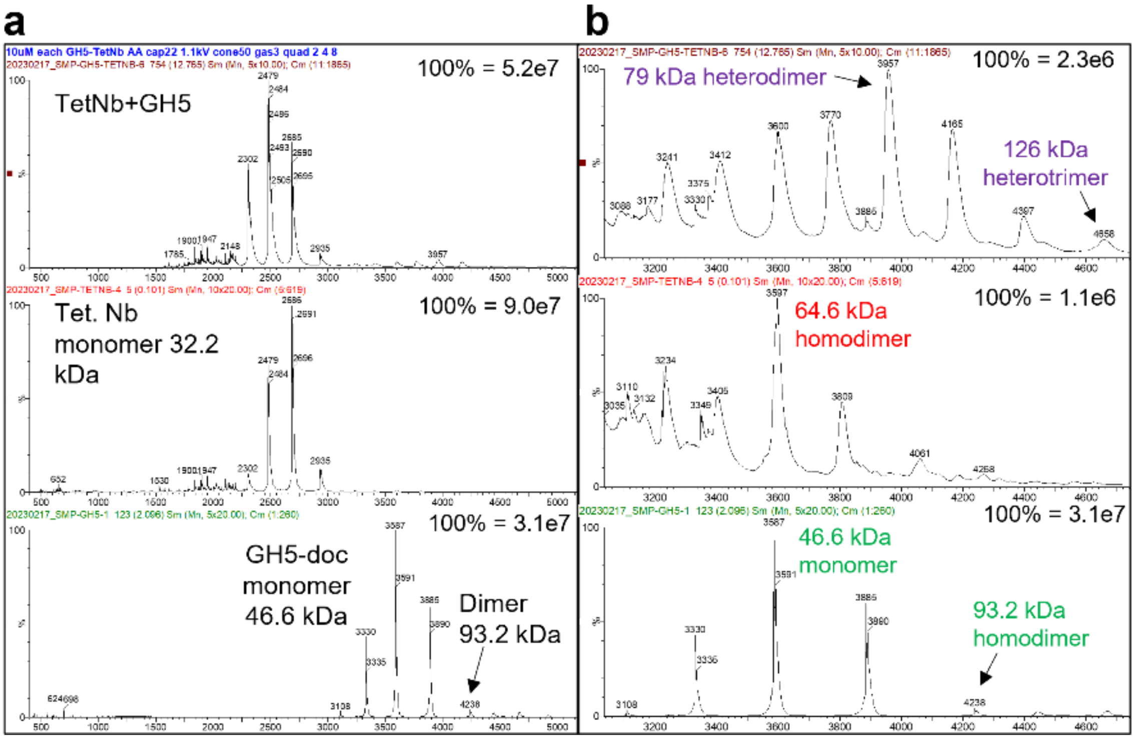


Figure A2. Native MS detects heterodimeric and heterotrimeric species in mixtures of Tethered NbE6 scaffoldin (TetNb) and a double dockerin-fused GH5 enzyme (GH5-doc). A) Full spectrum of the TetNb+GH5 mixture (top), TetNb only (middle), and GH5-doc only (bottom). B) Zoomed in view of the higher m/z range with annotated homo- and heterooligomer species of the full spectrum to the left. Top to bottom order is the same as in A).

Predicting dockerin-scaffoldin binding with AlphaFold

The identity of the protein that fungal dockerin binds to mediate assembly of the fungal cellulosome has been exceptionally difficult to uncover with great certainty, and this has made it impossible to reconstitute fungal cellulosomes for biochemical investigation and engineering. Projecting the structure and assembly mechanisms of bacterial cellulosomes onto those of anaerobic fungi, we expect that enzyme-fused dockerin domains non-covalently interact with partner cohesin domains that are repeated on a central scaffodin protein that brings many dockerin-containing proteins into a single complex. While several groups have identified putative cohesins that, by various affinity capture techniques, bind fungal dockerin, there is no conclusive evidence of a dockerin-protein interaction of ~100 nM affinity, as has been measured for dockerin against crude cellulosome extract [59]. With newly published whole genome sequences for several anaerobic fungi, Haitjema et al identified a class of large, repeat-rich proteins that fit the expected description of a fungal scaffoldin [82]. Furthermore, a fragment of one of these scaffoldins showed dockerin binding with a $K_d \cong 1\ \mu M$. The 1-2 orders of magnitude lower binding affinity lends some skepticism to whether this is the true dockerin binding partner that mediates cellulosome assembly. However, this may be due to missing post-translational modifications or misfolding of the dockerin and putative scaffoldin fragment when the proteins are produced in *E. coli*. Both proteins contain several disulfide bonds, and it's reasonable to expect that a sizable fraction of the proteins were misfolded, which could reduce the measured K_d by perhaps a factor of 2-4 (if 50-75% of the proteins are misfolded – single dockerin has 2 disulfides, double dockerin has 4, and the putative scaffoldin fragment has 2). Missing post-translational modifications that contribute to the binding enthalpy could easily contribute to 10x-lower binding affinity as the K_d is proportional to $e^{\frac{-\Delta H_{bind}+T\Delta S_{bind}}{RT}}$ (e.g. a 1.3 kcal/mol increase in ΔG_{bind}, a 16% change, yields a 10x increase in K_d).

