Ruminococcal cellulosome systems from rumen to human

Yonit Ben David[1], Bareket Dassa[1], Ilya Borovok[2], Raphael Lamed[2], Nicole M. Koropatkin[3], Eric C. Martens[3], Bryan A. White[4], Annick Bernalier-Donadille[5], Sylvia H Duncan[6], Harry J. Flint[6], Edward A. Bayer[1*†] and Sarah Moraïs[1*]

[1]*Department of Biological Chemistry, The Weizmann Institute of Science, Rehovot, Israel.*

[2]*Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Ramat Aviv, Israel.*

[3]*Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA.*

[4]*Department of Animal Sciences and Institute for Genomic Biology, University of Illinois, Urbana, Illinois, USA*

[5]*INRA Unite de Microbiologie UR454, CR de Clermont-Ferrand/Theix, Saint-Genes Champanelle , France.*

[6]*Microbiology Group, Rowett Institute of Nutrition and Health, University of Aberdeen, Aberdeen, United Kingdom.*

*Corresponding authors:
Edward A. Bayer, Department of Biological Chemistry, The Weizmann Institute of Science, 234 Herzl Street, Rehovot 7610001 Israel. Tel: (+972)-8-934-2373. Fax: (+972)-8-934-4118. Email: ed.bayer@weizmann.ac.il and
Sarah Moraïs, Department of Biological Chemistry, The Weizmann Institute of Science, 234 Herzl Street, Rehovot 7610001 Israel. Tel: (+972)-8-934-2729. Fax: (+972)-8-934-4118. Email: sarahv@weizmann.ac.il

Conflict of interest: †EAB is on the editorial board of *Environmental Microbiology*

Running title: Human ruminococcal cellulosome

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1462-2920.12868

Summary

A cellulolytic fiber-degrading bacterium, *Ruminococcus champanellensis*, was isolated from human faecal samples, and its genome was recently sequenced. Bioinformatic analysis of the *R. champanellensis* genome revealed numerous cohesin and dockerin modules, the basic elements of the cellulosome, and manual sequencing of partially sequenced genomic segments revealed two large tandem scaffoldin-coding genes that form part of a gene cluster. Representative *R. champanellensis* dockerins were tested against putative cohesins, and the results revealed three different cohesin-dockerin binding profiles which implied two major types of cellulosome architectures: (i) an intricate cell-bound system and (ii) a simplistic cell-free system composed of a single cohesin-containing scaffoldin. The cell-bound system can adopt various enzymatic architectures, ranging from a single enzyme to a large enzymatic complex comprising up to 11 enzymes. The variety of cellulosomal components together with adaptor proteins may infer a very tight regulation of its components. The cellulosome system of the human gut bacterium *R. champanellensis* closely resembles that of the bovine rumen bacterium *Ruminococcus flavefaciens*. The two species contain orthologous gene clusters comprising fundamental components of cellulosome architecture. Since *R. champanellensis* is the only human colonic bacterium known to degrade crystalline cellulose, it may thus represent a keystone species in the human gut.

Introduction

More than 100 trillion microorganisms colonize the human gut, with very high cell density ($>10^{11}$ cells/g) (Flint and Bayer, 2008). Their influence on the host is very significant, since they can affect nutrient absorption and production (Goodman et al., 2009), energy balance (Turnbaugh et al., 2006) and regulation of the immune system (Lee and Mazmanian, 2010). Moreover, the status of human gut microorganism is associated with many diseases, e.g., colonic cancer, diabetes, irritable bowel syndrome and inflammatory bowel disease (Young et al., 2005; Kerckhoffs et al., 2016; Vaarala, 2012). The major phyla that were detected in the human microbiota are the Gram-negative Bacteroidetes and the Gram-positive Firmicutes, while Actinobacteria, Proteobacteria and Verrucomicrobia have been also identified (Eckburg et al., 2005). In addition to bacteria, archaea and eukaryotes are in smaller numbers in the healthy human gut (Eckburg et al., 2005; Scanlan and Marchesi, 2008).

Among the gut microbiota, only a few species, particularly Firmicutes from the Clostridial cluster IV (Ruminococcaceae), have been recognized as cellulose-degrading bacteria (Chassard et al., 2010). Polysaccharide substrates in the large intestine are hydrolyzed by gut bacteria into smaller fragments that are fermented to short-chain fatty acids (mainly acetate, propionate and butyrate) and gases (H_2, CO_2) (Mackie et al., 1997; Flint et al., 2012). Herbivorous mammals get their main energy, up to 70%, from degradation of plant materials by gut microorganisms (Flint and Bayer, 2008). In humans, however, the energy contribution of gut microorganisms is relatively small (no more than 10%) (McNeil, 1984). Nevertheless, as mentioned above, they can have a great impact on human health.

Members of the Bacteroidetes phylum demonstrate a highly diverse ability for degradation of polysaccharide materials, including starch, xylan, pectin, galactomannan, arabinogalactan, etc (Bayliss and Houston, 1984; Xu et al., 2003; Martens et al., 2011). Nevertheless, only *Bacteroides cellulosilyticus,* is known to degrade certain forms of cellulose (Robert et al., 2007; McNulty et al., 2013). Members of the Firmicutes phylum can utilize starch, cellulose, xylan, galactomannan and other hemicelluloses and are considered to be more substrate-specific than the Bacteroidetes (Salyers et al., 1977; Chassard et al., 2007; Chassard et al., 2012; Ze et al., 2012) including species whose populations respond to specific dietary polysaccharides (Walker et al., 2011). The Firmicutes have been studied less intensively, and their role in polysaccharide breakdown is only now starting to be revealed. Despite this, a few species among them have been suggested to represent keystone species in polysaccharide degradation (Ze et al., 2013).

In many ways, the mechanisms of polysaccharide utilization by gut microorganisms remain unclear; yet, two main paradigms have been investigated widely, namely the starch utilization system (Sus) and the cellulosome system (White et al., 2014). The Sus and the Sus-like Polysaccharide Utilization Loci (PUL) are highly abundant and conserved in the Bacteroidetes phylum (Thomas et al., 2011). There are many different PUL systems, each of which may degrade a specific substrate, such as, pectin, xylan and galactomannan (Martens et al., 2011; McNulty et al., 2013). The archetypal Sus cluster of *Bacteroides thetaiotaomicron* is composed of eight genes, and four of these, SusDEFG, are localized to the outer membrane. SusD is an α-helical starch-binding protein that is required for glycan uptake via SusC, a TonB-dependent receptor in the outer membrane (Koropatkin et al., 2008; Cameron et al., 2014). A hallmark feature of PULs is the inclusion of homologs of *susCD* (Martens et al., 2009). The lipoproteins SusE and SusF are composed of tandem starch-binding domains, similar to carbohydrate-binding modules, yet lack enzymatic activity (Cameron et al., 2012). SusG is an α-amylase that has two non-catalytic starch-binding sites that enhance catalysis on solid substrates yet are dispensable for growth on soluble

starch, unless combined with a genetic knock-out for *susEF* (Koropatkin and Smith, 2010; Cameron et al., 2012). The SusCDEFG protein are believed to physically interact and work together to bind, degrade and import starch (Cho and Salyers, 2001; Karunatilaka et al., 2014). This separation of binding and catalytic functions among distinct polypeptides that work together as a multiprotein complex is somewhat analogous to the cellulosome. The other three Sus proteins include a regulator protein, SusR, and two periplasmic enzymes, SusA and SusB (D'Elia and Salyers, 1996; Shipman et al., 2000; Martens et al., 2009). That the Sus of *B. thetaiotaomicron* is a paradigm that describes glycan acquisition in the Bacteroidetes has been supported by recent in-depth studies of other Sus-like systems, encoded within PULs that target xyloglucan (Larsbrink et al., 2014), porphyran (Hehemann et al., 2010), and α-mannan (Cuskin et al., 2015). In contrast, the Gram-positive mechanisms of human gut bacteria in general have remained poorly explored, and the presence of cellulosome-producing bacteria has not been reported.

The cellulosome is an extracellular multi-enzyme complex, first discovered in the anaerobic, cellulolytic bacterium *Clostridium thermocellum* (Bayer et al., 1983), that is considered a very efficient cellulase system for plant cell-wall degradation. The "classical" cellulosome is composed of a non-catalytic “scaffoldin” subunit, and two interacting modules termed "cohesin" and "dockerin" that dictate cellulosome assembly (Bayer et al., 2008). Cellulosomal enzymes comprise mostly carbohydrate-active enzymes (CAZymes), i.e., glycoside hydrolases (GHs), carbohydrate esterases (CEs) and polysaccharide lyases (PLs). In addition to their catalytic modules, these enzymes contain a dockerin module, which interacts tightly with the cohesin modules found on the scaffoldin subunit (Bayer et al., 2004). The different scaffoldins contain various numbers of cohesins. They may also contain a carbohydrate-binding module (CBM), which mediates the interaction with the substrate, as well as either a dockerin or an anchoring motif involved in attachment to the bacterial cell surface. Cellulosome organization facilitates stronger synergism among the catalytic units. Additionally, the proximity between the cell-bound cellulosome and the

substrate minimizes the diffusion of the hydrolytic products and enzymes, providing the bacterium with a competitive advantage over non-cellulosomal organisms (Bayer et al., 1983; Shoham et al., 1999).

The assembly of cellulosome components into the mature complex relies on cohesin-dockerin interactions. These interactions are among the strongest protein-protein interactions found in nature (Mechaly et al., 2001; Stahl et al., 2012; Schoeler et al., 2014). Cohesin-dockerin interactions are considered to be species-specific, although divergent intraspecies interactions are evident in some bacteria and some cross-species interactions have also been observed (Pages et al., 1997b; Haimovitz et al., 2008). Three types of cohesins and dockerins have been defined according to phylogenetic sequence analysis (Bayer et al., 2004). Dockerins are relatively short protein modules characterized by two reiterated segments, each of which possesses a Ca^{+2}-binding loop and an α-helix, together termed F-hand motifs (Bayer et al., 2004). The binding of two calcium ions has been found to be crucial for appropriate dockerin folding (Karpol et al., 2008). In each segment, positions 1, 3, 5, 9 and 12 of the loop coordinate Ca^{+2} binding and are usually occupied by aspartic acid or asparagine (Carvalho et al., 2003; Handelsman et al., 2004). In addition, it has been proposed that positions 10, 11, 17, 18 and 22 recognize and mediate the binding of the cohesin (Pages et al., 1997b; Mechaly et al., 2001). Owing to the reiterated segments that form a pair of cohesin-binding surfaces on the dockerin, a dual mode of binding may ensue (Carvalho et al., 2007).

Ruminococcus champanellensis is a recently described (Chassard et al., 2012) anaerobic, mesophilic, Gram-positive bacterium found in the human colon, whose genome has been sequenced. It is the only human colonic bacterium so far reported to efficiently degrade pure cellulose (Avicel and filter paper). In addition, it can utilize xylan and cellobiose but not starch or glucose (Chassard et al., 2012; Ze et al., 2013). Phylogenetic analysis has revealed that the *R. champanellensis* genome is related to those of the cellulolytic rumen bacterium, *R. flavefaciens*

(<91.5% 16S rRNA gene sequence similarity) (Walker et al., 2008). Moreover, it is the only bacterium in the human colon reported so far whose genome has been found to encode for a wide variety of cellulosomal elements, i.e., dockerins and cohesins [this report]. These findings may reflect the formation of cellulosome system(s) in the human gut and suggest a new mechanism for carbohydrate utilization in the colon. Therefore, understanding their role in the human gut ecosystem is extremely interesting and can contribute to the development of strategies for microbial manipulation and personalized medicine.

In this study we describe the discovery of a cellulosome system in the human colon bacterium, *R. champanellensis*. Bioinformatic analysis of the genome of *R. champanellensis* has revealed 64 dockerin and 20 cohesin modules. All of the putative cohesins and 24 representative dockerins were cloned into matching fusion-protein cassettes and overexpressed. Different proteomic methods were performed in order to evaluate initial cohesin-dockerin interactions, the results of which served to predict numerous types of cellulosome architectures in *R. champanellensis*.

Results

Genomic analysis of *R. champanellensis* reveals potential cellulosomal genes

The 2.57-Mb draft genome sequence of *R. champanellensis* 18P13 has recently been published. Intriguingly, our initial bioinformatic analysis based on this sequence indicated genes consistent with cellulosomal components. In this early analysis, 11 putative cohesin and 62 putative dockerin sequences were revealed. In subsequent analyses, manual examination of the gaps of the draft genome sequence of *R. champanellensis* revealed two additional incomplete genes containing both cohesins and dockerins (*scaA* and *scaB*). These genes were part of a gene cluster that included a previously identified scaffoldin (*scaC*). This type of gene cluster has been found in several other cellulosome-producing bacteria (Bayer et al., 2008). The missing sequences, which included the

complete *scaA* and *scaB* genes (GenBank KP341766), were recovered by genome walking (Supplemental Figure S1), and a total of nine additional putative cohesins and 2 putative dockerins were thus detected. The genome of the bovine rumen bacterium *R. flavefaciens* contains an orthologous gene cluster with a similar gene arrangement (Rincon et al., 2005; Jindou et al., 2008).

All putative cohesin- and dockerin-containing proteins, except one *Rc*-Doc3550 (GI 291583550), carry N-terminal signal peptides, suggesting that these proteins are secreted. Analysis of the *Rc*-Doc3550 sequence has predicted a transmembrane domain in the middle of the protein, which would position the dockerin on the exterior of the membrane. The 20 cohesins were found on eleven different scaffoldin-like proteins, which were termed ScaA to ScaK (Figure 1). ScaA, ScaB and ScaJ scaffoldins carry more than one putative cohesin, and contain 2, 7 and 3 cohesin modules, respectively. ScaE has a putative C-terminal sortase signal motif, which is considered to be a cell wall anchoring sequence (Rincon et al., 2005). ScaC, ScaD, ScaF, ScaG and ScaH are small adaptor proteins that contain a single predicted cohesin module together with a dockerin module. In addition, ScaH carries a domain annotated as a putative lipase or esterase module. ScaK possesses a GH25 catalytic domain (putative lysozyme activity) in its C-terminal region, while ScaI has a region of unknown function.

Comparison of the *R. champanellensis* cohesin sequences to those of *C. thermocellum*, *Acetivibrio cellulolyticus* and *R. flavefaciens* was performed (Figure 2). It was revealed that most of the *R. champanellensis* cohesins cannot be classified into the two classical groups of cohesins, type I and type II. Instead, they are more similar to *R. flavefaciens* cohesins, most of which are classified as type III cohesins.

