Broad phylogeny and functionality of cellulosomal components in the bovine rumen microbiome

Lizi Bensoussan, Sarah Moraïs, Bareket Dassa, Nir Friedman, Bernard Henrissat, Vincent Lombard, Edward Bayer, Itzhak Mizrahi

environmental microbiology

Environmental Microbiology (2017) **19**(1), 185–197 doi:10.1111/1462-2920.13561

Broad phylogeny and functionality of cellulosomal components in the bovine rumen microbiome

Lizi Bensoussan,[1] Sarah Morais,[1] Bareket Dassa,[1] Nir Friedman,[2] Bernard Henrissat,[3] Vincent Lombard,[3] Edward A. Bayer[1] and Itzhak Mizrahi[2]*

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

[2]*The Faculty of Natural Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel.*

[3]*Architecture et Fonction des Macromolécules Biologiques, UMR6098, Aix-Marseille University, CNRS UMR7257, Marseille, France.*

Summary

The cellulosome is an extracellular multi-enzyme complex that is considered one of the most efficient plant cell wall-degrading strategies devised by nature. Its unique modular architecture, achieved by high affinity and specific interaction between protein modules (cohesins and dockerins) enables formation of various enzyme combinations. Extensive research has been dedicated to the mechanistic nature of the cellulosome complex. Nevertheless, little is known regarding its distribution and abundance among microbes in natural plant fibre-rich environments. Here, we explored these questions in bovine rumen microbial communities, specialized in efficient degradation of lignocellulosic plant material. We bioinformatically screened for cellulosomal modules in this complex environment using a previously published ultra-deep fibre-adherent rumen metagenome. Intriguingly, a large portion of the functions of the dockerin-containing proteins were related to alternative biological processes, and not necessarily to the classic fibre degradation function. Our analysis was experimentally validated by characterizing specific interactions between selected cohesins and dockerins and revealed that cellulosome is a more generalized strategy used by diverse bacteria, some of which were not previously associated with cellulosome production. Remarkably, our results provide additional proof of similarity among rumen microbial communities worldwide. This study suggests a broader and widespread role for the cellulosomal machinery in nature.

Received 23 June, 2016; accepted 29 September, 2016. *For correspondence. E-mail: imizrahi@bgu.ac.il; Tel. (+1972)8 647 98 36; Fax (+1972)8 647 79 859.

The copyright line for this article was changed on 16 April 2019 after original online publication

Introduction

The cellulosome is an enzymatic complex produced by anaerobic microorganisms (Bayer *et al.*, 2004; 2007; 2013). It is composed of a set of multi-modular components, partly structural and partly enzymatic, which assemble together to form an elaborate multi-enzyme complex. This complex has been extensively studied in the last several decades and has been shown to be highly efficient in degrading plant cell walls in various environments (Flint *et al.*, 2008). The core of the cellulosome structure is a non-catalytic subunit, termed scaffoldin, composed of one or several 'cohesin' modules that interact specifically with 'dockerin' modules fused to an enzyme that will generally exhibit functions related to fibre deconstruction (Bayer *et al.*, 1998; Shoham *et al.*, 1999). The latter characteristically include glycoside hydrolases (GHs), carbohydrate esterases (CEs) and polysaccharide lyases (PLs), which work synergistically to degrade the intricate lignocellulosic substrate. In addition, in some complexes, such as that of *Clostridium thermocellum*, the most studied cellulosome producer, the scaffoldin subunit contains a carbohydrate-binding module (CBM) which mediates the interaction of the complex with the plant fibres, and an X-dockerin modular dyad that participates in its anchoring to the bacterial cell surface.

The cohesin–dockerin interaction, is one of the strongest protein–protein interactions found in nature and serves to secure the different subunits into the complex (Bayer *et al.*, 1994; 2004; 2013). The interactions tend to be species-specific, meaning that a cohesin of one species will recognize and bind only to the dockerins of the same species, although some cross-species interactions have been observed (Pages *et al.*, 1997; Haimovitz *et al.*, 2008). Generally, cohesins and their complementary dockerins are classified into three main types (I, II and III) according to sequence, and hence structure, similarity. Historically, cohesins of the 'primary' (enzyme-integrating) scaffoldin

were termed type-I-cohesins whereas cohesins of cell-surface anchoring proteins were classified as type-II-cohesins, although in one known case, *Bacteroides (Pseudobacteroides) cellulosolvens*, the types are reversed (Ding *et al.*, 2000; Xu *et al.*, 2004). Distinctive cohesin modules that were found in the rumen bacterium, *Ruminococcus flavefaciens* and its substrains, led to their classification as type-III (Shoham *et al.*, 1999; Ding *et al.*, 2001; Rincon *et al.*, 2003; Bayer *et al.*, 2004; 2013; Haimovitz *et al.*, 2008). Dockerins are short protein modules, about ~70 amino acids long, characterized by a highly conserved duplicated domain of two Ca^{+2}-binding loop-helix motifs, each of about 22 amino acids residues, connected by a linker. These two loops are required by the dockerin module to fold into a stable tertiary structure and crucial for binding with the cohesin in two different configurations, which allows a dual-binding mode of interaction (Carvalho *et al.*, 2003; 2007). In each Ca^{+2}-binding segment, positions 1,3,5,9 and 12 are usually occupied by aspartate (D) or asparagine (N) (Pages *et al.*, 1997; Carvalho *et al.*, 2007).

Despite the deep functional and structural characterization of cellulosomal machineries, little is known about their natural distribution within a given ecosystem. This gap of knowledge stems from the fact that these complexes have been mainly studied in cultivated isolated microorganisms. Hence, a burning question arises regarding their abundance and distribution across different taxonomical lines in a given ecosystem. In this study, we have aimed to address this question within the context of the rumen microbiome. In this anaerobic ecosystem, there is an intricate relationship between multiple microorganisms which supports plant biomass degradation that results in increased microbial protein and fermentation products that are valuable to the host animal. We, therefore, selected the rumen microbiome for its highly complex nature, together with its main functionality of plant cell wall deconstruction (Qi *et al.*, 2010; Mizrahi, 2013; Dassa *et al.*, 2014).

Within the rumen, the plant cell wall is deconstructed by microbes utilizing specialized enzymes to hydrolyze the resident polysaccharides (Flint, 1997; Flint *et al.*, 2008; Mizrahi, 2013). So far, few rumen plant cell wall-degrading bacteria have been identified as primary degraders of plant fibre in this ecosystem. The cellulolytic *Ruminococcus* and *Fibrobacter* species are important cellulose-degrading bacteria that inhabit the rumen and are considered as the most active ones (Qi *et al.*, 2010). Two *Ruminococcus* species have been explored in depth: *Ruminococcus flavefaciens* (strains FD-1, 007c, 17 and others) and *Ruminococcus albus* (strains 7, 8 and SY3), and the studies demonstrated that the genomes of multiple *R. flavefaciens* strains encode numerous interacting scaffoldins, notably ScaA, B, C and E, that are contained in a single gene cluster, as opposed to only one scaffoldin in the *R. albus* strain 7 and SY3 and none in strain 8 (Ding *et al.*, 2001; Rincón *et al.*, 2004; Rincon *et al.*, 2005; Jindou *et al.*, 2006; Miller *et al.*, 2009; Dassa *et al.*, 2014). These organisms occupy only a small fraction of the rumen ecosystem (Jami and Mizrahi, 2012b) and are currently considered to be the only ones that carry cellulosomal components within the rumen environment.

In the present study, advances in sequencing technologies served to examine complex microbial environments, notably the rumen environment, and to gain insight into the taxonomical composition and protein functionality using metagenomics approaches. Hence, we have used an ultra-deep two bovine rumen metagenome database, previously described by Hess *et al.* (2011) as a basis to understand the complex microbial ecosystem of the rumen. We extensively analysed this database to describe the abundance, variability and functionality of cellulosomal elements (i.e., cohesin and dockerin-containing proteins), their resemblance to previously described cellulolytic components, and we examined biochemically their specific interactions.