In addition to the dockerin-binding nanobodies presented in this thesis, this scaffoldin fragment presents another tool for synthetic cellulosome development that may be better suited than those nanobodies for building synthetic cellulosomes. Improving the binding affinity of scaffoldin fragment for dockerin would improve its utility as a part for cellulosome construction. Towards this aim, I employed AlphaFold-based protein structure prediction and docking tools to predict the structure of this scaffoldin fragment and also predict the structure of the dockerin-scaffoldin fragment complex.

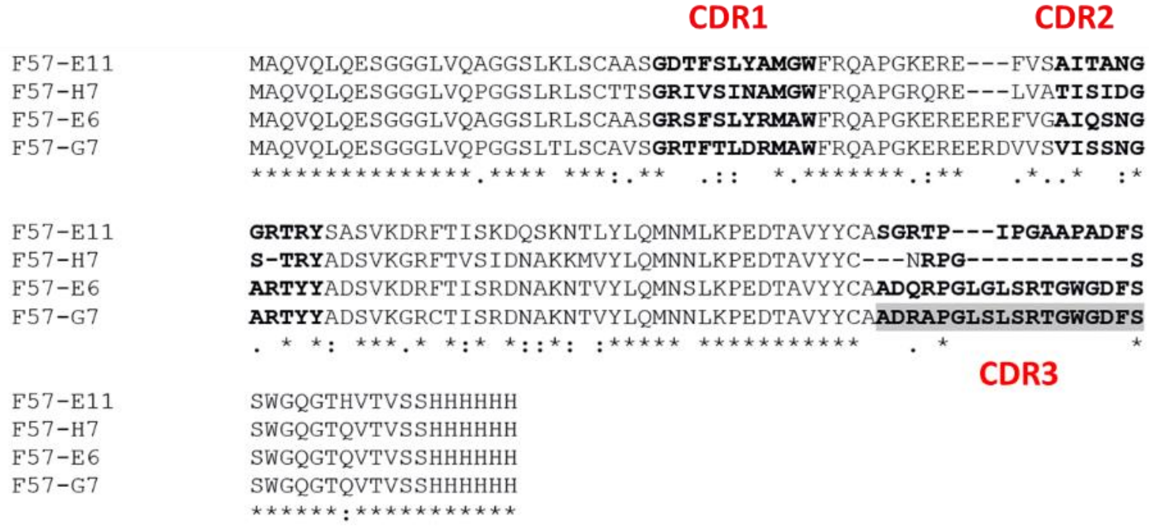
AlphaFold predicts an anti-parallel β-sheet structure with two disulfide bonds for the scaffoldin fragment. All conserved dockerin residues, which when mutated disrupted binding in an earlier study [59], are at the protein interaction interface in the AlphaFold model (Figure A3). Furthermore, conserved positions in the repeat motif present in the putative fungal scaffoldins, namely Y35, E59, K50, and N52, appear to form contacts with the important dockerin residues. The model's placement of sequence-conserved residues and those biochemically verified to be important to binding makes further investigation into this protein-protein interaction enticing. The structure provides a starting point for rationally mutating the scaffoldin fragment to bind dockerin more tightly, making it a better part for cellulosome construction. Analogously, the structure also suggests mutations to make towards verifying the fidelity of this model. For example, measuring binding affinity changes upon mutating away the predicted K50-D24 salt bridge would support or undermine the presence of that interaction at the protein-protein interface.

Figure A3. AlphaFold Multimer predicts a dockerin-scaffoldin fragment structure consistent with sequence and biochemical analyses. A representative AlphaFold Multimer prediction of single dockerin (magenta) binding the scaffoldin fragment from Haitjema et al, 2017 (green). Key residues on each protein are annotated in white, and potential inter-protein interactions are shown with yellow dashed lines.

Other nanobodies isolated against Neocallimastix californiae cellulosome

A major research objective of the Department of Energy Bioimaging project was to employ advanced, quantum-enabled imaging techniques to study the processes of fungal cellulosome production, trafficking, assembly, and localization in live cell systems. The cornerstone technology of this approach was the cellulosome targeting nanobody, which could be conjugated with quantum dots for live cell imaging applications. As described in Chapter IV, I collaborated with the Nanobody production core at the University of Kentucky (UK) School of Molecular Medicine to produce cellulosome-binding nanobodies. To do this, I purified milligram quantities of cellulosome complexes from *Neocallimastix californiae* cultures and sent them to UK, where they were used in an alpaca immunization campaign and subsequent panning to identify nanobodies produced by the alpaca that specifically bound the *N. californiae* cellulosome.