In terms of sequence similarities, the two cohesins of ScaA exhibit 98% protein sequence identity with each other, and they likely share the same dockerin specificity. Moreover, the ScaA architecture (an X-module, 2 cohesins and a dockerin) is similar to ScaA from *R. flavefaciens* FD1. The alignments of the cohesin sequences from ScaB form two major groups, based on sequence

similarity. The first contains CohB1, CohB2 and CohB3 (i.e., the first three cohesins from scaffoldin B), the latter two sharing 93% identity with each other and 77% identity relative to CohB1. The second group of ScaB cohesins comprises the remaining cohesins, where each pair is highly similar to each other: CohB4 and CohB5 (99% identity), and CohB6 and CohB7 (94% identity). The identity between the two pairs is 40% (54% similarity), which may indicate an additional subdivision of this group. The overall modular organization of ScaB (7 cohesins, an X-module and a dockerin module) is analogous to ScaB of *R. flavefaciens* strain 17 (as opposed to strain FD-1). The *R. champanellensis* ScaA and ScaB cohesins are classified together with CohH.

R. champanellensis CohC and CohD, which exhibit 54% identity to each other, are related to *R. flavefaciens* CohC, a type I-like cohesin. Consequently, these two cohesins can also be classified as type I. ScaC and ScaD of *R. champanellensis* also share the same modular arrangement (a single cohesin attached to dockerin), similar to that of *R. flavefaciens* ScaC. ScaF and ScaG cohesins share 35% identity (and 48% similarity). Concerning ScaJ cohesins, CohJ1 is related to CohJ2, sharing 32% identity (and 49% similarity); and the two additional cohesins of ScaJ, CohJ2 and CohJ3, share 35% identity (and 54% similarity) to each other. Thus, the predicted cohesin sequences show substantial similarity and divergence, which may well translate into corresponding similarities and differences in dockerin specificities. Curiously, *Rc*-ScaI has an enigmatic cohesin sequence comprising two inverted parts separated by a linker. Therefore, it was not included in the phylogenetic tree (Figure 2) and comparative analysis of the cohesins.

Based on the CAZy website, the *R. champanellensis* genome contains 107 CAZyme modules, more than half of which are found on dockerin-containing proteins. Among these modules, 54 are glycoside hydrolases belonging to 25 GH families, mainly cellulases from families 5 and 9 (Table 1). *R. champanellensis* also possesses GH8 and GH48 glycoside-hydrolase families, which are known to play a key role in cellulose hydrolysis and are often distinctive components of known cellulosomes (Bayer et al., 2013). In addition, three important xylanase families were

observed, namely, GH10, GH11 and GH43. These combined data suggest a distinctive role for *R. champanellensis* as a cellulose-degrading bacterium.

Many enzymes of *R. champanellensis* seem to have a complex multi-modular structure composed of more than one catalytic module, together with a CBM and/or dockerin module. For example, the protein *Rc*-GH10B (GI 291544573) contains GH10 and GH43 modules together with two CBM22 and one CBM6 modules. This complex modular structure is very common among enzymatic polypeptides from cellulolytic bacterial species (Bayer et al., 1998). By contrast, the glycoside hydrolases in the non-cellulolytic Bacteroidetes, were mainly found in a single-domain polypeptide. This may reflect the difference between the types of degraded carbohydrate substrates, i.e., complex and insoluble in comparison to small and soluble (Flint et al., 2008).

Selection of representative cohesins and dockerins

The specific interaction between the cohesin and dockerin pair involves many factors, which cannot be predicted by bioinformatic analysis alone. Therefore, all 20 predicted cohesins and a broad set of dockerins from *R. champanellensis* were selected for further investigation. In this manner, we can expect to receive a general understanding of cellulosome assembly in this bacterium. This is particularly true in a case like the cellulosome system in *R. champanellensis*, where the various dockerin sequences appear to be relatively divergent.

Dockerin modules are characterized by two reiterated segments, each consisting of a Ca^{2+}-binding loop followed by an α-helix. However, their internal sequence can vary greatly between different species and within the same species. Previous studies have shown that dockerins of similar sequence, especially in the putative cohesin-recognition residues, usually interact with the same cohesin (Mechaly et al., 2001; Pinheiro et al., 2009). Therefore, the 64 dockerins of *R. champanellensis* were aligned, and then clustered into four groups. The two dockerins from ScaA

and ScaB revealed unique sequences and were therefore not included in any of the latter groups (Figure 3 and Figure S2).

The dockerin sequences were clustered according to the conservation pattern of their internal Ca^{+2}-binding repeats and their putative helix regions. Sequence logos of the reiterated sequences of the different groups are presented in Figure 3b. Different patterns were observed for the putative cohesin-recognition residues (positions 10, 11, 17, 18 and 22) and for their flanking positions in the putative helix region. Group 1 dockerins exhibit a conserved Val and Leu residues at the putative binding positions 10 and 17. In addition, this group has very conserved Ala residues in positions 13 and 21. In Group 2 dockerins, the end portions of the putative helix, positions 18-22, are characterized by the conserved sequence RYVAQ in the first segment and RYLAH in the second. The dockerins in Groups 3 and 4 exhibit relatively high sequence variation, yet Group 3 can generally be recognized by positive amino acids in positions 17 and 18 in the first putative helix and Gln in position 17 of the second. Group 4 shows similar features but in opposite segment arrangements. DocA and DocB both have an additional amino acid at position 7 in the second segment and were thus not classified in either of the groups. However, the putative recognition residues of DocA are more similar to those of Group 2, while DocB is more similar to the Group 1 dockerins.

Representative dockerins from each group were selected according to several parameters: (1) Dockerins on cohesin-containing proteins (scaffoldins) were all selected, as these were presumed to be crucial for cellulosome architecture. (2) Dockerins from proteins having a catalytic module present (e.g., GH5, GH8, GH9, GH10, GH11, GH13, GH43 and GH48) were selected preferentially. (3) Dockerins with either high or low sequence conservation within the same group, especially in the putative recognition residues, were also preferentially selected. In total, 24 dockerins were selected and examined in this work (Table 1 and Figure S2).

The selected cohesins and dockerins were expressed in *E. coli* cells using two different cassettes for cohesins and dockerins, respectively. The cohesin modules were fused to a CBM3a from *C. thermocellum* (CBM-Coh) while the dockerin modules were fused to xylanase T6 from *Geobacillus stearothermophilus* (Xyn-Doc) with an added His tag on the N terminus. The use of fused proteins has been found to enhance the stability and the expression level of the cohesin and dockerin modules compared to their expression as part of the native protein or in the free state (Barak et al., 2005). Moreover, it allows a relatively simple way for detection of the different cohesin-dockerin interactions. Following expression, the cohesins and dockerins were purified on either cellulose beads or a Ni-NTA affinity column, respectively.

Cohesin-dockerin microarray

This study is the first to explore cohesin-dockerin interactions of *R. champanellensis*, and the number of possible interaction pairs among the 20 cohesins and 24 dockerins selected for this study was calculated at 480. Therefore, we used the CBM-based microarray method, which allowed us to examine every dockerin separately against a large number of cohesins in one reaction. The cellulose slides contained the 11 cohesins (as CBM-Cohs) of *R. champanellensi*s that were detected in the first bioinformatic analysis using the published sequenced genome. The nine additional cohesins of ScaA and ScaB that were detected by deep examination of the unsequenced parts of the genome were analyzed for their dockerin-specific interaction by ELISA assay. In addition, a set of 17 cohesins from the following bacterial species: *A. cellulolyticus, Bacteroides cellulosolvens, Clostridium acetobutylicum, Clostridium cellulolyticum, C. thermocellum, Ruminococcus bromii* and *R. flavefaciens*, were applied together on the slide to explore the possibility of cross-species interactions. The addition of cohesins from different species enabled us to examine the specificity of the cohesin-dockerin interaction, to explore possible cross-species interactions and to verify the

accuracy of the method. A protein containing only a CBM module was also expressed in order to be used as a negative control, whereby the CBM alone without the fused cohesin module, would not be expected to interact with the Xyn-Docs. In addition, a xylanase-CBM fusion protein was expressed for use as a positive control, to ensure that the anti-Xyn antibodies interact with the xylanase.

The cohesin-dockerin interactions were tested by exposing the different dockerins to the cellulose slides (CBM-Coh microarray), each dockerin to a separate slide. Each dockerin was tested in at least two separate experiments. The microarray was scanned against two fluorescence dyes, Cy3 and Cy5. The Cy3 dye was conjugated to rabbit α-xylanase primary antibody, to indicate the presence of Xyn-Doc proteins (a positive result indicated a positive reaction). In addition, a Cy5 dye was labeled with rabbit α-CBM antibody in order to examine the extent of binding of the test CBM-fused cohesin to the cellulose slide. In total, 24 dockerins were tested by the microarray method, taken from three species: 22 from *R. champanellensis*, one from *C. thermocellum* and one from *A. flavefaciens*. The last two were used as positive controls to ensure the specificity of the system. Representative slides are shown in Figure 4 (all slides are included in the Supplemental Figure S3).

These 22 dockerins of *R. champanellensis* were examined against 28 cohesins from different species. Table 2 summarizes the newly discovered cohesin-dockerin interactions in *R. champanellensis*. Interaction intensity was determined by the number of clearly seen rows among the three different concentrations, representing a semi-quantitative estimation of the cohesin-dockerin binding.

Evaluation of cohesin-dockerin binding affinities by ELISA

In order to confirm the microarray results, different ELISA tests were performed. At least one interaction from each dockerin group was thus examined. Figure 5 presents the results of

selected ELISA tests for *R. champanellensis*. ELISA experiments were performed either with cohesins or dockerins in the coating step. Cohesin-dockerin interactions are known to be calcium dependent (Yaron et al., 1995; Karpol et al., 2008). Therefore, in some cases, selected interactions were examined in the absence of calcium (removed upon addition of EDTA) in order to verify calcium dependency.

The ELISA method was also used for examination of the cohesin-dockerin binding interactions of the ScaA and ScaB scaffoldins (Table 2). The two cohesins of ScaA share 98% sequence identity, and we therefore presumed that they would interact with the same dockerin partners. Indeed, both CBM-CohA2 and ScaA (containing both A1 and A2 cohesin modules) interacted positively with several dockerins from Group 2 in a similar manner. The cohesins of ScaB can be divided in two groups, B1/B2/B3 and B4/B5/B6/B7 according to their sequence similarities (Figure 2). The first group B1/B2/B3 is closely related to the ScaA cohesins and shared the same binding profile as CohA2 and the recombinant ScaA. The second group B4/B5/B6/B7 is also related to ScaA cohesins but with a more distant connection. It appeared that CBM-CohB4 and CBM-CohB5/B6 interact with the same dockerins from Group 2 but with the addition of the ScaA dockerin (Table 2). Cohesins B6 and B7 share 94% sequence identity. Both were expressed separately but failed to interact with any of the dockerin partners, ostensibly due to incorrect modular folding. Nevertheless, we can assume that the both CohB6 and CohB7 are bona fide cohesins on the basis of sequence similarities, but their precise specificity is currently unknown.

In total 480 intra-species and 374 inter-species interactions were tested by microarray and ELISA techniques, among them 64 interactions were found to be positive (Table 2).

From the microarray data, the cohesin of ScaI appeared to have many interactions with dockerins from Groups 3 and 4, but the intensity of the signal was low in most cases. We therefore examined the interaction of CohI with several of the designated dockerins using indirect ELISA (iELISA), which has proved in the past to be a more sensitive method than the standard ELISA

(Slutzki et al., 2012a), and therefore it was used to examine a few selected CohI interactions to verify its interaction with designated dockerins (Figure 5c and 5d). The ELISA results were found to be generally consistent with the microarray results.

Dockerin-binding profile of *R. champanellensis*

Group 1 dockerins. The selected dockerins from Group 1 (DocJ, DocH, DocF, DocG and Doc3939) and DocB were found to interact strongly with CohE, which bears a sortase cell surface-attachment motif at its C terminus. However, as opposed to the other members in this group, dockerins DocJ and Doc3939 failed to interact with cohesin J1. It seems logical that DocJ would fail to interact with CohJ1, since both modules are located in the same protein. In both DocJ and Doc3939, the reason for this finding may be the presence of a negatively charged amino acid residue (Asp or Glu) instead of the uncharged Gln in position 18 of the dockerin's first duplicated segment (Supplemental Figure S4). This position was previously demonstrated to play an important role in cohesin-dockerin interactions (Pages et al., 1997b; Mechaly et al., 2001). In addition, DocG seems to bind to CohJ1 with higher affinity than DocH and DocF (Figure 5a, Table 2). This observation may reflect slight differences among the dockerin sequences. In any case, by virtue of the high degree of symmetry of the putative recognition residues in the duplicated dockerin segments (Figure 3), all of the interacting Group 1 dockerins would be expected to exhibit a dual-binding mode of action (Carvalho et al., 2007) with CohE and CohJ1.

Based on the above, it seems that the dockerins in Group 1 are critical for cellulosome assembly, since they mediate between the bacterium and the outer environment through the interaction with the cell wall-attached cohesin of ScaE. It is interesting to note that the parent proteins of all dockerins that interact with CohE appeared to be structural proteins and not enzymatic in nature (Table 1).

Group 2 dockerins. The dockerins of Group 2 exhibited specific interactions with cohesins H and I, the two cohesins of ScaA and the seven cohesins of ScaB, with a lower affinity to the ScaI cohesin (Table 2). Moreover, in the case of cohesin H, ELISA tests demonstrated the dependency on calcium ions in its interaction with DocC, since complex formation between them was significantly reduced by the addition of EDTA (Figure 5b). There is a striking lack of symmetry between the putative recognition residues in the duplicated dockerin segments (Supplemental Figure S2), which would strongly suggest a single mode of binding with the target cohesins. Sequence homology between the 17 dockerin sequences of this group, particularly in the two duplicated segments is highly conserved. Therefore, it can be assumed that all the proteins in this group interact with CohH and CohI, with a preference for cohesin H.

ScaA dockerin (DocA) could be related to this group in view of its interactions with CohH and cohesins B4, B5 and B6 (Table 2). As opposed to other members of this group, DocA failed to interact with its own cohesins A1 and A2 and cohesins B1, B2 and B3. It seems logical that DocA would fail to interact with its own cohesins, and since B1, B2 and B3 have strong similarity with ScaA cohesins, it may follow suit.