Materials and methods

Retrieval of dockerin and cohesin-containing sequences from the rumen metagenome

Local BLAST (Altschul *et al.*, 1997) searches were carried out among the contigs which were sequenced from a bovine rumen metagenome (Hess *et al.*, 2011), and sequences of known cohesin and dockerin modules were analysed against an in-house cohesin–dockerin database, which was constructed in the Bayer laboratory. All hits below *E*-value of 10^{-4} were retrieved and inspected individually by examining their characteristic sequence features.

Species and functional characterisation of cohesin/dockerin-bearing proteins

In order to obtain information about the species of the cohesin/dockerin protein sequences we used Blast2Go (Conesa *et al.*, 2005). In the process of characterizing the cohesin- and dockerin-containing sequences, we performed BLAST searches (blastp) of these protein sequences against the NCBI-nr database (non-redundant protein sequences) with *E*-value of 10^{-5}. Prediction of each protein function was carried out by MG-RAST (Meyer *et al.*, 2008), a pipeline which produces functional prediction of gene sequences by comparing it to proteins and nucleotide databases. The default maximum *E*-value cut-off used by MG-RAST was 10^{-5}.

Identification of carbohydrate-degrading enzymes

Dockerin-containing proteins were annotated by the Carbohydrate-Active enZymes (CAZy) database (http://www.cazy.org) (Lombard *et al.*, 2014) in order to bioinformatically analyze their catalytic modules. The CAZy database contains information about enzymes that assemble, modify and

breakdown polysaccharides. This includes identification of the catalytic modules and their classification for glycoside hydrolases (GHs), carbohydrate esterases (CEs), polysaccharide lyases (PLs) and carbohydrate-binding modules (CBM). Additional conserved domains of the proteins were analysed using the NCBI Conserved Domain Database (Marchler-Bauer *et al*., 2014) (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) which serve as a resource for the annotation of proteins with the location of conserved domain footprints.

Identification of signal peptide

The presence of secretion signals for the cohesin- and dockerin-containing proteins were analysed by the SignalP server that uses the neural network and hidden Markov model algorithms (Petersen *et al*., 2011), with the sensitive cut-off values both with gram-positive and gram-negative options.

Multiple sequence alignment and phylogenetic trees

Multiple sequence alignment (MSA) was conducted using the EBI website (http://www.ebi.ac.uk/Tools/msa/) by Clustal Omega and MUSCLE for the cohesins and dockerins respectively. The phylogenetic trees of the cohesins and dockerin were built by MEGA6 (Tamura *et al*., 2013) with the statistical method of maximum likelihood. The phylogeny test was bootstrapped, and the number of bootstrap replications was 100. The phylogenetic trees were visualized by the server Interactive Tree Of Life (iTOL) (Letunic and Bork, 2011) (http://itol.embl.de/). Sequence similarity between identified dockerins was calculated using the server SIAS (http://imed.med.ucm.es/Tools/sias.html).

Cloning of candidate proteins

XynDoc (fusion proteins between *Geobacillus stearothermophilus* T6 xylanase and the target dockerins) and CBM-Coh (fusion proteins between *C. thermocellum* scaffoldin CBM3a and the target cohesins) gene plasmid cassettes were designed for expression of dockerin and cohesin proteins, respectively, as described previously (Barak *et al*., 2005). The test dockerin was inserted into the plasmid between the KpnI and BamHI restriction sites of the pET9 vector (Novagen), and the cohesin between BamHI and XhoI restriction sites of the pET28 vector (Novagen).

Genomic DNA from 10 low-fibre diet bovine rumen samples as described in Jami and Mizrahi (2012a,b) and in Supporting Information Table S3 (30% fibre diet) (Jami and Mizrahi, 2012a) and from five different high-fibre diet bovine rumen samples was purified (Jami and Mizrahi, 2012b). Dockerin- and cohesin-harboring genes, containing desired restriction sites, were amplified from the genomic DNA using specific primers (Supporting Information Table S1). The PCR reaction mixture contained TaKaRa Ex Taq DNA polymerase (0.25 U μl^{-1}) and Ex Taq buffer supplemented with dNTPs (2.5 mM), forward and reverse primers (1 μM) and 10 ng metagenomic DNA. PCR products were purified using PCR purification kit (Qiagen).

Some gene fragments could not be cloned from the metagenomic DNA and were therefore ordered from Integrated DNA Technologies (IDT) and used as templates for PCR reactions.

PCR products and plasmids were digested with the appropriate FastDigest restriction enzymes (KpnI and BamHI for dockerin, BamHI and XhoI for cohesin) (Thermo scientific, Fermentas). The digested DNA fragments (PCR or plasmid) were purified from a DNA gel using a HiYieldTM Gel/PCR Extraction kit (Real Biotech Corporation, RBC, Taiwan). The digested DNA fragments were then ligated into the appropriate linearized vectors using T4 DNA ligase (Fermentas UAB Vilnius, Lithuania), at 4°C overnight.

The cloned dockerin and cohesin sequences are listed in the Supporting Information (Table S2).

Protein expression and purification

E. coli BL21 cells were transformed with the desired plasmid and plated onto kanamycin plates. Five millileter was added to the plate to collect the cells and to resuspend them into 500 ml LB medium (Luria-Bertani), supplemented with 50 $\mu g\ ml^{-1}$ kanamycin (Sigma-Aldrich Chemical Co, St. Louis, Missouri) and 2 mM $CaCl_2$ for dockerin-containing proteins (for proper folding). The cultures were grown at 37°C to $A_{600} \approx 0.8$–1. Induction for protein expression was performed by adding Isopropyl-1-thio-β-D-galactoside (IPTG) (Fermentas) at a final concentration of 0.1 mM, and growth was continued at 16°C for ~16 h. Xylanase-fused dockerins were purified by Ni-NTA batch purification as reported earlier (Caspi, 2006), and CBM-fused cohesins were purified by macroporous-beaded cellulose preswollen gel (IONTOSORB, Usti nad Labem, Czech Republic) as described by Ben David (Ben David, 2015).

The fractions containing relatively pure proteins were pooled and their concentrations were estimated by absorbance at 280 nm. Proteins were stored in 50% (v/v) glycerol at −20°C. Extinction coefficients were determined, based on the known amino acid composition of each protein using ExPASy-ProtParam (Gasteiger *et al*., 2005).

Affinity ELISA binding assay

An ELISA-based procedure was followed to determine cohesin–dockerin specificity of interaction. The procedure was followed as described previously (Barak *et al*., 2005).

Non-denaturing gel electrophoresis

In order to check the extent of a given cohesin–dockerin interaction, non-denaturing gel electrophoresis (Vazana *et al*., 2012) was performed for each pair of cohesin/dockerin fusion proteins. The stoichiometric molar ratio of each pair was calculated. The optimal interaction may form at a ratio somewhat different from the estimated 1:1 ratio. Therefore, incremental ratios were tested around the estimated value of 1, in order to determine the effective stoichiometric enzyme-to-scaffoldin ratio. In a 30-μl reaction (containing 15 μl of wash buffer), equimolar amounts (4–8 μg) of each protein was added. To allow complex formation, the reaction was incubated for 1 h at 37°C. Sample buffer (15 μl, without SDS) was added to 30 μl of the