From this work, we received three nanobodies, NbE6, which I worked with extensively as part of the protein complex engineering work, NbE11, and NbH7. A multiple sequence alignment of the three sequences is shown in Figure A4. E6 and E11 are highly homologous



```
F57-E11       MAQVQLQESGGGLVQAGGSLKLSCAASGDTFSLYAMGWFRQAPGKERE---FVSAITANG
F57-H7        MAQVQLQESGGGLVQPGGSLRLSCTTSGRIVSINAMGWFRQAPGRQRE---LVATISIDG
F57-E6        MAQVQLQESGGGLVQAGGSLRLSCAASGRSFSLYRMAWFRQAPGKEREEREFVGAIQSNG
F57-G7        MAQVQLQESGGGLVQPGGSLTLSCAVSGRTFTLDRMAWFRQAPGKEREERDVVSVISSNG
              ***************.**** ***:.**  .::  *.*******.:**   .*..*  :*

F57-E11       GRTRYSASVKDRFTISKDQSKNTLYLQMNMLKPEDTAVYYCASGRTP---IPGAAPADFS
F57-H7        S-TRYADSVKGRFTVSIDNAKKMVYLQMNNLKPEDTAVYYC---NRPG-----------S
F57-E6        ARTYYADSVKDRFTISRDNAKNTVYLQMNSLKPEDTAVYYCAADQRPGLGLSRTGWGDFS
F57-G7        ARTYYADSVKGRCTISRDNAKNTVYLQMNNLKPEDTAVYYCAADRAPGLSLSRTGWGDFS
              . * *: ***.* *:* *::*: :***** ***********   . *            *

F57-E11       SWGQGTHVTVSSHHHHHH
F57-H7        SWGQGTQVTVSSHHHHHH
F57-E6        SWGQGTQVTVSSHHHHHH
F57-G7        SWGQGTQVTVSSHHHHHH
              ******:***********
```

Figure A4. Multiple sequence alignment of cellulosome-binding nanobodies isolated by UK. We received E6, E11, and H7 in house.

and have similar CDR3 lengths, suggesting they may bind the same cellulosome epitope. H7 has a much shorter CDR3 and less homology to E6, suggesting it binds a different epitope.

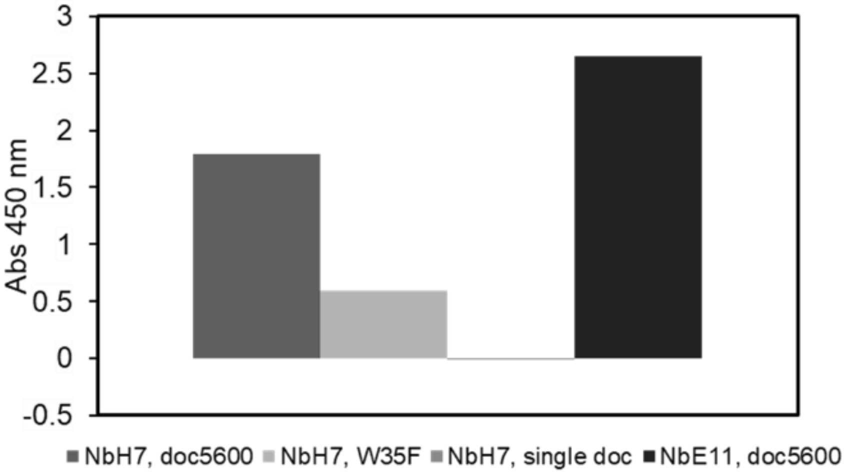


Figure A5. A single replicate indirect ELISA experiment with immobilized nanobody and soluble, purified double dockerin shows NbE11 and NbH7 also bind double dockerin. Double and single dockerins contained Strep tags and were detected using an HRP-conjugated antibody against Strep tag.

Indirect ELISA binding experiments indicate all three nanobodies bind purified double dockerin (Figure A5). Competitive binding experiments or epitope binning would be an interesting next step to characterize how these nanobodies bind double dockerin and could directly address the hypothesis of whether similar sequences breed overlapping epitopes in these nanobodies. This would be highly useful information for the proposed use of quantum dot labeled nanobodies to track cellulosome assembly dynamics, in which nanobodies labeled with different fluorophores bind different species to track cellulosome assembly. The work described in Chapter IV of this thesis suggests it is likely the NbE6 double dockerin binding epitope overlaps with cohesin's dockerin binding epitope, meaning NbE6 could not be used to track dockerin-cohesin assembly. Perhaps NbE11 or NbH7 do not compete with cohesin for dockerin binding and would thus be better suited for cellulosome imaging. This sort of competitive binding experiment between a nanobody putative dockerin binding partners would be very interesting to perform.