Group 3 and 4 dockerins. Dockerins of Groups 3 and 4 were found to share the same binding profile (Table 2). In total, 12 dockerins were selected from both groups. Six dockerins, from the GH98, GH10B, GH43C, 4116, 4559 and 4133 proteins, interacted with the three designated cohesins, CohC, CohD and CohI. Dockerins GH98 and GH11 reacted only with CohC and CohD, while dockerin GH43A interacted exclusively with CohD. These results were quite unexpected since the two dockerin groups appeared to have relatively different sequences. However, between the two groups, the two sets of duplicated putative recognition residues showed a lack of symmetry between them. Therefore, as in the case of Group 2, this may indicate a single mode of binding for

Groups 3 and 4, which would allow a wider range of combinations among the cohesin-dockerin pairs.

The dominant glycoside hydrolase family in Groups 3 and 4 is GH43, while families GH8, GH9, GH10 and GH11 are also present (Table 1). GH43, GH10 and GH11 are families known to exhibit hemicellulose-degrading activity, where the latter two exhibit xylanase activity. As a result, the enzymes associated with these groups of dockerins may be more involved in the degradation of hemicellulosic substrates than cellulose. In addition, many proteins in these groups contain regions of LRR motifs and unknown function. As mentioned for Group 2, the proteins in these two groups may be integrated into the cell surface-attached cellulosome complex via the ScaC and ScaD adaptor proteins, or, alternatively, they may bind to ScaI and act in a cell-free manner.

Based on the above-described findings, cell-bound and cell-free cellulosome architectures were proposed for *R. champanellensis*. The two schematic models are presented in Figure 6.

In many cellulosome-producing bacteria, the cohesin-dockerin interaction appears to be largely species specific. However, a study by Haimovitz *et al*. (Haimovitz et al., 2008) has also demonstrated interspecies recognition in selected cases both for type I and type II interactions. Here we have examined possible cross-interaction between *R. champanellensis* dockerins to 17 cohesins from different species. Interestingly, three interactions were detected: *Rc*-DocGH11 interacted with *Ct*-CohOlpC, *Rc*-DocGH9B interacted with *Rf*-CohC and *Ct*-DocS interacted with *Rc*-CohC (Figure S3). It is likely that the cross-reactivity between *R. champanellensis* and *C. thermocellum* is a result of spurious interaction due to coincidental similarity in their sequence motifs rather than a true functional interaction, since these two bacteria exist in very different environments and temperature conditions. In this context, the Lys-Arg motif is prevalent in both *C. thermocellum* dockerins as well as in the *R. champanellensis* dockerins of Group 3 and 4. The interaction between the *R. champanellensis* dockerin GH9b to *R. flavefaciens* CohC is probably based on its phylogenetic connection to *R. champanellensis* CohC and CohD.

groupings were defined previously for the 223 dockerins detected in the *R. flavefaciens* FD1 genome, based largely on sequence relationships (Rincon et al., 2010), but it is not possible at present to correlate these with the dockerin groupings that we have defined here in *R. champanellensis* based on their binding specificities. Nevertheless we can note that dockerins associated with common GH families, including GH10, GH11, GH9 and GH43, were distributed across several dockerin groupings in both species.

Inactive cohesin and dockerin modules

Some of the modules examined in this work failed to recognize any of the tested cohesins or dockerins. Among the 20 selected *R. champanellensis* cohesins, seven appeared to be inactive (namely, B6, B7, F, G, J2, J3 and K). Although representative dockerins were selected carefully, dockerins with specific recognition for these cohesins may exist but were not selected for this study. Moreover, folding anomalies of the cohesins modules should also be taken into account.

All of the predicted cohesins of *R. champanellensis*, derived from the draft genome sequence, were tested in this study. Thus, it was surprising to find that four dockerins failed to interact with any of the cohesins; especially dockerins *Rc*-GH5B and *Rc*-GH8 whose sequences are very similar to those of active dockerins. Three of the inactive dockerins (GH5B, GH8 and GH9G) were therefore expressed as the intact wild-type protein rather than as Xyn-Doc chimaeras. Thus, although the *Rc*-GH9G and *Rc*-GH5B dockerins failed to interact with any of the cohesin partners when inserted in the Xyn cassette, they successfully interacted with their respective group-specific cohesins (Table 2). The same was not true for the GH8 dockerin and the CohJ2 and CohJ3 cohesin modules, which remained inactive even when expressed as full proteins. Gel filtration experiments have showed folding irregularities for CohG and DocGH8 (data not shown), which can explain their failure to interact with appropriate dockerin or cohesin. Dockerin *Rc*-3550 is markedly

different in its putative recognition residues compared to the other dockerins, this dockerin may thus be able to interact with one of the inactive cohesins. Moreover, the currently available draft genome sequence of *R. champanellensis* is incomplete with numerous gaps. Consequently, it is still possible that not all of the cohesin and dockerin modules have yet been detected.

In any case, as a rule, the dockerin sequences are generally identifiable with a very high degree of confidence. Positive identification of the cohesin sequences, on the other hand, is often more obscure. Therefore, unless a predicted cohesin sequence is irrefutably similar to a previously identified and confirmed cohesin, its definitive classification as such can be verified only upon conclusive experimental evidence.

Discussion

R. champanellensis is the first cellulolytic bacterium found in the human gut to have genes associated with cellulosomal components, i.e., cohesin and dockerin modules. Cellulosomal subunits interconnect to form an efficient multi-enzyme cellulose-degrading machine through cohesin-dockerin interactions. In doing so, they represent the fundamental components of the cellulosome assembly. In this study, initial structures of cellulosome complexes in this bacterium were predicted based on the 64 newly discovered cohesin-dockerin interactions.

By piecing together the puzzle of cohesin-dockerin interactions and the modular arrangement of their parent molecules, we can predict that the overall architecture of the cellulosome system in *R. champanellensis* is very complex, and somewhat reminiscent of that of *R. flavefaciens* in the cow rumen (Dassa et al., 2014). The cell-bound cellulosome of *R. champanellensis* is anchored to the cell surface by ScaE via its sortase signal motif (Figure 6**)**. This scaffoldin is the only scaffoldin identified to bear a recognizable segment consistent with a cell-anchoring function. ScaE can then

interact with ScaB to form a major enzymatic complex by incorporating a maximum of three enzymes or adaptor scaffoldins (ScaC- and ScaD-mediated enzymes) on its first three cohesins and two ScaA scaffoldins, each bearing two enzymes, on cohesins 4 and 5. The exact involvement of cohesins 6 and 7 is currently undefined.

The cohesin of ScaE can also interact directly with dockerins of adaptor proteins from Group 1, namely, ScaF, ScaG, ScaH and ScaJ. Three of these proteins, ScaF, ScaG and ScaH, can also attach to CohJ1. Of these scaffoldins, only ScaH, can, in turn, interact directly with dockerin-containing enzymes (Group 2), either alone or via ScaC and ScaD adaptor proteins, to attach single enzymes to the cell surface. Alternatively, the ScaA dockerin can also interact with ScaH to form a two-enzyme cell-bound complex. In addition, the enzyme-related function of the ScaH scaffoldin is underscored by its resident SGNH-hydrolase module, which has been reported to facilitate hydrolysis of ester and amide bonds in a wide range of substrates including complex polysaccharides (Dalrymple et al., 1997; Reina et al., 2007). Finally, ScaC and ScaD would presumably serve in a regulatory role by selective integration of alternative dockerin-containing proteins, e.g., mainly hemicellulases, CBM modules and peptidases.

The ScaB dockerin and dockerins of Group 1 may be of particular interest, since they were found to interact directly with the cell-anchoring scaffoldin, ScaE. ScaB, in particular, with its multiplicity of cohesins, provides the major basis for cellulosome structure. ScaE can thus mediate the proximity between the bacterial cell wall and the enzymes. However, the major mechanism for attachment of the cell to the substrate has yet to be determined. One possible candidate would be protein 3939 whose dockerin interacts directly with ScaE. This protein contains multiple FN3 (fibronectin type III) domains and two PKD (polycystic kidney disease) domains, both of which are relatively common components in bacterial cellulase systems and may be involved in protein-protein or protein-carbohydrate interactions (Lohning et al., 1996). This protein may therefore have an important role in carbohydrate degradation. Interestingly, an untested member of the Group 1

dockerins (protein 3199), contains a cadherin-like domain which may also suggest a carbohydrate-binding function (Fraiberg et al., 2011), thereby mediating a possible connection between the bacterium and the cellulosic substrate.

Most of the proteins in Group 2 represent glycoside hydrolase enzymes, mainly cellulases or closely associated enzymes of families 5, 9, 44, 48 and 74; some of which also contain a CBM module (Table 1). Hence, the proteins that bear Group 2 dockerins would appear to play a major role in cellulose degradation. In addition, two cohesin-containing proteins, ScaC and ScaD, are also included in this group. Intriguingly, the two latter monovalent scaffoldins likely play an adaptor role (Rincon et al., 2004), since they bind to Group 3 and 4 enzymes, many of which appear to be hemicellulases. The integration of ScaC and ScaD into the cellulosomal system of *R. champanellensis* may therefore serve in a regulatory capacity to alter the repertoire of enzymes that then act on selected hemicellulosic substrates that emerge during plant cell wall hydrolysis. However, some of the dockerin-containing proteins, mainly from Groups 3 and 4, lack confirmed carbohydrate-degrading components, thus indicating that some of the cohesin-dockerin interactions in this bacterium serve in a non-cellulosomal context, as previously suggested for other organisms (Peer et al., 2009). One possible role for these interactions is to enhance the interaction between the bacterium and the host epithelium cells.

As opposed to the above-described interactions among the *R. champanellensis* scaffoldins, ScaI represents a protein with a single unusual cohesin module and a region of unknown function. This may suggest the assembly of a cell-free cellulosome-like architecture, albeit in most cases, only a weak interaction would be expected between ScaI and the various proteins. A ScaI-mediated cell-free cellulosome-like system may be released into solution to degrade carbohydrates farther away from the bacterium. The concept of free cellulosome was described before for *A. cellulolyticus* and *C. cellulolyticum*, and was assumed to allow efficient degradation in cases where the substrate is abundant and remote from the bacterium (Artzi et al., 2014). In *A. cellulolyticus* and

C. cellulolyticum the main cellulosome scaffoldin consists of more than one cohesin and CBM modules, in contrast to the simple monovalent nature of the ScaI modular architecture. Alternatively, ScaI may either protect a free dockerin from adverse environmental conditions or play a role as a transient molecular shuttle, to transfer dockerin-bearing components to a more permanent position within the cellulosome complex (Pages et al., 1997a; Pinheiro et al., 2009).

Unlike more complex cellulosomes, this bacterium has a relatively simple cellulosome that could assemble up to 11 enzymes. The intricacy of cellulosome architecture may be related to the importance of dietary fibers in the diet of the host. While recalcitrant dietary fibers are the main energy source of herbivorous animals, transit times and conditions in the human large intestine are less conducive to the extensive fermentation of such material, with the result that humans, in common with other omnivores, select more accessible forms of fiber in their diets. This can be expected to have an impact both on the microbial community and on microbial metabolism in the colon (Flint et al., 2008). Although *R. champanellensis* was isolated using spinach cell walls and is able to degrade filter paper cellulose (Chassard et al., 2012), this species may be adapted to degrading dietary fiber that is less recalcitrant than that available to *R. flavefaciens* in the rumen. The relatively compact cellulosome of *R. champanellensis* may, nevertheless, explain why this species is, so far, unique among isolated human gut bacteria in its ability to degrade insoluble filter paper cellulose. It is thus possible that this species plays a key role in releasing energy from certain types of dietary fiber. Breakdown products from dietary fiber have a great impact on human health, and the efficiency of this breakdown may depend on the populations of specialist bacteria such as *R. champanellensis*. Mechanistic understanding will therefore contribute to the development of strategies for microbial manipulation, in order to prevent and/or treat health disorders and consequent metabolic processes. Moreover, the study of these special bacteria will help improve our understanding of the ecology and metabolism of the gut microbiota.

Since *Ruminococcus* is one of the major genera found in the adult human microbiota (Eckburg et al., 2005), we could expect that additional human gut bacteria could potentially express cellulosomal genes. In this context, two additional strains, *Ruminococcus* sp. CAG:379 and *Ruminococcus* sp. CAG:624, were also isolated from the human gut. The former closely resembles *R. champanellensis* and the latter seems to be strongly related to *R. flavefaciens* strain FD1. All four strains contain a gene cluster containing several scaffoldins with similar gene arrangements (Figure 7). *R. champanellensis* and *Ruminococcus* sp. CAG:379 exhibit 96% and 99% sequence similarity between their *scaC* and *scaE* genes, respectively. The third human gut isolate, *Ruminococcus* sp. CAG:624, and the bovine rumen *R. flavefaciens* FD1 contain very similar clusters with the addition of a *cttA* gene (Rincon et al., 2007) that is apparently lacking in *R. champanellensis* and *Ruminococcus* sp. CAG:379 genomes. Moreover, the genomes of both *R. champanellensis* and *Ruminococcus* sp. CAG:379 possess a *scaE* gene, phylogenetically similar to those that appear immediately downstream of the *cttA* gene in *Ruminococcus* sp. CAG:624 and *R. flavefaciens*, but apparently located outside of the *sca* gene cluster. More studies in this direction could provide further insight into cellulosome involvement in the human gut microbiota and its possible connection to the *R. flavefaciens* cellulosome in ruminants.

Anaerobic microbial communities demonstrate extensive metabolic cross-feeding, which involve fermentation products like hydrogen and lactate, as well as partial substrate degradation products. Primary degraders, like *R. champanellensis*, can break down insoluble complex carbohydrates into soluble polysaccharides which in turn can be utilized by non-cellulolytic bacteria (Flint et al., 2007). Robert and Bernalier-Donadille (Robert and Bernalier-Donadille, 2003) have suggested that the presence and development of methanogens in the colon are strongly dependent on H_2-producing genera, like *Ruminococcus* and *Enterococcus*. In turn, efficient growth of H_2-producing cellulolytic bacteria is increased, due to the removal of H_2 by methanogens, acetogens and sulphate-reducing species (Latham and Wolin, 1977). Therefore the discovery of a

cellulosome system in this bacterium could provide it with a critical advantage over other species in the human gut ecosystem.