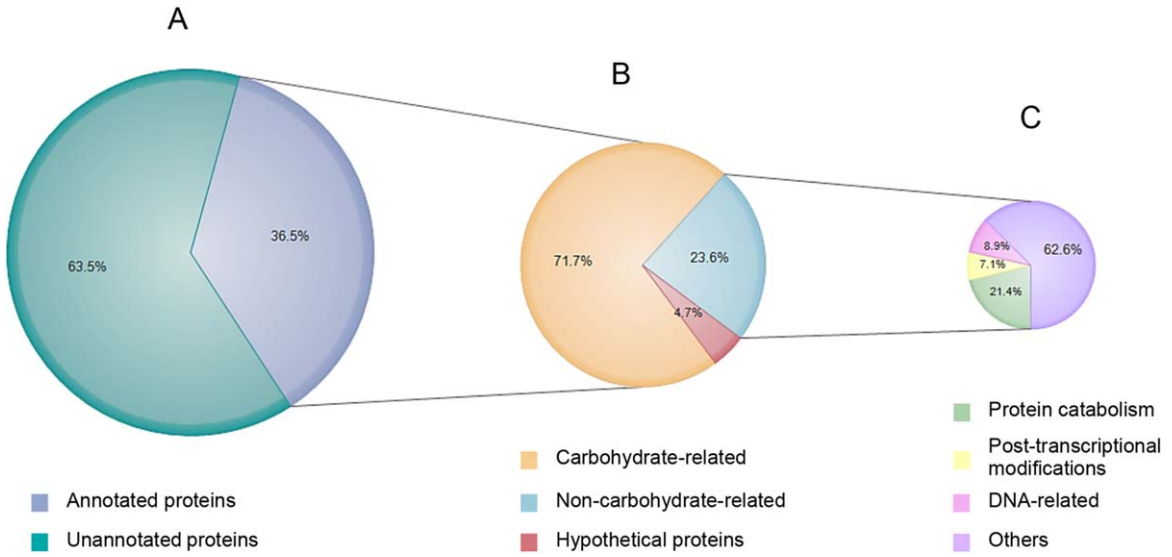

Fig. 1. Functionality of the dockerin-containing proteins found in the metagenome. Pie charts illustrating the distribution of functional categories at the highest level supported by these functional hierarchies. Each slice indicates the percentage of reads with predicted protein functions annotated to the specified category for the given source.
A. Pie chart showing the distribution of the dockerin-containing protein sequences; 237sequences (36.5%) contain predicted proteins with known functions, whereas the functions of 412 sequences (63.5%) remain unknown.
B. Pie chart showing the distribution of the predicted proteins to three distinct functional categories: carbohydrates-related (71.7%), non-carbohydrates-related (23.6%) and hypothetical proteins (4.6%).
C. Pie chart showing the distribution of the predicted proteins classified with functions unrelated to carbohydrates (non-carbohydrates-related) into functions related to protein catabolism (21.4%), post-transcriptional modification (7.1%) DNA-related (8.9%) and others (62.5%).

reaction mixture, and the samples were loaded onto the non-denaturing gels (4.3%-stacking/9%-separating gels).

Results

The role of cellulosomal complexes is not exclusive to plant fibre degradation

In order to identify cohesin-and dockerin-containing proteins in the rumen metagenome, we applied BLAST searches against a database of known dockerins and cohesins, 649 dockerins and 61 cohesins were thus revealed.

Dockerin modules are usually fused to an enzyme that serves a polysaccharide-degrading function within the cellulosomal complex. Therefore, we sought to understand the functional attributes of the dockerin-containing proteins found in the rumen metagenome. The putative dockerin-containing proteins were subjected to annotation using the Carbohydrate-Active enZymes (CAZy) database, in an attempt to identify catalytic modules related to carbohydrate degradation. This database is specific and curated for catalytic modules, such as glycoside hydrolases (GHs), carbohydrate esterases (CEs) or polysaccharide lyases (PLs) and the non-catalytic modules, carbohydrate-binding modules (CBMs). Among the 649 dockerin-containing proteins, 158 (24.3%) contained a variety of predicted carbohydrate-active catalytic units and CBMs. These include 135 GHs from 21 different families, 21 CEs, 5 PLs and 65 CBMs. The list of dockerin-bearing proteins and associated CAZy modules is presented in the Supporting Information (Table S4).

In addition, we submitted the 491 proteins which did not contain carbohydrate-active catalytic units or CBMs according to CAZy, to MG-RAST analysis (Meyer *et al.*, 2008). This server produces automated functional assignments of sequences in the metagenome by comparing them to a large number of protein databases (GenBank, IMG, KEGG, PATRIC, RefSeq, SEED, SwissProt, TrEMBL, eggNOG, M5NR). Our analysis revealed that out of 491 dockerin-containing proteins 16.1% were predicted as proteins of known function with a total of 79 different functions (Supporting Information Table S5). Interestingly, only 12 annotated proteins had carbohydrate-related enzymatic functions, 56 proteins were annotated with functions that are not associated with lignocellulosic degradation and 11 were annotated as hypothetical proteins. Figure 1A and B summarizes the functions obtained by both the CAZY and MG-RAST servers.

Among the 56 proteins that were annotated with functions not related to carbohydrate degradation (Fig. 1C), 12 functions out of 47 (21.4%) were related to protein catabolism. This finding suggests a wider role of cellulosomal complexes in macromolecule degradation and scavenging, and the dockerin-bearing proteins thus appear not to be

exclusive to plant fibre degradation. Interestingly, two predicted peptidases were annotated as D-alanyl-D-alanine carboxypeptidases.

Some of the proteins (7 in total) were annotated with both a carbohydrate-related function by CAZY and non-carbohydrate-related function by MG-RAST (Supporting Information Table S6), suggesting two putative functions for the same protein. Blast analysis of these sequences confirmed that six of these proteins contain two putative functions, whereas in the seventh protein the Blast analysis confirmed only the CAZy annotation with about 300 amino acids at the C-terminus of the protein remaining unannotated.

None of the 61 cohesin-containing proteins were annotated with predicted functions using the same databases, which is in accordance with the fact that cohesins, and therefore scaffoldins, are non-enzymatic subunits of the cellulosomes.

Since the dockerin-containing proteins are supposed to be secreted extracellularly, we further analysed our dataset for the existence of a secretion signal with the SignalP server. Out of the overall dataset, only 55% of the proteins contained a signal peptide. The absence of signal peptide in 45% of the proteins could be the result of performance problems of the detection method or truncated 5′ ends of the ORFs in the metagenome. As for the cohesin-bearing proteins, signal peptides were detected in 65% of the cases. In order to assess whether the average distribution of signal peptide-containing proteins is different from that of the dockerin- and cohesin-containing proteins, we further calculated the distribution of signal peptides on proteins residing in contigs that contained the identified cohesin modules. Of the 4,636 ORFs existing on these contigs, only 10.3% (479 proteins) were predicted to have signal peptides. This difference further strengthens our findings regarding the high prevalence of signal peptides in the dockerin- and cohesin-containing proteins.

Cellulosome strategy is not exclusive to a specific phylogeny but scattered over large phylogenetic space

Phylogenetic trees were generated for the cohesin and dockerin sequences. They were used together with Blast2Go to characterize the taxonomical composition of the bacterial origin encoding these cellulosomal components in the rumen. The protein sequences were annotated using BLAST according to the best hit (lowest *E*-value) and supplemented with 70% bootstrap support of maximum likelihood trees (as described in the material and methods section). A large number of the newly identified cohesins and dockerins belonged to the *Ruminococcus* genus and were similar to those of *R. flavefaciens*. We also observed taxonomical diversity within the cohesin- and dockerin-containing proteins, which extended beyond the *Ruminococcus* genus (Fig. 2). Like the *Ruminococcus* genus, these genera were mostly members of the *Firmicutes* phylum, including *Roseburia*, *Clostridium*, as well as other microbes from this phylum which eluded precise annotation. We also identified dockerin-containing proteins from the *Bacteroidetes* phylum (8.8%), specifically for the *Prevotella* (7.2%) and *Bacteroides* (1.6%) genera that were not documented before. As can be seen in Fig. 2B, we observed 2 events of lateral gene transfer between the Bacteroidetes and Firmicutes phyla, associated together on the same clades (with bootstrap confidence of a minimum of 70%), this would suggest lateral gene transfer of dockerin modules from the Firmicutes to the Bacteroidetes phyla. In addition, to further confirm these modules as proper dockerins, they were aligned with known dockerins and the predicted structures of some of these dockerins were obtained using the Swiss model (Biasini *et al.*, 2014) (Supporting Information Figure 1S). The predicted structures, the alignment and the known positions of the calcium-coordinating and recognition residues of the aligned dockerins are all consistent with the established type I dockerins. No significant species hits were found for 62 dockerins and 2 cohesins. These findings indicate that the cellulosomal strategy is adopted by a surprisingly wide range of phylogenetic groups and by a number of uncharacterized bacteria.