Non-digestible carbohydrates are considered to comprise the main energy source for microbial growth in the human colon (Duncan et al., 2007). Hence, the human diet has a major impact on the microbial population and metabolism in the colon (Flint et al., 2008). *R. champanellensis* could thus represent a keystone species in the human gut (Ze et al., 2013), since is the only human colonic bacterium so far reported to degrade crystalline cellulosic substrates and might therefore be expected to initiate degradation of a wide range of plant material. The presence of a cellulosome system in this bacterium would support this argument. Such a keystone role has been proposed previously with respect to starch fermentation for the related species *R. bromii*, which is a highly specialized degrader of particulate starch, in view of evidence that human volunteers lacking this species fail to fully ferment resistant starch present in their diet (Walker et al., 2011; Ze et al., 2012).

Understanding the molecular basis for novel cohesin-dockerin interactions will extend our knowledge of cellulosome organization in different species. The cellulosomal elements that form the relatively simple architecture of the largest *R. champanellensis* cellulosome (11 enzymes) could thus be used in designer cellulosomes to integrate select copies of desired enzymes. The different cohesin and dockerin pairs can thus be included as components of designer cellulosomes, which can be used as a tool for understanding cellulosome action and for future biotechnological application, such as production of biofuels and waste management (Bayer et al., 2007).

Experimental Procedures

Bioinformatic analysis

The genome sequence of *R. champanellensis* (strain 18P13 = JCM 17042) was obtained from GenBank (FP929052.1). The genome was sequenced by the Pathogen Genomics group at the Wellcome Trust Sanger Institute (UK) as part of the EU MetaHit project (http://www.sanger.ac.uk/resources/downloads/bacteria/metahit/). Prediction of cohesins and dockerins modular sequences were performed using the BLASTP and TblastN algorithm (Altschul et al., 1997), employing known cohesin and dockerin sequences as queries. Hits of E-value higher than 10^{-4} were examined individually. Analysis of Carbohydrate-Active Enzymes (Cazymes) was performed using the CAZy database (http://www.cazy.org). Sequences were then further analyzed to identify additional modular structures using the aid of CD-search (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) (Marchler-Bauer and Bryant, 2004). Multiple sequence alignments of cohesins and dockerins were generated using ClustalW2 [http://www.ebi.ac.uk/Tools/msa/clustalw2/]. Phylogenic trees were created using the Robust Phylogenetic Analysis (Dereeper et al., 2008) tool from the Phylogeny.fr website. Analysis was accomplished using the default bootstrapping "one click" mode and then visually edited using the TreeGraph2 software (Stöver and Müller, 2010). Signal peptide sequences were predicted using the SignalP server [http://www.cbs.dtu.dk/services/SignalP/]. Logos of the dockerin sequences were created with Weblogo v.2.8.2 (http://weblogo.berkeley.edu/).

Cloning of CBM-fused cohesins and xylanase-fused dockerins

Cohesin and dockerin genes were amplified by PCR from the *R. champanellensis* 18P13 genomic DNA, which was prepared from cell pellets using the FastDNA spin kit for soil (MP Biomedicals, France), using specific primers. The list of primers used in this study is provided in the Supplementary Materials (Table S1). Cohesin genes were designed to have BamHI and XhoI restriction sites. Dockerin genes were designed to have KpnI and BamHI restriction sites. In cases where a BamHI sequence was found in the desired gene, the BglII sequence was inserted instead,

since their cleavage sites produce compatible cohesive ends. DNA samples were purified using a PCR purification kit (Real Biotech Corporation, RBC, Taiwan) and double-digested by appropriate FastDigest restriction enzymes (Thermo scientific, Fermentas UAB, Vilnius, Lithuania). The different modules were assembled in linearized pET28a-CBM-Coh or pET9d-Xyn-Doc cassettes. The CBM-Coh gene cassette (Barak et al., 2005) consists of a family 3a CBM from the *C. thermocellum* CipA scaffoldin cloned into plasmid pET28a (Novagen Inc., Madison, WI, USA), into which any cohesin gene can be introduced between BamHI and XhoI restriction sites of the plasmid. The Xyn-Doc gene cassette (Barak et al., 2005) consists of xylanase T6 from *G. stearothermophilus* with an N-terminal His-tag cloned into plasmid pET9d (Novagen Inc., Madison, WI, USA), into which any dockerin-encoding sequence can be introduced between the KpnI and BamHI restriction sites of the plasmid.

Protein expression

E. coli BL21 (DE3) cells were transformed with the desired plasmid and grown at 37°C in 300-500 ml LB medium, supplemented with 50 μg/ml kanamycin (Sigma-Aldrich Chemical Co, St. Louis, Missouri), with the inclusion of 2 mM $CaCl_2$ for dockerin-containing proteins, to A_{600}≈0.8-1. Protein expression was induced by addition of 0.1 mM Isopropyl-1-thio-β-D-galactoside (IPTG) (Fermentas UAB), and the growth was continued either at 37°C for 3 h or at 16°C for ~16 h (according to predetermined conditions). Cells were harvested by centrifugation (5000 rpm, 15 min) and resuspended in 30 ml TBS (Tris-buffered saline, 137 mM NaCl, 2.7 mM KCL, 25 mM Tris-HCl, pH=7.4) or TBS supplemented with 5 mM imidazole for dockerin-containing proteins (Merck KGaA, Darmstadt, Germany), and stored at -20°C. Immediately before purification, the thawed cells were sonicated and then centrifuged (14,000 rpm, 30 min, 4°C). The supernatant fluids were used for further steps for protein purification.

Purification of CBM-containing cohesin

Supernatant fluids containing the cohesin-containing proteins were added to 2 g of macroporous beaded cellulose preswollen gel (IONTOSORB, Usti nad Labem, Czech Republic), and incubated for 1 h, with rotation at 4°C. The mixture was then loaded onto a column by gravity, washed with 100 ml of TBS containing 1 M NaCl and then with 100 ml TBS. Three 5 ml elutions of 1% triethanolamine (TEA) were then collected. The fractions were subjected to SDS-PAGE in order to assess protein purity, and then dialyzed against TBS overnight at 4°C.

Purification of Xyn-containing dockerin

The supernatant fluids containing the dockerin-bearing proteins were mixed with ~4 ml Ni-NTA for 1 h on a 20-ml Econo-pack column, on a rotator at 4°C (batch purification system). The column was then washed by gravity flow with 50-100 ml wash buffer (TBS, 15 mM imidazole). Elution was performed first using 10 ml 100 mM imidazole, followed by 10 ml 250 mM imidazole. Fractions (2 ml) were collected and subjected to SDS-PAGE. The fractions containing relatively pure proteins were pooled, and $CaCl_2$ (10 mM), as well as protease-inhibitor cocktail, was added. The proteins were dialyzed overnight at 4°C with TBS supplemented with 5 mM $CaCl_2$.

Protein concentration

Protein concentrations were estimated by absorbance at 280 nm. Extinction coefficient was determined based on the known amino acid composition of each protein using VectorNTI version 11 computer program. Some proteins were concentrated using Amicon ultra concentrators (Millipore, Ireland). Proteins were stored in 50% (v/v) glycerol at -20°C.

CBM-based microarray

A manual spotter MicroCASTer (Schleicher & Schuell) and a Micro Grid 610 (DIGILAB) were utilized to print proteins onto the cellulose-coated glass slides (Type-GSRC-1 from Advanced Microdevices pvt. Ltd.). Protein samples were diluted in TBS, pH 7.4, to concentrations of 9, 3, 1, 0.3 and 0.1 µM and applied in quadruplicate to the cellulose slides. The printed microarrays were kept at 4°C prior to application.

The printed microarrays were quenched by incubating the slides in blocking buffer (1% BSA in TBS with 10 mM $CaCl_2$ and 0.05% Tween 20) at room temperature for 30 min. The slides were then incubated at room temperature with the desired Xyn-Doc sample at a concentration of 3 nM in blocking buffer for 30 min. After washing 3 times (5 min each) with washing buffer (TBS with 10 mM $CaCl_2$ and 0.05% Tween 20), fluorescent staining was accomplished by adding Cy3-labeled anti-Xyn T6 antibody and Cy5-labeled anti-CBM3a antibody (diluted 1:1000) in blocking buffer, and the slides were incubated for 30 min. The probed slides were washed again 3 times, air-dried and scanned for fluorescence signals using a Typhoon 9400 Variable Mode Imager GE Healthcare Bio-Sciences AB (Uppsala, Sweden).

The labeling of the fluorescent antibodies was performed using GE Healthcare's N-hydroxysuccinimide-ester-activated Cy-5 dye and Cy-3 kits. The dyes were resuspended in 0.1 M sodium carbonate buffer, pH 9, and mixed with the antibody (1 mg in 1 ml), according to the manufacturer's instructions. Free dye was removed by dialysis against TBS. The fluorescence-labeled antibody was stored in 50% glycerol at –20°C.

ELISA affinity assay

The standard affinity-based ELISA procedure was performed as described previously (Barak et al., 2005). The coating step was performed with 10-30 nM of the desired proteins. A concentration gradient of Xyn-Doc or CBM-Coh (0.01-1000 nM) was then applied to the coated Maxisorp 96-well plate (Greiner Bio-One, Belgium). In some cases, 10 mM EDTA was substituted

for the $CaCl_2$ in all solutions to determine calcium dependence of the interaction. The dose-response curve was fitted to the data using GraphPad Prism 5 (GraphPad Software, Inc., La Jolla, CA).

Indirect ELISA (iELISA)

The indirect ELISA-based method, iELISA, is more sensitive than conventional ELISA, since the procedure is performed under conditions of much lower dockerin concentrations, and the interaction takes place in the soluble phase. Maxisorp ELISA plates (Greiner Bio-One, Belgium) were coated overnight at 4°C with 30 nM of desired CBM-Coh protein in 0.1 M Na_2CO_3 (pH 9), 100 µl/well. The wells were blocked with 100 µl/well of blocking buffer (TBS, 10 mM $CaCl_2$, 0.05% Tween 20, 2% BSA) for 1 h at 37°C, and the blocking solution was then discarded. In parallel, a pre-equilibration step was preformed; a concentration gradient of CBM-Coh (0.01-1000 nM) was prepared in non-absorbing 96-well plates. To all of the wells, Xyn-Doc was added to a final concentration of 1-20 nM in a total volume of 150 µl. The pre-equilibration step was allowed to proceed for 1 h. Afterwards, 100 µl samples from the interaction in previous step were transferred to the wells of the MaxiSorp plate and incubated for 20 min. The solution was then discarded, and the plate was washed once with Washing Buffer (TBS, 10 mM $CaCl_2$, 0.05% Tween 20). The antibody interaction steps and the chromogenic substrate reaction were performed as described for the ELISA (Barak et al., 2005). A detailed description of the method can be found in Slutzki *et al*. (Slutzki et al., 2012a, b).

Analytical gel filtration chromatography

Prepacked SuperdexTM 200 10/300 GL column was obtained from GE Healthcare Bio-Sciences (Pittsburgh, PA). Samples of 200 µl were injected into the column using an autosampler.

Tris-buffered saline (TBS), pH 7.4, containing 10 mM $CaCl_2$ was used as running buffer at a flow rate of 0.5 $ml \cdot min^{-1}$. Proteins were detected using a UV detector at a wavelength of 280 nm.

Acknowledgements

The authors appreciate the kind assistance of Miriam Lerner (ImmunArray Ltd. Company, Rehovot, Israel) with experiments involving the MicroGrid II arrayer. This research was supported by a grant (No. 1349) to EAB also from the Israel Science Foundation (ISF) and a grant (No. 24/11) issued to RL by The Sidney E. Frank Foundation also through the ISF. Additional support was obtained from the establishment of an Israeli Center of Research Excellence (I-CORE Center No. 1527/14) managed by the Israel Science Foundation, from the United States-Israel Binational Science Foundation (BSF), Jerusalem, Israel, by the Weizmann Institute of Science Alternative Energy Research Initiative (AERI) and the Helmsley Foundation. The authors also appreciate the support of the European Union, Area NMP.2013.1.1-2: Self-assembly of naturally occurring nanosystems: CellulosomePlus Project number: 604530 and an ERA-IB Consortium (EIB.12.022), acronym FiberFuel. HF and SHD acknowledge support from the Scottish Government Food Land and People programme and from BBSRC grant no. BB/L009951/1. In addition, EAB is grateful for a grant from the F. Warren Hellman Grant for Alternative Energy Research in Israel in support of alternative energy research in Israel administered by the Israel Strategic Alternative Energy Foundation (I-SAEF). E.A.B. is the incumbent of The Maynard I. and Elaine Wishner Chair of Bio-organic Chemistry.

References

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

Artzi, L., Dassa, B., Borovok, I., Shamshoum, M., Lamed, R., and Bayer, E.A. (2014) Cellulosomics of the cellulolytic thermophile *Clostridium clariflavum*. *Biotechnology for Biofuels* **7**: 100.

Barak, Y., Handelsman, T., Nakar, D., Mechaly, A., Lamed, R., Shoham, Y., and Bayer, E.A. (2005) Matching fusion-protein systems for affinity analysis of two interacting families of proteins: The cohesin-dockerin interaction. *Journal of Molecular Recognition* **18**: 491-501.

Bayer, E.A., Kenig, R., and Lamed, R. (1983) Adherence of *Clostridium thermocellum* to cellulose. *Journal of Bacteriology* **156**: 818-827.

Bayer, E.A., Lamed, R., and Himmel, M.E. (2007) The potential of cellulases and cellulosomes for cellulosic waste management. *Curr Opin Biotechnol* **18**: 237-245.

Bayer, E.A., Shoham, Y., and Lamed, R. (2013) The Prokaryotes: Lignocellulose-Decomposing Bacteria and Their Enzyme Systems. In *The Prokaryotes, Fourth Edition*. Rosenberg, E. (ed). Berlin: Springer-Verlag, pp. 216-266.

Bayer, E.A., Shimon, L.J.W., Lamed, R., and Shoham, Y. (1998) Cellulosomes: structure and ultrastructure. *J Struct Biol* **124**: 221-234.

Bayer, E.A., Belaich, J.-P., Shoham, Y., and Lamed, R. (2004) The cellulosomes: Multi-enzyme machines for degradation of plant cell wall polysaccharides. *Annual Review of Microbiology* **58**: 521-554.

Bayer, E.A., Lamed, R., White, B.A., and Flint, H.J. (2008) From cellulosomes to cellulosomics. *Chem Rec* **8**: 364-377.

Bayliss, C.E., and Houston, A.P. (1984) Characterization of plant polysaccharide- and mucin-fermenting anaerobic bacteria from human feces. *Appl Environ Microbiol* **48**: 626-632.

Cameron, E.A., Maynard, M.A., Smith, C.J., Smith, T.J., Koropatkin, N.M., and Martens, E.C. (2012) Multidomain Carbohydrate-binding Proteins Involved in *Bacteroides thetaiotaomicron* Starch Metabolism. *J Biol Chem* **287**: 34614-34625.

Cameron, E.A., Kwiatkowski, K.J., Lee, B.H., Hamaker, B.R., Koropatkin, N.M., and Martens, E.C. (2014) Multifunctional nutrient-binding proteins adapt human symbiotic bacteria for glycan competition in the gut by separately promoting enhanced sensing and catalysis. *MBio* **5**: e01441-01414.

Carvalho, A.L., Dias, F.M., Prates, J.A., Nagy, T., Gilbert, H.J., Davies, G.J. et al. (2003) Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. *Proc Natl Acad Sci U S A* **100**: 13809-13814.

Carvalho, A.L., Dias, F.M., Nagy, T., Prates, J.A., Proctor, M.R., Smith, N. et al. (2007) Evidence for a dual binding mode of dockerin modules to cohesins. *Proc Natl Acad Sci U S A* **104**: 3089-3094.

Chassard, C., Delmas, E., Robert, C., and Bernalier-Donadille, A. (2010) The cellulose-degrading microbial community of the human gut varies according to the presence or absence of methanogens. *FEMS Microbiol Ecol* **74**: 205-213.

Chassard, C., Goumy, V., Leclerc, M., Del'homme, C., and Bernalier-Donadille, A. (2007) Characterization of the xylan-degrading microbial community from human faeces. *FEMS Microbiol Ecol* **61**: 121-131.

Chassard, C., Delmas, E., Robert, C., Lawson, P.A., and Bernalier-Donadille, A. (2012) *Ruminococcus champanellensis sp. nov.*, a cellulose-degrading bacterium from human gut microbiota. *Int J Syst Evol Microbiol* **62**: 138-143.

Cho, K.H., and Salyers, A.A. (2001) Biochemical analysis of interactions between outer membrane proteins that contribute to starch utilization by *Bacteroides thetaiotaomicron*. *J Bacteriol* **183**: 7224-7230.

Cuskin, F., Lowe, E.C., Temple, M.J., Zhu, Y., Cameron, E.A., Pudlo, N.A. et al. (2015) Human gut Bacteroidetes can utilize yeast mannan through a selfish mechanism. *Nature* **517**: 165-169.

D'Elia, J.N., and Salyers, A.A. (1996) Contribution of a neopullulanase, a pullulanase, and an alpha-glucosidase to growth of *Bacteroides thetaiotaomicron* on starch. *J Bacteriol* **178**: 7173-7179.

Dalrymple, B.P., Cybinski, D.H., Layton, I., McSweeney, C.S., Xue, G.P., Swadling, Y.J., and Lowry, J.B. (1997) Three *Neocallimastix patriciarum* esterases associated with the degradation of complex polysaccharides are members of a new family of hydrolases. *Microbiology* **143 (Pt 8)**: 2605-2614.

Dassa, B., Borovok, I., Ruimy-Israeli, V., Lamed, R., Flint, H.J., Duncan, S.H. et al. (2014) Rumen cellulosomics: divergent fiber-degrading strategies revealed by comparative genome-wide analysis of six ruminococcal strains. *PLoS One* **9**: e99221.

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F. et al. (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* **36**: W465-469.

Duncan, S.H., Belenguer, A., Holtrop, G., Johnstone, A.M., Flint, H.J., and Lobley, G.E. (2007) Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces. *Appl Environ Microbiol* **73**: 1073-1078.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M. et al. (2005) Diversity of the human intestinal microbial flora. *Science* **308**: 1635-1638.

Flint, H.J., and Bayer, E.A. (2008) Plant cell wall breakdown by anaerobic microorganisms from the Mammalian digestive tract. *Ann N Y Acad Sci* **1125**: 280-288.

Flint, H.J., Duncan, S.H., Scott, K.P., and Louis, P. (2007) Interactions and competition within the microbial community of the human colon: links between diet and health. *Environ Microbiol* **9**: 1101-1111.

Flint, H.J., Scott, K.P., Louis, P., and Duncan, S.H. (2012) The role of the gut microbiota in nutrition and health. *Nat Rev Gastroenterol Hepatol* **9**: 577-589.

Flint, H.J., Bayer, E.A., Rincon, M.T., Lamed, R., and White, B.A. (2008) Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nat Rev Microbiol* **6**: 121-131.

Fraiberg, M., Borovok, I., Bayer, E.A., Weiner, R.M., and Lamed, R. (2011) Cadherin domains in the polysaccharide-degrading marine bacterium *Saccharophagus degradans* 2-40 are carbohydrate-binding modules. *J Bacteriol* **193**: 283-285.

Goodman, A.L., McNulty, N.P., Zhao, Y., Leip, D., Mitra, R.D., Lozupone, C.A. et al. (2009) Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* **6**: 279-289.

Haimovitz, R., Barak, Y., Morag, E., Voronov-Goldman, M., Shoham, Y., Lamed, R., and Bayer, E.A. (2008) Cohesin-dockerin microarray: Diverse specificities between two complementary families of interacting protein modules. *Proteomics* **8**: 968-979.

Handelsman, T., Barak, Y., Nakar, D., Mechaly, A., Lamed, R., Shoham, Y., and Bayer, E.A. (2004) Cohesin-dockerin interaction in cellulosome assembly: A single Asp-to-Asn mutation disrupts high-affinity cohesin-dockerin binding. *FEBS Letters* **572**: 195-200.

Hehemann, J.H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M., and Michel, G. (2010) Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**: 908-912.

Jindou, S., Brulc, J.M., Levy-Assaraf, M., Rincon, M.T., Flint, H.J., Berg, M.E. et al. (2008) Cellulosome gene cluster analysis for gauging the diversity of the ruminal cellulolytic bacterium *Ruminococcus flavefaciens*. *FEMS Microbiol Lett* **285**: 188-194.

Karpol, A., Barak, Y., Lamed, R., Shoham, Y., and Bayer, E.A. (2008) Functional asymmetry in cohesin binding belies inherent symmetry of the dockerin module: Insight into cellulosome assembly revealed by systematic mutagenesis. *Biochemical Journal* **410**: 331-338.

Karunatilaka, K.S., Cameron, E.A., Martens, E.C., Koropatkin, N.M., and Biteen, J.S. (2014) Superresolution imaging captures carbohydrate utilization dynamics in human gut symbionts. *mBio* **5**: e02172.

Kerckhoffs, A.P., Ben-Amor, K., Samsom, M., van der Rest, M.E., de Vogel, J., Knol, J., and Akkermans, L.M. (2011) Molecular analysis of faecal and duodenal samples reveals significantly higher prevalence and numbers of *Pseudomonas aeruginosa* in irritable bowel syndrome. *J Med Microbiol* **60**: 236-245.

Koropatkin, N.M., and Smith, T.J. (2010) SusG: a unique cell-membrane-associated alpha-amylase from a prominent human gut symbiont targets complex starch molecules. *Structure* **18**: 200-215.

Koropatkin, N.M., Martens, E.C., Gordon, J.I., and Smith, T.J. (2008) Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. *Structure* **16**: 1105-1115.

Larsbrink, J., Rogers, T.E., Hemsworth, G.R., McKee, L.S., Tauzin, A.S., Spadiut, O. et al. (2014) A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature* **506**: 498-502.

Latham, M.J., and Wolin, M.J. (1977) Fermentation of cellulose by *Ruminococcus flavefaciens* in the presence and absence of *Methanobacterium ruminantium*. *Appl Environ Microbiol* **34**: 297-301.

Lee, Y.K., and Mazmanian, S.K. (2010) Has the microbiota played a critical role in the evolution of the adaptive immune system? *Science* **330**: 1768-1773.

Lohelig, C., Pohlschmidt, M., Glucksmann-Kuis, M.A., Duyk, G., Bork, P., Schneider, M.O. et al. (1996) Structural motifs of the PKD1 protein. *Nephrol Dial Transplant* **11 Suppl 6**: 2-4.

Macfarlane, R., White, B., Macfarlane, G., and Gibson, G. (1997) Carbohydrate Fermentation, Energy Transduction and Gas Metabolism in the Human Large Intestine. In *Gastrointestinal Microbiology*: Springer US, pp. 269-318.

Marchler-Bauer, A., and Bryant, S.H. (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* **32**: W327-331.

Martens, E.C., Koropatkin, N.M., Smith, T.J., and Gordon, J.I. (2009) Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm. *J Biol Chem* **284**: 24673-24677.

Martens, E.C., Lowe, E.C., Chiang, H., Pudlo, N.A., Wu, M., McNulty, N.P. et al. (2011) Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biol* **9**: e1001221.

McNeil, N.I. (1984) The contribution of the large intestine to energy supplies in man. *Am J Clin Nutr* **39**: 338-342.

McNulty, N.P., Wu, M., Erickson, A.R., Pan, C., Erickson, B.K., Martens, E.C. et al. (2013) Effects of diet on resource utilization by a model human gut microbiota containing *Bacteroides cellulosilyticus WH2*, a symbiont with an extensive glycobiome. *PLoS Biol* **11**: e1001637.

Mechaly, A., Fierobe, H.-P., Belaich, A., Belaich, J.-P., Lamed, R., Shoham, Y., and Bayer, E.A. (2001) Cohesin-dockerin interaction in cellulosome assembly: A single hydroxyl group of a dockerin domain distinguishes between non-recognition and high-affinity recognition (Erratum). *Journal of Biological Chemistry* **276**: 19678.

Pages, S., Gal, L., Belaich, A., Gaudin, C., Tardif, C., and Belaich, J.P. (1997a) Role of scaffolding protein CipC of *Clostridium cellulolyticum* in cellulose degradation. *J Bacteriol* **179**: 2810-2816.

Pages, S., Belaich, A., Belaich, J.-P., Morag, E., Lamed, R., Shoham, Y., and Bayer, E.A. (1997b) Species-specificity of the cohesin-dockerin interaction between *Clostridium thermocellum* and *Clostridium cellulolyticum:* Prediction of specificity determinants of the dockerin domain. *Proteins* **29**: 517-527.

Peer, A., Smith, S.P., Bayer, E.A., Lamed, R., and Borovok, I. (2009) Noncellulosomal cohesin- and dockerin-like modules in the three domains of life. *FEMS Microbiol Lett* **291**: 1-16.

Pinheiro, B.A., Gilbert, H.J., Sakka, K., Sakka, K., Fernandes, V.O., Prates, J.A. et al. (2009) Functional insights into the role of novel type I cohesin and dockerin domains from *Clostridium thermocellum*. *Biochem J* **424**: 375-384.

Reina, J.J., Guerrero, C., and Heredia, A. (2007) Isolation, characterization, and localization of AgaSGNH cDNA: a new SGNH-motif plant hydrolase specific to *Agave americana L. leaf epidermis*. *J Exp Bot* **58**: 2717-2731.

Rincon, M.T., Cepeljnik, T., Martin, J.C., Lamed, R., Barak, Y., Bayer, E.A., and Flint, H.J. (2005) Unconventional mode of attachment of the *Ruminococcus flavefaciens* cellulosome to the cell surface. *J Bacteriol* **187**: 7569-7578.

Rincon, M.T., Cepeljnik, T., Martin, J.C., Barak, Y., Lamed, R., Bayer, E.A., and Flint, H.J. (2007) A novel cell surface-anchored cellulose-binding protein encoded by the sca gene cluster of *Ruminococcus flavefaciens*. *J Bacteriol* **189**: 4774-4783.

Rincon, M.T., Martin, J.C., Aurilia, V., McCrae, S.I., Rucklidge, G.J., Reid, M.D. et al. (2004) ScaC, an adaptor protein carrying a novel cohesin that expands the dockerin-binding repertoire of the *Ruminococcus flavefaciens* 17 cellulosome. *J Bacteriol* **186**: 2576-2585.

Rincon, M.T., Dassa, B., Flint, H.J., Travis, A.J., Jindou, S., Borovok, I. et al. (2010) Abundance and diversity of dockerin-containing proteins in the fiber-degrading rumen bacterium, *Ruminococcus flavefaciens* FD-1. *PLoS One* **5**: e12476.

Robert, C., and Bernalier-Donadille, A. (2003) The cellulolytic microflora of the human colon: evidence of microcrystalline cellulose-degrading bacteria in methane-excreting subjects. *FEMS Microbiol Ecol* **46**: 81-89.

Robert, C., Chassard, C., Lawson, P.A., and Bernalier-Donadille, A. (2007) *Bacteroides cellulosilyticus sp. nov.*, a cellulolytic bacterium from the human gut microbial community. *Int J Syst Evol Microbiol* **57**: 1516-1520.

Salyers, A.A., West, S.E., Vercellotti, J.R., and Wilkins, T.D. (1977) Fermentation of mucins and plant polysaccharides by anaerobic bacteria from the human colon. *Appl Environ Microbiol* **34**: 529-533.

Scanlan, P.D., and Marchesi, J.R. (2008) Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. *Isme J* **2**: 1183-1193.

Schoeler, C., Malinowska, K.H., Bernardi, R.C., Milles, L.F., Jobst, M.A., Durner, E. et al. (2014) Ultrastable cellulosome-adhesion complex tightens under load. *Nat Commun* **5**: 5635.

Shipman, J.A., Berleman, J.E., and Salyers, A.A. (2000) Characterization of four outer membrane proteins involved in binding starch to the cell surface of *Bacteroides thetaiotaomicron*. *J Bacteriol* **182**: 5365-5372.

Shoham, Y., Lamed, R., and Bayer, E.A. (1999) The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides. *Trends Microbiol* **7**: 275-281.

Slutzki, M., Barak, Y., Reshef, D., Schueler-Furman, O., Lamed, R., and Bayer, E.A. (2012a) Measurements of relative binding of cohesin and dockerin mutants using an advanced ELISA technique for high-affinity interactions. *Methods Enzymol* **510**: 417-428.

Slutzki, M., Barak, Y., Reshef, D., Schueler-Furman, O., Lamed, R., and Bayer, E.A. (2012b) Indirect ELISA-based approach for comparative measurement of high-affinity cohesin-dockerin interactions. *J Mol Recognit* **25**: 616-622.