Usually more than one cohesin was discovered on a scaffoldin module. Indeed, when we investigated the 61 presumed cohesin-containing proteins, they were distributed among 40 different putative scaffoldins. Some of the cohesin-containing scaffoldins were found to carry also dockerin modules. Phylogenetic analysis of the cohesin-containing proteins with a database of known cohesins indicated that some of these scaffoldins were closely related to known scaffoldins from *R. flavefaciens* and shared similar molecular arrangement (Fig. 3).

We asked further whether the taxonomical composition of the rumen cellulosomal components is similar in terms of abundance to that of the fibre-adherent rumen microbiome. To answer this question, we compared the rumen cellulosomal components with the 16SrRNA data of the two bovines described by Hess *et al*. (2011). The analysis of the rumen microbiome composition of these samples revealed a different distribution for the cohesins and dockerins compared to that of the rumen microbiome (Table 2). In this analysis, 41–43% of the predicted ribosomal RNA genes were assigned to the *Bacteroidetes* phylum and 19–21% to the *Firmicutes* phylum, whereas the phylum distribution of cohesins and dockerins was 1–9% assigned to *Bacteroidetes* phylum and 70–75% for *Firmicutes* phylum, belonging mainly to *R. flavefaciens*, a well-characterized cellulosome-producing species (Table 2 and Supporting Information Figure 2S). This finding confirms that cellulosomal components are prevalent in *Firmicutes* phylum as opposed to the Bacteroidetes and that the taxonomic

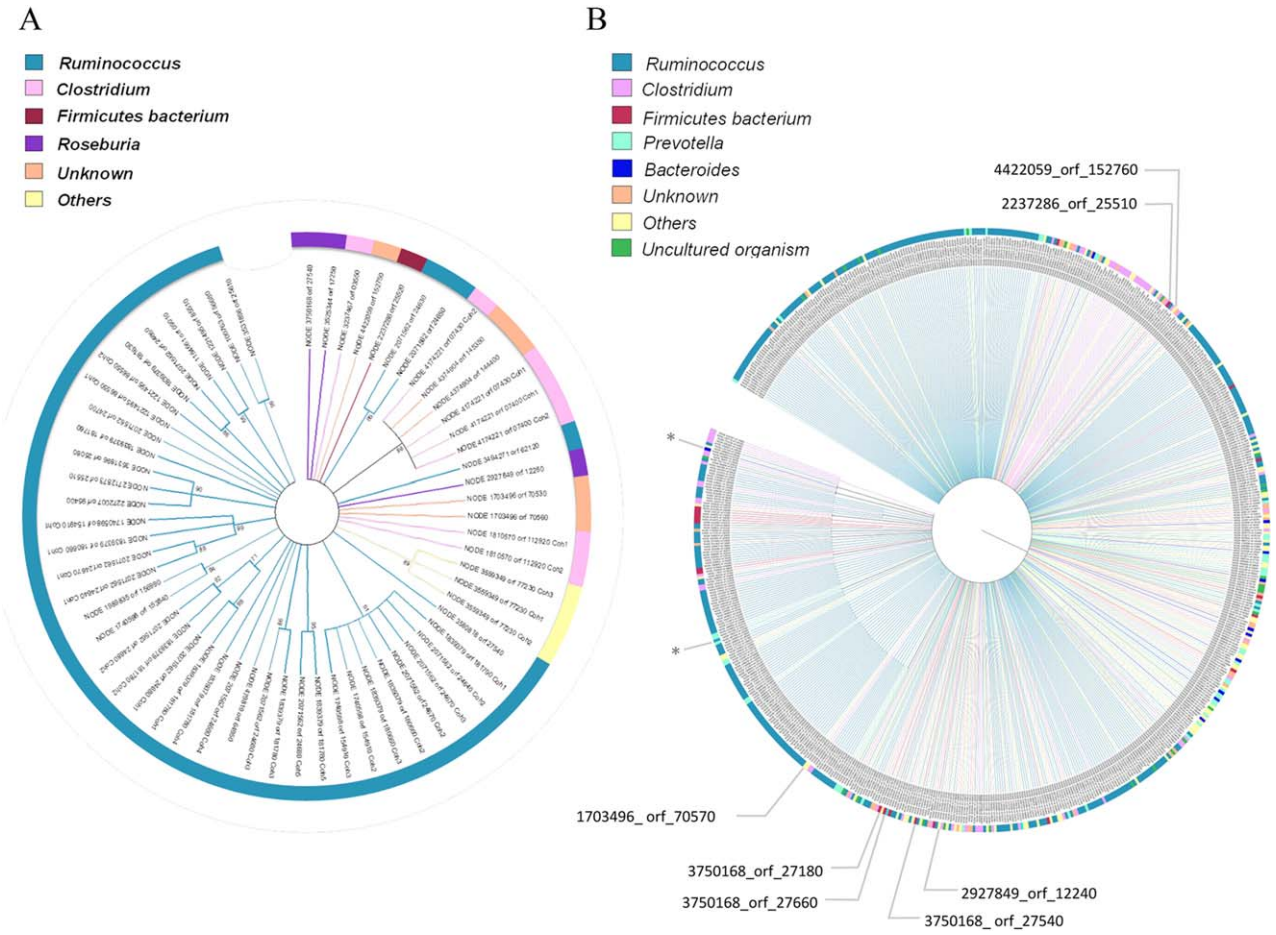

Fig. 2. Taxonomic diversity of cellulosome-related modules.
A. Phylogeny of the retrieved cohesins. A set of 61 cohesin modules was aligned using the CLUSTAL-Omega tool in the EBI-website, which served to reconstruct a phylogenetic tree by MEGA6.0 software. The cohesins distributed into four major putative species by BLAST comparison of the cohesin modules, and the rest were defined as unknown or less prevalent species that were grouped as 'others'. The displayed tree was condensed whereby each branch with less than 70% statistical significance was collapsed, and the numerical values above the nodes indicate the bootstrap percentiles. The putative species are coloured according to the legend.
B. Phylogeny of the retrieved dockerins. A set of 649 dockerin modules was aligned using the MUSCLE tool in the EBI-website, and the alignment served to reconstruct a phylogenetic tree by MEGA6.0 software. The dockerins distributed into five major putative species by BLAST comparison of the dockerin modules, and rest were defined as unknown, uncultured bacteria unknown or less prevalent species that were grouped as 'others'. The displayed tree was condensed in which each branch with less than 70% statistical significance was collapsed. The putative species are coloured according the legend. The seven selected dockerins are shown with their full annotations (contig and ORF numbers). The asterisk indicates possible lateral gene transfer events between Firmicutes and Bacteroidetes phyla.

distribution of cellulosomal components is different from the taxonomic distribution of the rumen microbiome. This would be anticipated from the known sources of cohesins and dockerins in both cellulosome-producing species as well as non-producers (Peer *et al*., 2009).

Finally, the 61 putative cohesin sequences were aligned with those from the database of known cohesins to examine sequence conservation. The metagenome-derived cohesin sequences were classified according to two major types (Supporting Information Figure 3S). Eight cohesin modules classified as type I and 23 as type III. The rest, 30 (49.18%) cohesin modules, could not be classified, and no significant similarity to type II cohesins was found. The unclassified cohesins exhibited similarity in their length and sequences to known cohesins. However, when the tree was collapsed by 70% statistical significance (Supporting Information Figure 3S), and no significant branching with any of the types was observed. This analysis shows that the sequence variability of the cohesins in the rumen is high and the type classification is much more diverse than previously considered.