Stahl, S.W., Nash, M.A., Fried, D.B., Slutzki, M., Barak, Y., Bayer, E.A., and Gaub, H.E. (2012) Single-molecule dissection of the high-affinity cohesin-dockerin complex. *Proc Natl Acad Sci U S A* **109**: 20431-20436.

Stöver, B.C., and Müller, K.F. (2010) TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. **11**: 7.

Thomas, F., Hehemann, J.H., Rebuffet, E., Czjzek, M., and Michel, G. (2011) Environmental and gut bacteroidetes: the food connection. *Front Microbiol* **2**: 93.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., and Gordon, J.I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**: 1027-1031.

Vaarala, O. (2012) Gut microbiota and type 1 diabetes. *Rev Diabet Stud* **9**: 251-259.

Walker, A.W., Duncan, S.H., Harmsen, H.J., Holtrop, G., Welling, G.W., and Flint, H.J. (2008) The species composition of the human intestinal microbiota differs between particle-associated and liquid phase communities. *Environ Microbiol* **10**: 3275-3283.

Walker, A.W., Ince, J., Duncan, S.H., Webster, L.M., Holtrop, G., Ze, X. et al. (2011) Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J* **5**: 220-230.

White, B.A., Lamed, R., Bayer, E.A., and Flint, H.J. (2014) Biomass utilization by gut microbiomes. *Annu Rev Microbiol* **68**: 279-296.

Xu, J., Bjursell, M.K., Himrod, J., Deng, S., Carmichael, L.K., Chiang, H.C. et al. (2003) A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. *Science* **299**: 2074-2076.

Yaron, S., Morag, E., Bayer, E.A., Lamed, R., and Shoham, Y. (1995) Expression, purification and subunit-binding properties of cohesins 2 and 3 of the *Clostridium thermocellum* cellulosome. *FEBS Lett* **360**: 121-124.

Young, G.P., Hu, Y., Le Leu, R.K., and Nyskohus, L. (2005) Dietary fibre and colorectal cancer: a model for environment-gene interactions. *Mol Nutr Food Res* **49**: 571-584.

Ze, X., Duncan, S.H., Louis, P., and Flint, H.J. (2012) *Ruminococcus bromii* is a keystone species for the degradation of resistant starch in the human colon. *ISME J* **6**: 1535-1543.

Ze, X., Le Mougen, F., Duncan, S.H., Louis, P., and Flint, H.J. (2013) Some are more equal than others: the role of "keystone" species in the degradation of recalcitrant substrates. *Gut Microbes* **4**: 236-240.

Figures Legends

Fig. 1. Schematic representation of the cohesin-bearing scaffoldin proteins in *R. champanellensis* based on the respective genome sequences. SGNH, hydrolase-type esterase domain (IPR013830); GH25, a putative GH25-family domain sharing similarity to lysozyme.

Fig. 2. Phylogenetic relationship of *R. champanellensis* cohesins with previously defined, selected cohesins from other cellulosome-producing bacteria. Dendrogram of type I, II and III cohesin modules. The tree was constructed from cohesins selected from four different species, *R. champanellensis* (*Rc*, red), *R. flavefaciens* (*Rf*-FD1, blue), *C. thermocellum* (*Ct*, green) and *A. cellulolyticus* (*Ac*, pink). Bootstrapping confidence values higher than 0.8 are shown in black.

Fig. 3. Dockerin sequences of *R. champanellensis*. (**A**) Sequences of the duplicated segments of the ScaA and ScaB dockerins. (**B**) Sequence logos of the additional 62 *R. champanellensis* dockerins, divided into four groups by sequence homology. In each group, the two duplicated segments (1 and 2) are aligned, where the positions of calcium-binding residues are highlighted in cyan, and putative recognition residues are highlighted in yellow. The alignment of the complete set of dockerin sequences organized into the different groups, including the additional two *R. champanellensis* dockerins from ScaA and ScaB, is shown in Figure S2.

Fig. 4. Representative cohesin-dockerin recognition analyses using protein microarray. (**A**) Interaction of the *R. champanellensis* ScaF dockerin (*Rc*-XynDocF) with *R. champanellensis* ScaJ1 and ScaE cohesins (*Rc*-J1 and *Rc*-E) as CBM-Coh fusion proteins. (**B**) Preferential interaction of *R. champanellensis* GH10B dockerin (*Rc*-XynDocGH10B) with *R. champanellensis* ScaC, ScaD and (weakly) ScaI cohesins (*Rc*-C, *Rc*-D and *Rc*-I). Fluorescence scan showing Cy3-conjugated anti-Xyn antibody, indicating cohesin-dockerin binding. (**C**) Scan showing Cy5-conjugated anti-CBM antibody, indicating the relative amount of the different CBM-Coh samples applied to the slide. Selected cohesins from other species *A. cellulolyticus (Ac), B. cellulosolvens (Bc), C. acetobutylicum (Ca), C. cellulolyticum (Cc), C. thermocellum (Ct), R. bromii (Rb)* and *R. flavefaciens (Rf)* were included as controls. A Xyn-CBM fusion-protein served as a positive control (+) and as a marker, which indicates the relative location of all samples on the cellulose slide.

Fig. 5. *R. champanellensis* cohesin-dockerin binding measured by ELISA and iELISA assays. (**A**,**B**) ELISA experiments demonstrating different interaction specificities between selected cohesins and dockerins. CohJ1 interacted with DocG, weakly with DocF and DocH, and failed to interact with its own dockerin (DocJ). In (**B**), CohH interacts strongly with DocC, DocD and DocGH48, but failed to interact with DocGH10B. The interaction with DocC was calcium dependent and was abolished upon chelation with EDTA. (**C**,**D**) iELISA experiments demonstrated that DocGH10B interacted strongly with CohC, CohD and somewhat weaker with CohI. In (**D**), DocA132 showed moderate, weak and negligible binding to CohC, CohD and CohI, respectively. Error bars indicate the standard deviation from the mean of triplicate (ELISA) or duplicate (iELISA) samples from one experiment.

Fig. 16. Proposed cell-bound and cell-free cellulosome complexes in *R. champanellensis*. Different types of cohesin-dockerin interactions are color-coded. The binding specificities of cohesin modules of ScaB6/7, ScaJ2/3, ScaF and ScaG (shown in light gray) are yet to be determined. SGNH stands for lipase/esterase. Only the GH9B dockerin bound strongly to the ScaI cohesin (Table 1); other dockerins displayed comparatively weak binding.

Fig. 17. Comparison of *sca* gene clusters in four different ruminococcal strains. Organization of the *sca* gene clusters in (**A**) *R. champanellensis* strain 18P13, (**B**) *Ruminococcus* sp. CAG:379 (GenBank PRJNA222131), (**C**) *Ruminococcus* sp. CAG:624 (GenBank PRJNA222208) and (**D**) *R. flavefaciens* strain FD1 (GenBank PRJNA37767). The organization of the cluster in *R. flavefaciens* FD-1 as shown in (D) is indicative of those of all other known *R. flavefaciens* strains (i.e, 17, C94, B34b, C1a, JM1 and 007c). Grey rectangles represent unsequenced regions of the respective genome. Percentages of sequence identity of ScaC and ScaE proteins are indicated.

Table legends

Table 1. Dockerin-containing proteins of *R. champanellensis.*

Table 2. Cohesin-dockerin interactions in *R. champanellensis*: summary of cellulose microarray experiments. Twenty-four dockerins (rows), including the ScaA and ScaB dockerins and representatives of the four different groups, were checked against 20 cohesins (columns). Each dockerin was examined in a different slide containing all the test cohesins and relevant controls. Interaction intensity (number of pluses) was defined as the number of clearly labeled rows among the five different concentrations (See Supplemental Figure S3 for raw data). The two ScaA cohesins and the seven cohesins of ScaB were tested separately by ELISA tests. Only positive interactions appear in the table. See Table 1 for description of dockerin-bearing proteins that contain CAZy domains. In others, the numbers refer to the last 4 digits of the respective full GI number (i.e., 291XXXX).

Supporting information Legends

Table S1. Primers list

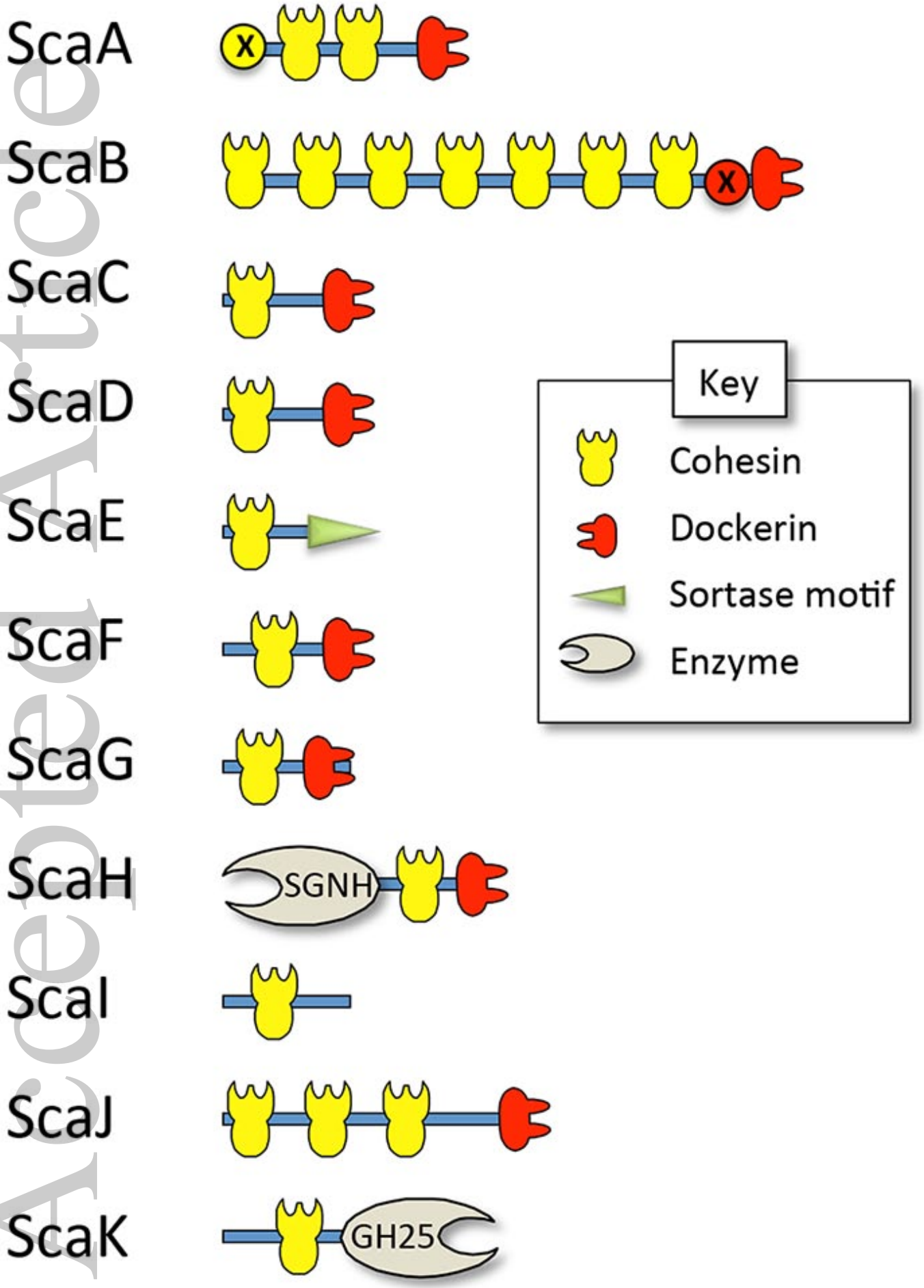
Fig. S1. Nucleotide sequence of the *Ruminococcus champanellensis* 18P13 ScaA/B Region of the Sca gene cluster coding for the cohesin-containing scaffoldins ScaA and ScaB. GenBank accession number KP341766. The coding sequence is shown in lowercase and the short intergenic region in highlighted uppercase.

Fig. S2. *R. champanellensis* dockerin alignment groups. The 64 dockerin sequences of *R. champanellensis* divided into 4 groups, using bioinformatics-based criteria. Each group is marked in a different color. Dockerins selected for this study are highlighted in green (see Table 1 for GI numbers of the parent proteins). Positions of calcium binding residues are shown in cyan, and putative recognition residues are shown in yellow.

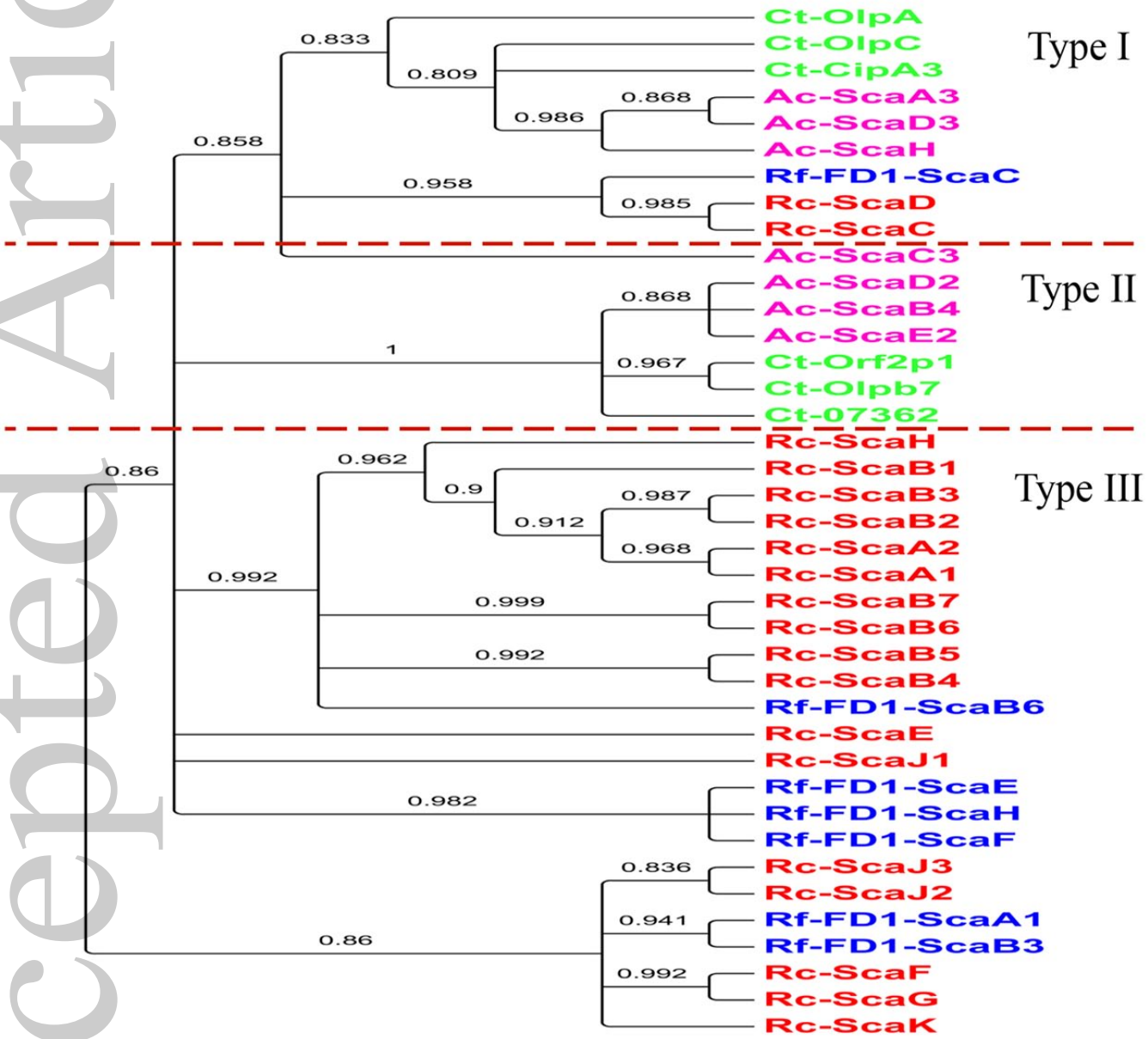
Fig. S3. Cellulose microarray results

The cellulose slides contained the 11 cohesins (as CBM-Cohs) of *R. champanellensi*s that were detected in the first bioinformatic analysis and 17 cohesins from different bacterial species. Every dockerin was tested on the cellulose slide. Fluorescence scanning, showing Cy3-conjugated anti-Xyn antibody, indicates cohesin-dockerin binding. Xyn-CBM proteins served as a positive control (+) and as a marker, which indicated the location of the samples on the cellulose slide.

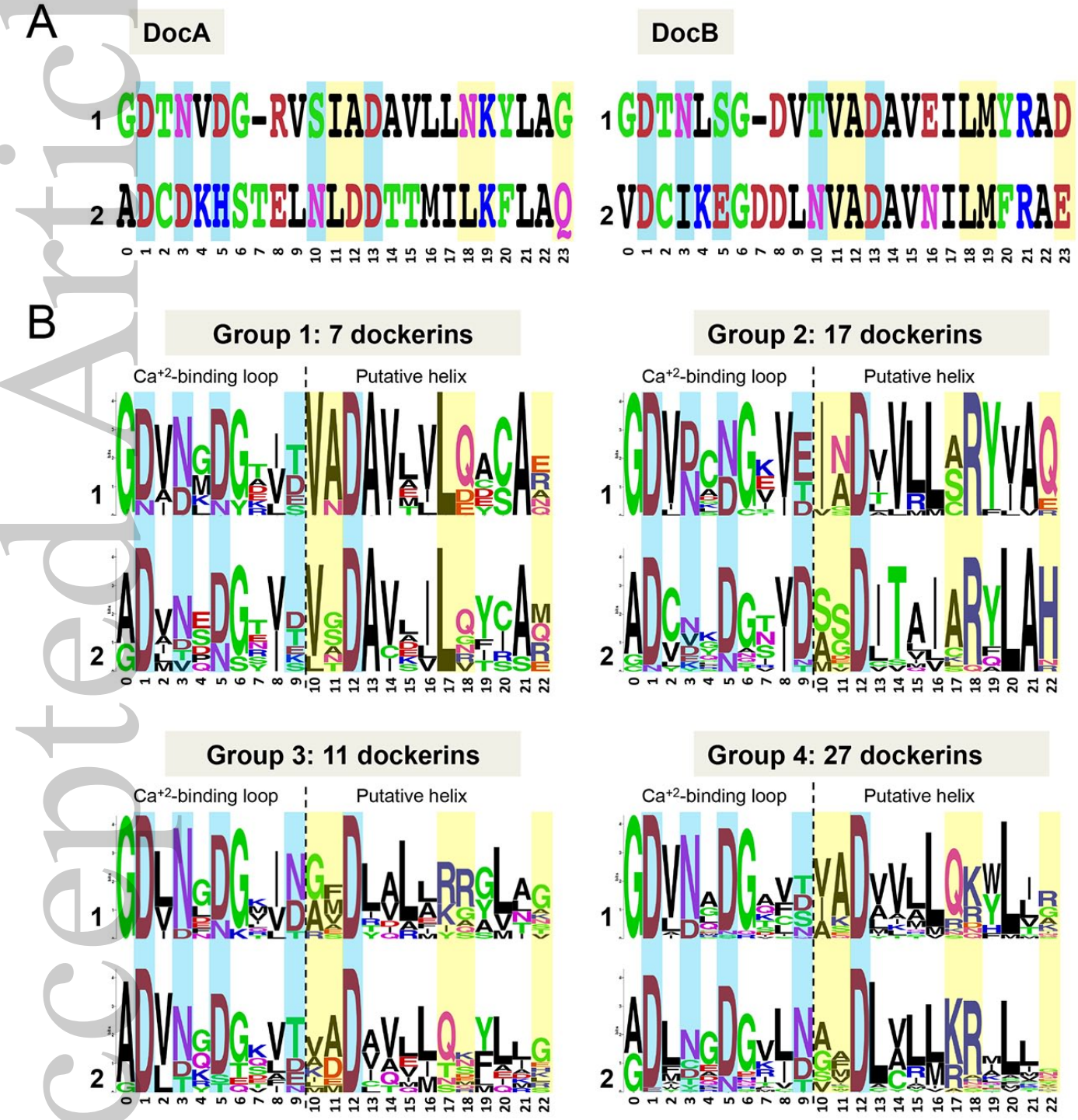
Fig. S4. Sequence alignment of *R. champanellensis* Group 1 dockerins that bind *Rc*-CohJ1 and/or *Rc*-CohE. The box indicates the proposed residues in position 18 of the first duplicated segment that may be involved in the differential binding profiles between *Rc*-Doc3939 and *Rc*-DocJ versus *Rc*-DocH, *Rc*-DocF and *Rc*-DocG. Numbering indicates the residue positions in the two duplicated segments. See Table 1 for id