Interaction studies of cohesins and dockerins

Cohesin–dockerin interactions are essential for the cellulosomal machinery to assemble. We, therefore, sought to

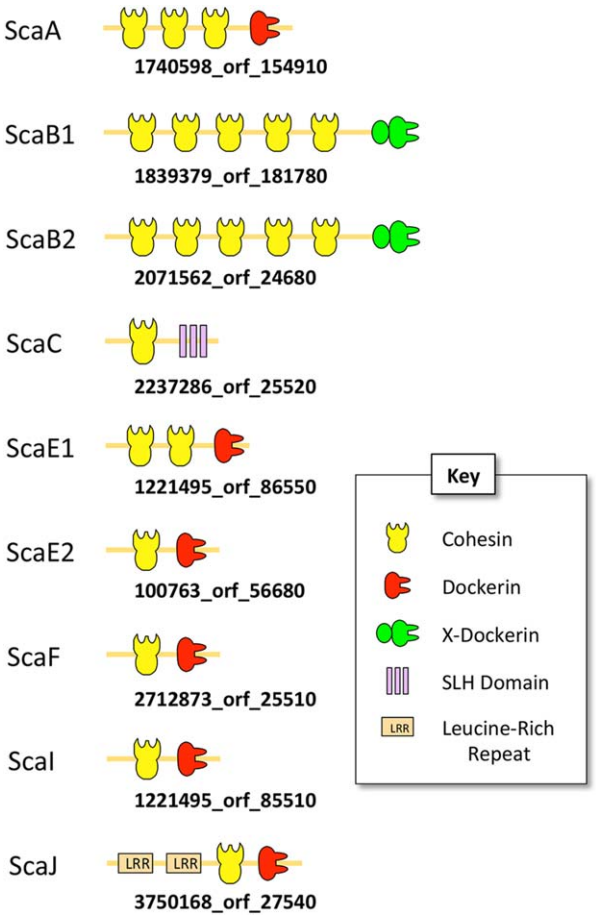

Fig. 3. Molecular arrangement of putative scaffoldins.
Nine putative scaffoldins were identified bioinformatically, and the phylogenetic tree of its cohesins against known cohesins was used to determine to which known scaffoldins they could be associated [the names of scaffoldins ScaA, ScaB refer to their resemblance to cohesins of the cellulosomal system of *Ruminococcus flavefaciens* (Ding *et al*., 2001)].

examine whether we could locate cohesin and dockerin modules among the identified putative cellulosomal components which would actively interact with each other in a specific manner. In addition, we wanted to determine whether these interactions persist in rumen microbiomes of animals from different breeds and geographies. We specifically challenged our hypothesis that the cellulosomal strategy is widely used by a wide range of taxa. In this context, we focused on components whose taxonomical association belonged to the less explored phylogenies. As cohesin–dockerin interactions tend to be species-specific (Pages *et al*., 1997), we selected candidate cellulosomal components residing on the same contig, assuming that they are coded by the same genome and may thus interact together. We found 18 such contigs that carry cohesin and dockerin modules. Their phylogenetic association was determined by taking into account the phylogenetic association of all their carried genes (Table 1), five of these contigs that belonged to the less explored phylogenies where chosen for further analysis. These are herein termed contigs 1 to 5, and Fig. 4A illustrates the modular arrangement of the selected contigs. Six dockerins and six cohesins from these contigs were designed for cloning and activity assays. Although the amino acid sequence similarity between them ranges from 12% to 66%, all dockerins shared known common patterns of dockerin-related residues, such as repeated Ca^{2+}-binding loop α-helices, putative recognition residues that mediate binding to the cohesin. Most of the selected sequences also show a conserved Gly residue, adjacent to the initial calcium-binding aspartate (D) immediately prior to the putative Ca^{2+}-binding loop, with the exception of one dockerin (Dockerin 5c) which possesses an arginine (R) at that position (Supporting Information Figure 4S).

As our analysis was performed on fibre-adherent rumen metagenomes from animals fed on switchgrass in the United States, we examined whether the cellulosomal proteins will also be present in rumen microbiomes of animals fed on a different diet and from a different geography. We, therefore, assessed the prevalence of these proteins in the rumen microbiome of 15 animals that were situated in Israel and fed on two different rations. Ten bovines were fed on a 25% fibre diet and five on a 95% fibre diet (Supporting Information Table S5). None of the 12 examined cellulosomal proteins could be identified in the rumen microbiomes of the animals that were fed on the 25% fibre diet but seven proteins (1a, 2c, 3b, 4a, 4b, 5aC, 5c) were identified in the rumen microbiomes of the animals that were fed on 95% fibre diet, which is a similar diet to that on which the animals in the United States were fed.

We next sought to examine the functionality of the identified proteins. We succeeded in cloning the seven identified genes directly from the metagenome, whereas the five gene fragments that could not be cloned in the rumen were therefore synthetized (1b, 2a, 2b, 3a, 5b). One gene (5aD) was amplified from the metagenomic DNA, but its sequence was different than the one expected, and this gene was therefore excluded from further experimentation.

The five chosen dockerins and the six cohesin modules were expressed with two different cassettes: each dockerin module was fused to the C terminus of xylanase T6, and the product was termed Xyn-Doc; and each cohesin module was N-terminally fused to a CBM3a module, and the product termed CBM-Coh. After purification of the 11 proteins, we used an affinity-based ELISA approach that takes advantage of the complementary Xyn- and CBM-fused modules (Barak *et al*., 2005) to measure their potential interactions.

All six cohesins were thus examined for their interaction with dockerins 1b, 2a, 3b, 4b and 5c. According to the ELISA results, cohesin 4a exhibited significant interaction

Table 1. Contigs which contain both cohesins and dockerins.

Contig number	No. of dockerins	No. of cohesins	Putative species
NODE_100763	2	1	*Ruminococcus*
NODE_1184661	8	1	*Ruminococcus*
NODE_1221495	10	3	*Ruminococcus*
NODE_1703496	**1**	**2**	***Firmicutes bacterium***
NODE_1740598	23	4	*Ruminococcus*
NODE_1810570	1	2	*Clostridia (Faecalibacterium prausnitzii)*
NODE_1839379	17	11	*Ruminococcus*
NODE_1989835	39	1	*Ruminococcus*
NODE_2071562	6	14	*Ruminococcus*
NODE_2237286	**1**	**1**	***Firmicutes bacterium***
NODE_2272007	6	1	*Ruminococcus flavefaciens*
NODE_2712873	25	1	*Ruminococcus*
NODE_2927849	**1**	**1**	***Clostridium***
NODE_3525344	2	1	*Ruminococcaceae bacterium*
NODE_3531896	10	2	*Ruminococcus*
NODE_3750168	**3**	**1**	***Firmicutes bacterium***
NODE_4422059	**1**	**1**	***Phylum Firmicutes***
NODE_479819	1	1	*Butyrivibrio*

Eighteen contigs containing cellulosomal proteins, with both dockerin and cohesin modules have been found. The total number of cohesin and dockerin modules along with the putative species for each contig, based on BLAST searches of the full contig, is shown. The selected contigs for cloning are in boldface font, highlighted in purple.

with dockerin 4b and cohesin 3a exhibited significant interaction with dockerin 3a and dockerin 4b (Fig. 4B and Supporting Information Figure 5S). In addition, cohesin 2b also showed somewhat weaker interaction with dockerin 4b (Fig. 4B). Figure 4B presents the interactions between the modules according to ELISA results. When verifying