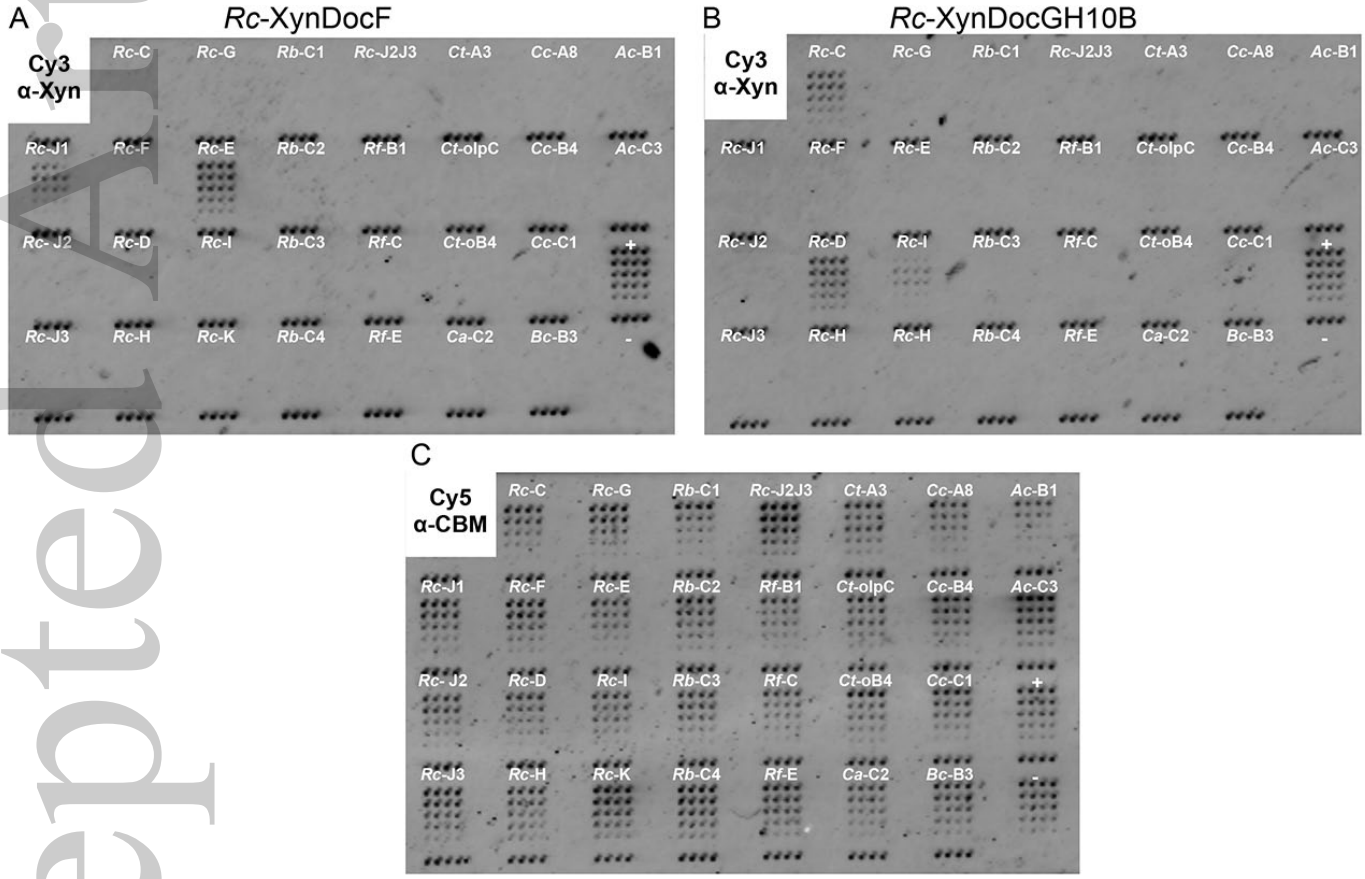
EMI_12868_F1.tif



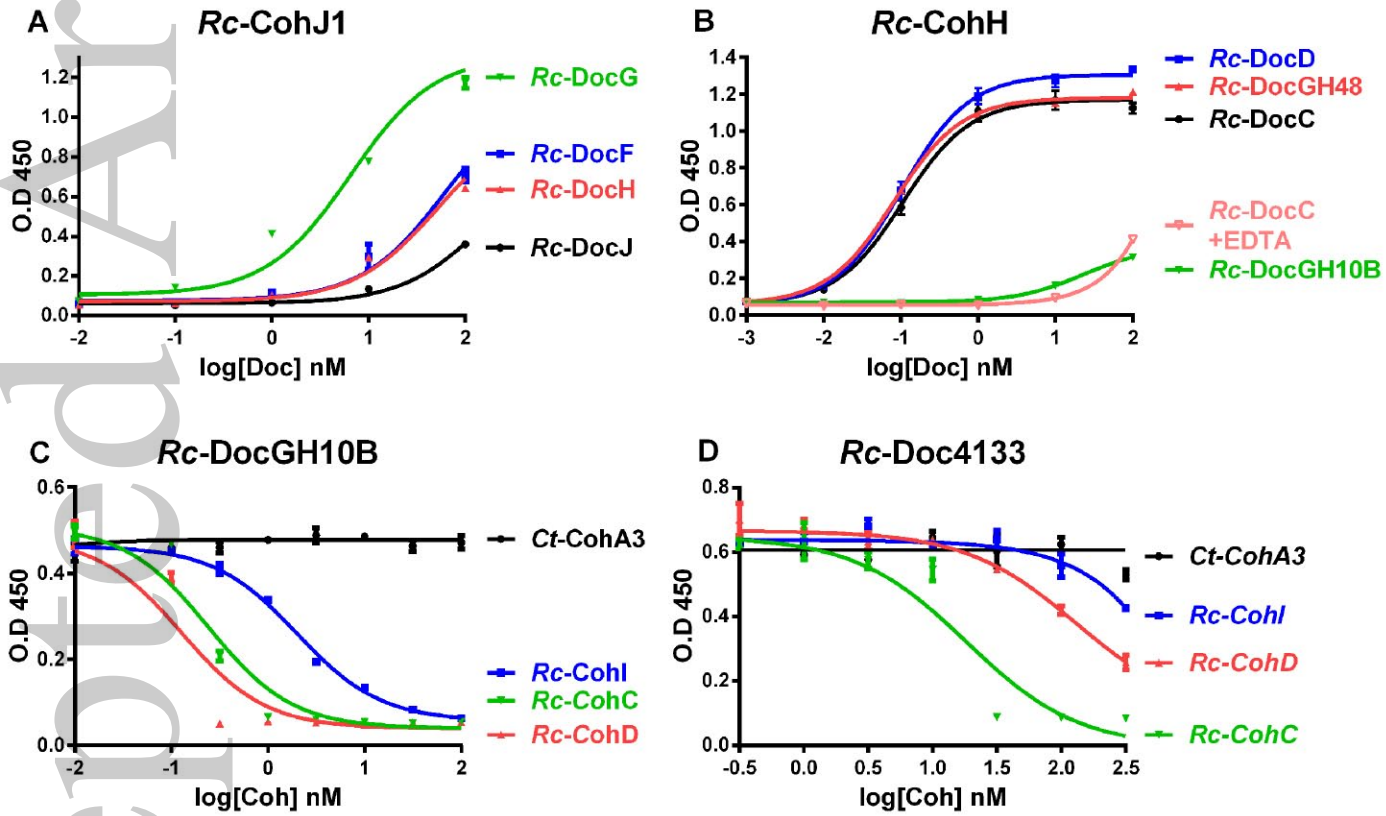
EMI_12868_F2.tif



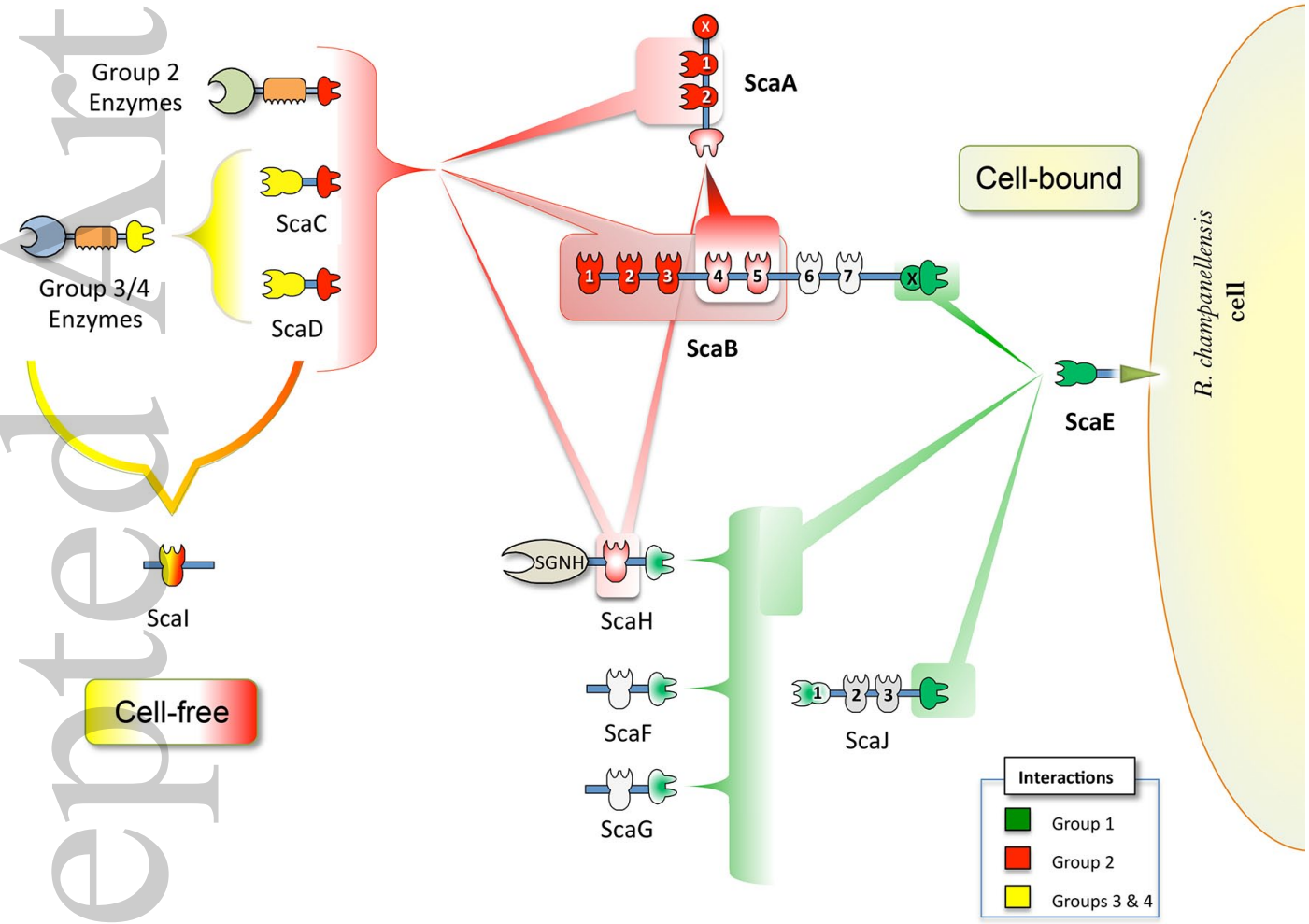
EMI_12868_F3.tif



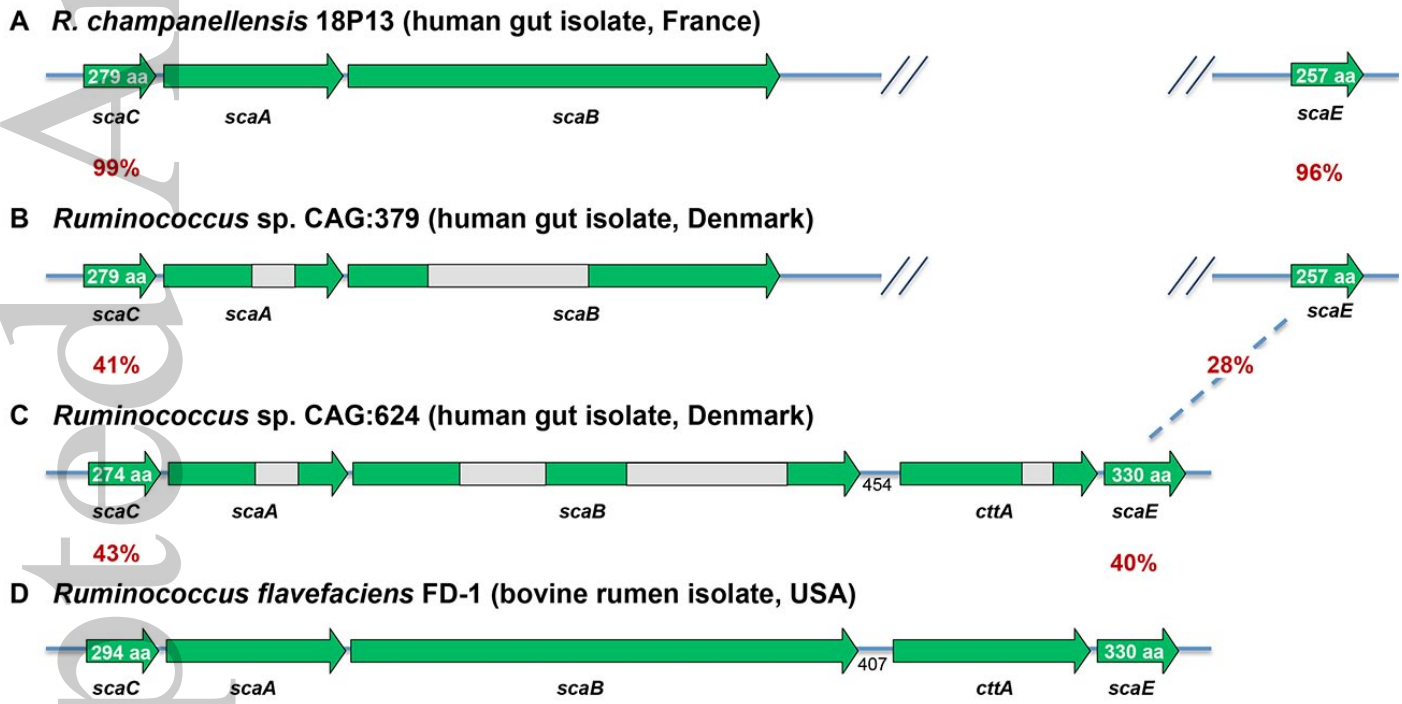
EMI_12868_F4.tif



EMI_12868_F5.tif



EMI_12868_F6.tif



EMI_12868_F7.tif

Table 1. Dockerin-containing proteins of *R. champanellensis.*

GI number	Protein name[a]	Dockerin group	Modular arrangement[b]
KP341766	ScaA		SIGN **X Coh Coh Doc**
KP341766	ScaB		SIGN **Coh Coh Coh Coh Coh Coh Coh X Doc**
291545285	ScaJ	1	SIGN **Coh Coh Coh Doc**
291544538	ScaF	1	SIGN **Coh Doc**
291545095	ScaH	1	SIGN SGNH **Coh Doc**
291545197	ScaG	1	SIGN **Coh Doc**
291543939	3939	1	SIGN FN3 PKD FN3 FN3 FN3 FN3 PKD **Doc**
291543199		1	SIGN Cadherin-like **Doc**
291544999		1	SIGN LRR **Doc**
291543801	ScaC	2	SIGN **Coh** UNK **Doc**
291544607	ScaD	2	SIGN **Coh Doc**
291543938	GH9C	2	SIGN UNK **GH9 CBM3** UNK **Doc**
291543738	GH5B	2	SIGN **GH5 Doc**
291544207	GH48	2	SIGN **GH48** UNK **Doc**
291543186		2	SIGN UNK **Doc**
291543282	GH9A	2	SIGN UNK **GH9 CBM3 Doc**
291543413	GH74	2	SIGN **GH74 Doc**
291543414	GH5A	2	SIGN UNK **GH5** UNK **Doc**
291543470	GH10A	2	SIGN **CBM22 GH10 Doc**
291543699	GH44	2	SIGN **GH44** UNK **Doc**
291544214	PL1/PL9	2	SIGN **PL1 PL9 Doc**
291544445	GH9D	2	SIGN **GH9 Doc**
291544446		2	SIGN UNK **Doc**
291544575	GH9F	2	SIGN UNK **CBM4** UNK **GH9 Doc**
291545037	GH26B	2	SIGN **CBM35** UNK **GH26 Doc**
291545071	GH5C	2	SIGN UNK **GH5 Doc** UNK
291544973	GH98	3	SIGN UNK **GH98 CBM35** UNK X157 **Doc** UNK
291544122	GH43C	3	SIGN **GH43** UNK X19 **CBM22 Doc CE1**
291543994	GH43A	3	SIGN UNK **GH43 CBM61** UNK X157 **Doc**
291544573	GH10B	3	SIGN **CBM22 GH10** UNK **CBM22 Doc** UNK **GH43 CBM6**
291543550	3550	3	TMH **Doc**
291543665		3	SIGN **Doc CBM35** X128
291543673	GH9B	3	SIGN **CBM4** X229 **GH9 Doc GH16**
291543830		3	SIGN SH3 SH3 **Doc**
291544608	PL11	3	SIGN UNK **Doc** UNK **CBM35** UNK **PL11**
291544794	GH30	3	SIGN UNK **GH30 CBM22 Doc** UNK **CE1**

291544870	CE12	3	SIGN FN3 **CE12 CBM13** Doc **CBM35** UNK **CE12**
291545280	GH9G	4	SIGN **GH9 CBM3** UNK Doc
291543899	GH8	4	SIGN UNK **GH8** Doc
291545196	GH11	4	SIGN **GH11** UNK **CBM22** UNK Doc UNK **CBM22 CE4**
291544559	4559	4	SIGN LRR LRR LRR LRR LRR Doc
291544133	4133	4	SIGN DUF187 Doc
291544116	4116	4	SIGN FN3 CotH Doc
291543187	PL11	4	SIGN **PL11 CBM13** X157 Doc
291543191		4	SIGN Doc X259 UNK X259 UNK
291543643		4	SIGN UNK Doc
291543758	PL1	4	SIGN **CBM13 PL1 CBM13 CBM13** Doc
291543946		4	SIGN X134 UNK Doc
291543991	GH43B	4	SIGN **GH43** UNK **CBM13** Doc
291544094		4	SIGN UNK Doc
291544107		4	SIGN LRR LRR LRR LRR Doc
291544109		4	SIGN LRR LRR LRR Doc
291544115		4	SIGN UNK LRR Doc
291544187		4	SIGN UNK LRR Doc
291544250	Lipase	4	SIGN Lipase Doc
291544365	PL1/PL9	4	SIGN Doc **PL1 PL9**
291544405	GH43D	4	SIGN UNK **GH43** UNK **CBM6** Doc
291544406	PL1	4	SIGN UNK **PL1** UNK X157 Doc
291544408	PL1	4	SIGN UNK **PL1** X149 **CBM13** X157 Doc
291544414	Peptidase	4	SIGN Peptidase Doc
291544512	GH26A	4	SIGN **CBM35** UNK **GH26** UNK **CBM35** Doc
291544542	PL1	4	SIGN **CBM13 PL1 CBM13** Doc
291544574	GH9E	4	SIGN UNK **GH9 CBM3** Doc
291544817		4	SIGN UNK Doc

[a]Chosen names for this study.

[b]Abbreviations: SIGN, signal peptide; Doc, dockerin; Coh, cohesin; GH, glycoside hydrolase; CBM, carbohydrate-binding module; PL, polysaccharide lyases; CE, carbohydrate esterases; SGNH, lipases or esterases; FN3, fibronectin type III; PKD, polycystic kidney disease; DUF187, Glycoside hydrolase-like GH101; CotH, spore coat protein H; LRR, leucine-rich repeat; UNK, X, unknown. Selected dockerins for this study are highlighted in green.

Table 2. Cohesin-dockerin interactions in *R. champanellensis*: summary of cellulose microarray experiments. Twenty-four dockerins (rows), including the ScaA and ScaB dockerins and representatives of the four different groups, were checked against 20 cohesins (columns). Each dockerin was examined in a different slide containing all the test cohesins and relevant controls. Interaction intensity (number of pluses) was defined as the number of clearly labeled rows among the five different concentrations (See Supplemental Figure S4 for raw data). The two ScaA cohesins and the seven cohesins of ScaB were tested separately by ELISA tests. Only positive interactions appear in the table. See Table 1 for description of dockerin-bearing proteins that contain CAZy domains. In others, the numbers refer to the last 4 digits of the respective full GI number (i.e., 29154XXXX).

		A1/ A2	A2	B1/B2/ B3	B4	B5/ B6	B6	B7	C	D	E	F	G	H	I	J1	J 2	J 3	K
Group 1	DocF										+++++					+++			
	Doc3939										++++								
	DocG										+++++					++++			
	DocH										+++++					+++			
	DocJ										+++++								
	DocB										++++								
Group 2	DocC	+++++	+++++	+++++	+++++	+++++								+++++	++				
	DocD	+++++	+++++	+++++	++++	++++								+++++	++				
	DocGH48	+++++	+++++	+++++	+++++	+++++								+++++	++				
	DocGH9C	+++++	+++++	+++++	+++++	+++++								+++++	++				
	DocGH5B*	+++++	+++++	+++++	+++++	+++++								+++++					
	DocA				+++++	+++++													
Group 3	DocGH9B								+++++	+++++					++++				
	DocGH10B								++++	+++++					++				
	DocGH43C								+++++	+++++					++				
	DocGH98								+	+++									
	DocGH43A									++									
	Doc3550																		
Group 4	Doc4116								++++	+++++					++				
	Doc4559								+++++	+++++					+				
	Doc4133								+++++	+									
	DocGH11								++++	+++++									
	DocGH9G*								++++	+++					++				
	DocGH8																		

* Tested as an intact wild-type protein (instead of Xyn-Doc chimaera).