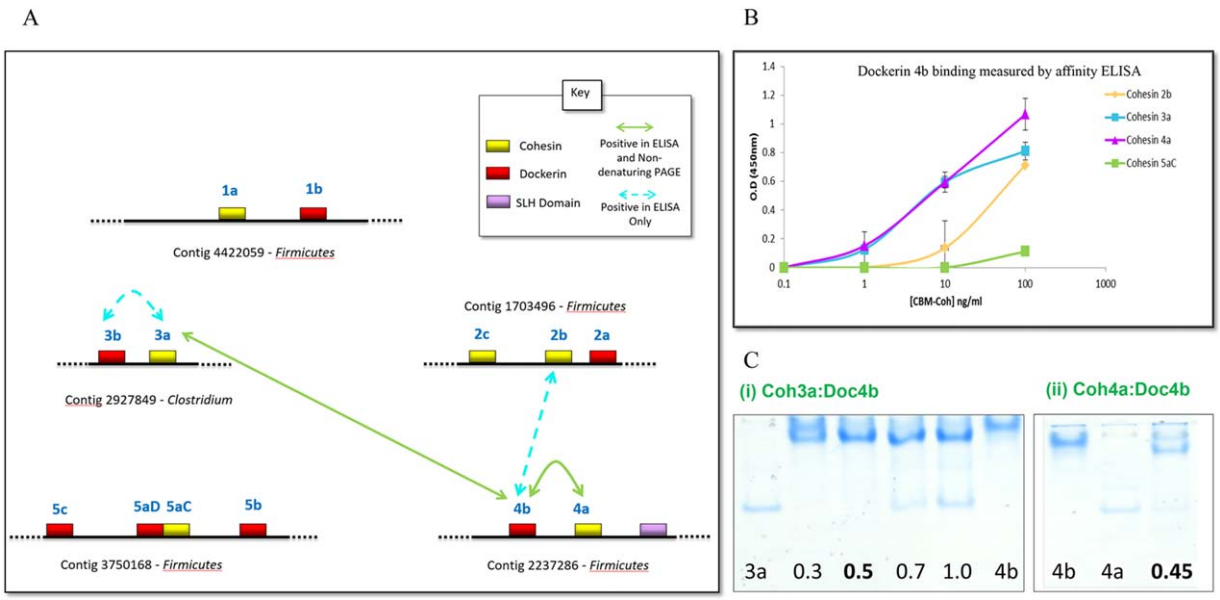

Fig. 4. A. Identification of specific interactions between cohesin and dockerin modules from the five chosen contigs. The molecular arrangement of the contigs is shown. Positive interactions for both in the ELISA assay and the native gel are designated by green arrows, and positive interactions observed only in the ELISA assay are shown in dashed cyan-coloured arrows.
B. Cohesin-dockerin binding measured by ELISA. ELISA plates were coated with the desired Xyn-Doc proteins. Positive interactions of dockerin 4b with cohesins 2b, 3a and 4a were observed. For comparison, none of the dockerins in the graph showed interaction with cohesin 5aC. Error bars indicate SD from the mean of duplicate samples from one experiment.
C. Non-denaturing gels confirming positive interactions between dockerin 4b and (i) cohesin 3a and (ii) cohesin 4a that showed positive interaction by ELISA. The numbers represent the predetermined molar-ratio estimates of the cohesin and dockerin modules. The experimentally determined optimal ratio of the interaction is given in boldface.

the results on non-denaturing PAGE, only the interactions between dockerin 4b with cohesin 3a and cohesin 4a were confirmed (Fig. 4C and Supporting Information Figure 6S). This difference may perhaps be due to the more sensitive nature of the ELISA technique.

Intriguingly, dockerin 4b interacted strongly with two cohesins located on two different contigs, which raises the speculation whether they originated from the same species, as we know that the cohesin–dockerin interaction is generally species-specific. The genomic GC-content of bacteria varies from less than 20% to more than 70% (Hildebrand *et al*., 2010). Similar genomic GC content between contigs might imply the same bacterial species of origin. The GC content of these two contigs are 58.4% and 56.9% for contig 3 and contig 4, respectively, thereby suggesting that these two contigs could be associated with the same bacterial species. For comparison, the GC content of the genome of *R. flavefaciens* strain AE3010 is 46% and 44.6% for strain FD1.

Discussion

The sophisticated and efficient nature of the cellulosomal machinery and its primary role in plant fibre degradation raises the question of whether it is also associated with other microbial extracellular functions and whether it is restricted to previously isolated and characterized microbes. Here, we sought to explore these questions in the rumen ecosystem by analysing ultra-deep fibre-adherent metagenome data, previously published by Hess *et al*. (2011).

This ecosystem houses a very efficient fibre-degrading microbiome (Kruger Ben Shabat *et al*., 2016), and some of the well-characterized cellulosome-producing bacteria were isolated therein (Dassa *et al*., 2014). Exploring metagenomes allowed us to more broadly assess these questions, since the sampled material provides access to genomes of uncultivated microbes as well. In this context, our analysis uncovered 649 dockerins and 61 cohesins that were related to known sequences in the metagenome of the bovine rumen. We discovered that 71.7% of the annotated sequences were related to carbohydrates and the remainder (23.6%) was annotated as non-carbohydrate-related. In those cases, the dockerin modules were found to be attached to a variety of enzymes that are inconsistent with classical cellulosome action involving fibre degradation. Large proportions of these proteins were assigned to protein catabolism functions, suggesting a potential role for cellulosomal machinery in extracellular protein degradation. Some of these enzymes were assigned to functions with a potential role in microbial interaction, such as dockerin-containing proteins annotated as D-alanyl-D-alanine carboxypeptidase (Supporting Information Table S5). D-Alanyl-D-alanine peptide is an important structural component of the bacterial cell wall, usually targeted by various antibiotics. This could suggest a defense or attack mechanism for the cellulosomal machinery in the complex rumen environment that could be involved in microbial competition. This hypothesis is further strengthened by the finding of dockerin modules associated with catalytic modules that encode putative lysozyme activity (Supporting Information Table S4 and Table S5) which could target the sugar bonds of the bacterial cell wall peptidoglycan. These findings imply broad involvement of dockerin modules in a wide variety of different biological and cellular processes that are unrelated to classical cellulosome function.

In terms of phylogeny, a large number of the newly discovered metagenome-derived cohesins and dockerins were similar to those of *R. flavefaciens*. Nevertheless, we observed a remarkable diversity of predicted species and many unknown proteins from ostensibly uncultured bacteria (Fig. 2). This finding would reflect the complexity of the rumen microbiome as well as the diversity of the bacterial species that contain cellulosomal genes. The taxonomic composition of the examined fibre-adherent microbiome fraction compared to that of the known cellulosomal components indicated contrasting species distribution and a clear enrichment of the cellulosome-containing bacteria in the predicted Firmicutes phylum compared to the Bacteroidetes phylum that was underrepresented (Table 2).

The approach followed in the present work relied on the major assumption that important functional components of the microbiome would be preserved across geography and time (Jindou *et al*., 2008; Jami and Mizrahi, 2012a; Henderson *et al*., 2014; Weimer, 2015). We, therefore, assumed that the same sequences from data collected in the United States would exist in Israeli bovines. In this context, we succeeded to clone from Israeli bovines genes identical to those identified in the metagenomic data collected from the bovines in the United States (Hess *et al*., 2011). Four new pairs of selected cohesins and dockerins that originated from uncharacterized microbes were thus expressed and are clearly different from the previously characterized rumen bacterium *R. flavefaciens*. These results support the hypothesis that bovines around the world share similar bacterial communities and similar cellulosome-producing groups of bacteria. The results further strengthen with the hypothesis that these cellulosome-containing microorganisms fill an important function in the rumen ecosystem and occupy a specific ecological niche (Jindou *et al*., 2008; Grinberg *et al*., 2015).

To assess the possibility of dietary influence on the abundance of cellulolytic bacteria in the current study, we used two groups of bovine rumen samples that differed in diet formulation given to the animals. Most genes could

Table 2. Phylogenetic distribution in % according to the 16S rRNA genes in the metagenomes of the two bovines, or the cellulosomal components that were therein identified (ND: not detected).

	16S RNA			
Phyla	Bovine 1	Bovine 2	Cohesins	Dockerins
Firmicutes	21.9	19.6	70.8	74.5
Bacteroidetes	41.1	43.3	1.7	8.8
Actinobacteria	0.7	3.1	3.4	3.1
Arthropoda	3.2	5.2	5.2	0.9
Proteobacteria	2	0.9	ND	1.9
Apicomplexa	ND	ND	1.7	0.5
Ascomycota	ND	ND	5.2	0.1
Chordata	ND	ND	1.7	0.3
Cnidaria	ND	ND	1.7	0.3
Euryarchaeota	ND	ND	1.7	2.4
Fibrobacteres	4.3	3.8	ND	ND
Spirochaetes	2	4.3	ND	ND
Tenericutes	1.5	1.8	ND	ND
unclassified (derived from Eukaryota)	4.6	4.7	3.4	3.2
unclassified (derived from Viruses)	0.02	ND	1.7	0.3
unclassified (derived from Bacteria)	14.5	7.5	ND	ND
unclassified (derived from unclassified sequences)	2.4	4.3	ND	1.2
Others	1.78	1.5	1.8	2.5

only be amplified from rumen microbiomes originating from high-fibre diet and not low-fibre diet (Supporting Information Table S3). This phenomenon further emphasizes the cardinal role of diet in determining not only the composition of the microbiome but also its functionality.

Our findings provide evidence that the complementary cellulosomal components revealed in this study interact and are potentially involved in broad and diverse extracellular functions in the rumen. It still remains to be determined how these enzymes carry out their functions and how they serve microbial physiology in a specified ecological niche. By understanding the microbial interactions in the bovine rumen, by investigating to what extent and abundance the cellulosomal machinery is being exploited by rumen microorganisms, and by exploring the relative activity of novel rumen cellulosomal components to known ones, we can provide fundamental knowledge for biotechnological application. In addition, the novel cohesin–dockerin pairs identified in this study extend our knowledge of cellulosome organisation in different species and could enrich the cohesin–dockerin repertoire for production of designer cellulosomes that are used as a powerful tool for understanding cellulosome action and for studying synergism among enzymes in this supramolecular complex (Fierobe *et al.*, 2005; Caspi *et al.*, 2009; Moraïs *et al.*, 2012; Arfi *et al.*, 2014).

The broader and generalized role of the cellulosomal machinery as revealed in this study opens new prospects for further research for its usage in environments which are not rich in fibre but require extracellular enzymatic assemblies. As in nature, the cellulosomal machinery can be harnessed for biotechnological applications other than lignocellulose degradation.

Acknowledgements

The research described in this communication was supported by grants from the Israel Science Foundation (No. 1313/13 to I.M. and 1349/13 to E.A.B.), by the Israeli Chief Scientist Ministry of Agriculture and Rural Development Fund (No. 362-0426) to E.A.B. and I.M., by the Israeli Chief Scientist Ministry of Science Foundation (No. 3-10880) to E.A.B. and I.M., by the European Research Council under the European Union's Horizon 2020 research and innovation program, project number 640384 to I.M. and E.A.B., by the United States-Israel Binational Science Foundation (BSF), Jerusalem, Israel and by the establishment of an Israeli Center of Research Excellence (I-CORE Center No. 152/11, to E.A.B.) managed by the ISF.

Conflict of Interest

The authors declare no conflict of interest

References

Altschul, S.F., Madden, T.L., Schäffer, 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.

Arfi, Y., Shamshoum, M., Rogachev, I., Peleg, Y., and Bayer, E.A. (2014) Integration of bacterial lytic polysaccharide monooxygenases into designer cellulosomes promotes enhanced cellulose degradation. *Proc Natl Acad Sci USA* **111:** 9109–9114.

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. *J Mol Recognit* **18:** 491–501.

Bayer, E.A., Belaich, J.P., Shoham, Y., and Lamed, R. (2004) The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annu Rev Microbiol* **58:** 521–554.

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., Morag, E., and Lamed, R. (1994) The cellulosome—a treasure-trove for biotechnology. *Trends Biotechnol* **12:** 379–386.

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

Bayer, E.A., Shoham, Y., and Lamed, R. (2013) Lignocellulose-decomposing bacteria and their enzyme systems. In *The Prokaryotes*. Springer, pp. 215–266.

Ben David, Y., Dassa, B., Borovok, I., Lamed, R., Koropatkin, N.M., Martens, E.C., *et al*. (2015) Ruminococcal cellulosome systems from rumen to human. *Environ Microbiol* **17:** 3407–3426.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., *et al*. (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res* **42:** W252–W258.

Caspi, J., Irwin, D., Lamed, R., Shoham, Y., Fierobe, H.P., Wilson, D.B., and Bayer, E.A.. (2006) *Thermobifida fusca* family-6 cellulases as potential designer cellulosome components. *Biocatal Biotransform* **24:** 3–12.

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

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 USA* **100:** 13809–13814.

Caspi, J., Barak, Y., Haimovitz, R., Irwin, D., Lamed, R., Wilson, D.B., and Bayer, E.A. (2009) Effect of linker length and dockerin position on conversion of a *Thermobifida fusca* endoglucanase to the cellulosomal mode. *Appl Environ Microbiol* **75:** 7335–7342.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21:** 3674–3676.

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.

Ding, S.Y., Bayer, E.A., Steiner, D., Shoham, Y., and Lamed, R. (2000) A scaffoldin of the *Bacteroides cellulosolvens* cellulosome that contains 11 type II cohesins. *J Bacteriol* **182:** 4915–4925.

Ding, S.Y., Rincon, M.T., Lamed, R., Martin, J.C., McCrae, S.I., Aurilia, V., *et al*. (2001) Cellulosomal scaffoldin-like proteins from *Ruminococcus flavefaciens*. *J Bacteriol* **183:** 1945–1953.

Fierobe, H.P., Mingardon, F., Mechaly, A., Bélaïch, A., Rincon, M.T., Pagès, S., *et al*. (2005) Action of designer cellulosomes on homogeneous versus complex substrates. *J Biol Chem* **280:** 16325–16334.

Flint, H.J. (1997) The rumen microbial ecosystem—some recent developments. *Trends Microbiol* **5:** 483–488.

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.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., and Bairoch, A. (2005) Protein identification and analysis tools on the ExPASy server. In The Proteomics Protocols Handbook. Walker, J.M. (ed). Totowa, NJ 612 Humana Press, pp. 571–607.

Grinberg, I.R., Yin, G., Borovok, I., Miller, M.E.B., Yeoman, C.J., Dassa, B., *et al*. (2015) Functional phylotyping approach for assessing intraspecific diversity of *Ruminococcus albus* within the rumen microbiome. *FEMS Microbiol Lett* **362:** 1–10.

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.

Henderson, G., Cox, F., Ganesh, S., Jonker, A., Young, W., and Janssen, P. (2014) Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Sci Rep* **5:** 14567–14567.

Hess, M., Sczyrba, A., Egan, R., Kim, T.W., Chokhawala, H., Schroth, G., *et al*. (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331:** 463–467.

Hildebrand, F., Meyer, A., and Eyre-Walker, A. (2010) Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet* **6:** e1001107.

Jami, E., and Mizrahi, I. (2012a) Composition and similarity of bovine rumen microbiota across individual animals. *PLoS One* **7:** e33306.

Jami, E., and Mizrahi, I. (2012b) Similarity of the ruminal bacteria across individual lactating cows. *Anaerobe* **18:** 338–343.

Jindou, S., Borovok, I., Rincon, M.T., Flint, H.J., Antonopoulos, D.A., Berg, M.E., *et al*. (2006) Conservation and divergence in cellulosome architecture between two strains of *Ruminococcus flavefaciens*. *J Bacteriol* **188:** 7971–7976.

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.

Kruger Ben Shabat, S., Sasson, G., Doron-Faigenboim, A., Durman, T., Yaacoby, S., Berg Miller, M.E., *et al*. (2016) Specific microbiome-dependent mechanisms underlie the energy harvest efficiency of ruminants. *ISME J* [In press]. doi: 10.1038/ismej.2016.62

Letunic, I., and Bork, P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39:** W475–W478.

Lombard, V., Ramulu, H.G., Drula, E., Coutinho, P.M., and Henrissat, B. (2014) The carbohydrate-active enzymes

database (CAZy) in 2013. *Nucleic Acids Res* **42:** D490–D495.
Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., *et al*. (2014) CDD: NCBI's conserved domain database. *Nucleic Acids Res* **43:** D222–D2226.
Meyer, F., Paarmann, D., D'souza, M., Olson, R., Glass, E.M., Kubal, M., *et al*. (2008) The metagenomics RAST server-a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform* **9:** 386.
Miller, M.B., Antonopoulos, D.A., Rincon, M.T., Band, M., Bari, A., Akraiko, T., *et al*. (2009) Diversity and strain specificity of plant cell wall degrading enzymes revealed by the draft genome of *Ruminococcus flavefaciens* FD-1. *PLoS One* **4:** e6650–e6650.
Mizrahi, I. (2013) Rumen symbioses. In *The Prokaryotes*. Springer: Berlin Heidelberg, pp. 533–544.
Moraïs, S., Barak, Y., Lamed, R., Wilson, D.B., Xu, Q., Himmel, M.E., and Bayer, E.A. (2012) Paradigmatic status of an endo- and exoglucanase and its effect on crystalline cellulose degradation. *Biotechnol Biofuels* **5:** 78.
Pages, S., Belaich, A., Belaich, J.P., Morag, E., Lamed, R., Shoham, Y., and Bayer, E.A. (1997) Species-specificity of the cohesin-dockerin interaction between *Clostridium thermocellum* and *Clostridium cellulolyticum*: prediction of specificity determinants of the dockerin domain. *Proteins Struc Funct Genet* **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.
Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8:** 785–786.
Qi, M., Jakober, K., and McAllister, T. (2010) Rumen microbiology. In *Animal and Plant Productivity. Encyclopedia of Life Support Systems*. Oxford: EOLSS, pp. 161–176.
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., Ding, S.Y., McCrae, S.I., Martin, J.C., Aurilia, V., Lamed, R., *et al*. (2003) Novel organization and divergent dockerin specificities in the cellulosome system of *Ruminococcus flavefaciens*. *J Bacteriol* **185:** 703–713.
Rincón, 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.
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.
Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* **30:** 2725–2729.
Vazana, Y., Moraïs, S., Barak, Y., Lamed, R., and Bayer, E.A. (2012) Designer cellulosomes for enhanced hydrolysis of cellulosic substrates. *Methods Enzymol* **510:** 429–452.
Weimer, P.J. (2015) Redundancy, resilience, and host specificity of the ruminal microbiota: implications for engineering improved ruminal fermentations. *Front Microbiol* **6:** 296.
Xu, Q., Bayer, E.A., Goldman, M., Kenig, R., Shoham, Y., and Lamed, R. (2004) Architecture of the *Bacteroides cellulosolvens* cellulosome: description of a cell surface-anchoring scaffoldin and a family 48 cellulase. *J Bacteriol* **186:** 968–977.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. 1S. Predicted structures of selected putative dockerins (A) NODE_1919510_orf_160580 (predicted as GH53-DOC1), (B) NODE_3656756_orf_139070 (predicted as DOC), (C) NODE_3702081_orf_56400 (predicted as DOC), (D) NODE_4135119_orf_128410 (predicted as DOC), (E) NODE_1110409_orf_44820 (predicted as Streptopain precursor/peptidase C10 family protein-DOC), (F) NODE_3589230_orf_87890 (predicted as DOC), (G) NODE_2608643_orf_97770 (predicted as DOC) and (H) NODE_3565304_orf_106370 (predicted as DOC) obtained using the Swiss model (Biasini *et al*., 2014). (A), (B), (C) and (D) were phylogenetically associated with the *Bacteroides* genera and (E), (F), (G) and (H) were associated with the *Prevotella* genera. The structures were modeled on dockerin type I (2ccl), which was one of the selected templates. Calcium ions are represented in pink in A an E. Clustal Omega alignment of these and additional putative dockerins against known type 1 dockerins is presented on the following pages. Established type-1 dockerins whose 3D structures from *Clostridium thermocellum* XynY(10B) (PDB Code 2CCL) and Cel48S (PDB Code 2MTE), *Clostridium cellulolyticum* Cel5A (PDB Code 2VN5) and *Acetivibrio cellulolyticus* ScaB (PDB Code 4UYP) were used as standards in the alignment. Calcium-coordinating residues (dockerin positions 1, 3, 5, 9 and 12) are highlighted in cyan, and reputed recognition residues (dockerin positions 10, 11, 17, 18 and 22) are highlighted in yellow. Some of the putative dockerins are abridged at the C terminus, owing to incomplete sequencing.
Fig. 2S. Ratio of taxonomical composition between the rumen cellulosomal components to the fibre adherent rumen microbiome. The ratios within a common phylum were calculated and represented as blue bars, displayed in relation to a ratio of 1.0.
Fig. 3S. Predicted phylogeny and types of the retrieved cohesins. A set of 61 cohesin modules was aligned using the CLUSTAL-Omega tool in the EBI-website, which served to reconstruct a phylogenetic tree by MEGA6.0 software. The cohesins distributed into 2 major cohesin types, I and III (pink and purple respectively). Thirty cohesins had no significant similarity to known cohesin types, and therefore are marked as unknown (green). The displayed tree was condensed in which each branch with less than 70% statistical significance was collapsed, and the numerical values above the nodes indicate the bootstrap percentiles.
Fig. 4S. Sequence alignment of dockerins from the chosen contigs for cloning. Sequence alignment was performed using the CLUSTALW program. Consensus residues are coloured; Ca^{2+}-binding residues are highlighted in cyan, and putative recognition residues are highlighted in yellow.

Residues are numbered relative to the highly conserved glycine (designated 0), which is positioned adjacent to the initial calcium-binding (residue 1).

Fig. 5S. Cohesin-dockerin binding measured by ELISA. ELISA plates were coated with the desired Xyn-Doc proteins. Cohesin 3a exhibited positive interaction with dockerin 3b and not cohesin 5aC. The absorbance was measured at 450 nm using a tunable microplate reader. Error bars indicate SD from the mean of duplicate samples from one experiment.

Fig. 6S. Two non-denaturing gels showing negative interaction between (A) dockerin 3b and cohesin 3a and (B) dockerin 4b and cohesin 2b. The numbers represent the predetermined molar-ratio estimates of the cohesin and dockerin modules.

Table S1: List of primers. Cloned protein names and designed primers. Restriction sites appear in red upper case fonts. Each protein includes its name based on the contig number (1 to 5) and the original name (in parentheses) as provided in the metagenome database.

Table S2: Five selected contigs containing one or more cohesins and dockerins. The contigs are numbered from 1 to 5 and the names of the cellulosomal components on each of them start with this number. The letters after the numbers distinguish between the cohesin and dockerins. Where relevant, the sequence of the N-terminal signal peptide appears in yellow font. Putative calcium-binding residues (D,N,S in positions 1, 3, 5, 9 and 12 of the duplicated calcium-binding loop) of the dockerin are highlighted in cyan. Putative cohesin-binding residues are highlighted in yellow.

Table S3: Bovine's diet compositions given to the two groups. The amount in kg of each ingredient and its percentages from the total composition is shown.

Table S4: CAZy modules of the identified dockerin-containing proteins

Table S5: Protein annotations as determined by MGRAST server

Table S6: Proteins annotated with a carbohydrate-related function in CAZy and non-carbohydrate-related function in MGRAST