nature microbiology

LETTERS

PUBLISHED: 30 MAY 2017 | VOLUME: 2 | ARTICLE NUMBER: 17087

A parts list for fungal cellulosomes revealed by comparative genomics

Charles H. Haitjema[1], Sean P. Gilmore[1], John K. Henske[1], Kevin V. Solomon[1†], Randall de Groot[1], Alan Kuo[2], Stephen J. Mondo[2], Asaf A. Salamov[2], Kurt LaButti[2], Zhiying Zhao[2], Jennifer Chiniquy[2], Kerrie Barry[2], Heather M. Brewer[3], Samuel O. Purvine[3], Aaron T. Wright[4], Matthieu Hainaut[5,6], Brigitte Boxma[7], Theo van Alen[7†], Johannes H. P. Hackstein[7], Bernard Henrissat[5,6,8], Scott E. Baker[3], Igor V. Grigoriev[2,9] and Michelle A. O'Malley[1*]

Cellulosomes are large, multiprotein complexes that tether plant biomass-degrading enzymes together for improved hydrolysis[1]. These complexes were first described in anaerobic bacteria, where species-specific dockerin domains mediate the assembly of enzymes onto cohesin motifs interspersed within protein scaffolds[1]. The versatile protein assembly mechanism conferred by the bacterial cohesin–dockerin interaction is now a standard design principle for synthetic biology[2,3]. For decades, analogous structures have been reported in anaerobic fungi, which are known to assemble by sequence-divergent non-catalytic dockerin domains (NCDDs)[4]. However, the components, modular assembly mechanism and functional role of fungal cellulosomes remain unknown[5,6]. Here, we describe a comprehensive set of proteins critical to fungal cellulosome assembly, including conserved scaffolding proteins unique to the Neocallimastigomycota. High-quality genomes of the anaerobic fungi *Anaeromyces robustus*, *Neocallimastix californiae* and *Piromyces finnis* were assembled with long-read, single-molecule technology. Genomic analysis coupled with proteomic validation revealed an average of 312 NCDD-containing proteins per fungal strain, which were overwhelmingly carbohydrate active enzymes (CAZymes), with 95 large fungal scaffoldins identified across four genera that bind to NCDDs. Fungal dockerin and scaffoldin domains have no similarity to their bacterial counterparts, yet several catalytic domains originated via horizontal gene transfer with gut bacteria. However, the biocatalytic activity of anaerobic fungal cellulosomes is expanded by the inclusion of GH3, GH6 and GH45 enzymes. These findings suggest that the fungal cellulosome is an evolutionarily chimaeric structure—an independently evolved fungal complex that co-opted useful activities from bacterial neighbours within the gut microbiome.

The release of fermentable sugars from lignocellulose is a major bottleneck in the development of sustainable chemicals from plant biomass and agricultural waste. It has recently been uncovered that early branching anaerobic fungi, which are known degraders of lignocellulose, encode the largest number of biomass-degrading enzyme transcripts yet found in nature[7]. The vast majority of these enzymes carry non-catalytic dockerin domains (NCDDs), which mediate assembly into large multiprotein complexes, or cellulosomes[7]. Cellulosomes were first described in anaerobic bacteria, and the modular interaction scheme native to these complexes has rapidly been exploited for biomass conversion and enzyme tethering applications[8–10]. However, since the first report of cellulosome-like structures in anaerobic fungi more than twenty years ago[11], the identification of dockerin-binding cohesins, or scaffoldins that mediate assembly, has yet to be accomplished due to a lack of genomic data, functional proteomics and characterized strains.

Genomic analysis of five unique anaerobic fungi revealed the presence of almost 1,600 total dockerin domain proteins (DDPs), proteins that contain NCDDs, across genera (Supplementary Table 1) with diverse functionality, primarily related to plant carbohydrate binding and biomass degradation (Fig. 1a,b and Supplementary Table 2). These include 15 glycoside hydrolase (GH) families, 5 distinct carbohydrate-binding domains and other functions implicated in plant cell wall modification and deconstruction including pectin-modifying enzymes and expansins (Supplementary Table 2). Approximately 20% of DDPs belong to spore coat protein CotH and are also present in bacterial cellulosomes. Their role in cellulosomes remains uncertain, but they have been speculated to be involved in plant cell wall binding[12]. Conversely, about 13% represent additional GH activities that are not present in bacterial cellulosomes (GH3, GH6 and GH45). The additional β-glucosidase conferred by GH3, in particular, enables fungal cellulosomes to convert cellulose directly to fermentable monosaccharides, whereas Clostridial cellulosomes produce low-molecular-weight oligosaccharides[13].

To find structural proteins that mediate the assembly of DDPs, we isolated the supernatant and cellulosome fractions from three of these isolates growing on reed canary grass as a sole carbon substrate. Size-exclusion chromatography (SEC) of the cellulosome fraction showed complex formation well within the MDa range (Supplementary Fig. 1), and SDS–PAGE analysis revealed the

[1]Department of Chemical Engineering, University of California, Santa Barbara, California 93106, USA. [2]US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA. [3]Environmental Molecular Sciences Laboratory, Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington 99354, USA. [4]Biological Sciences Division, Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, Washington 99354, USA. [5]Architecture et Fonction des Macromolécules Biologiques, Centre National de la Recherche Scientifique, Aix-Marseille Université, 13288 Marseille, France. [6]INRA, USC 1408 AFMB, Marseille, France. [7]Department of Evolutionary Microbiology, Radboud University, 6525 AJ Nijmegen, The Netherlands. [8]Department of Biological Sciences, King Abdulaziz University, 23218 Jeddah, Saudi Arabia. [9]Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA. [†]Present address: Agricultural and Biological Engineering, Purdue University, West Lafayette, Indiana 47907, USA (K.V.S.); Department of Microbiology, Faculty of Science, Radboud University, PO Box 9010, 6500 GL Nijmegen, The Netherlands (T.v.A.). *e-mail: momalley@engineering.ucsb.edu

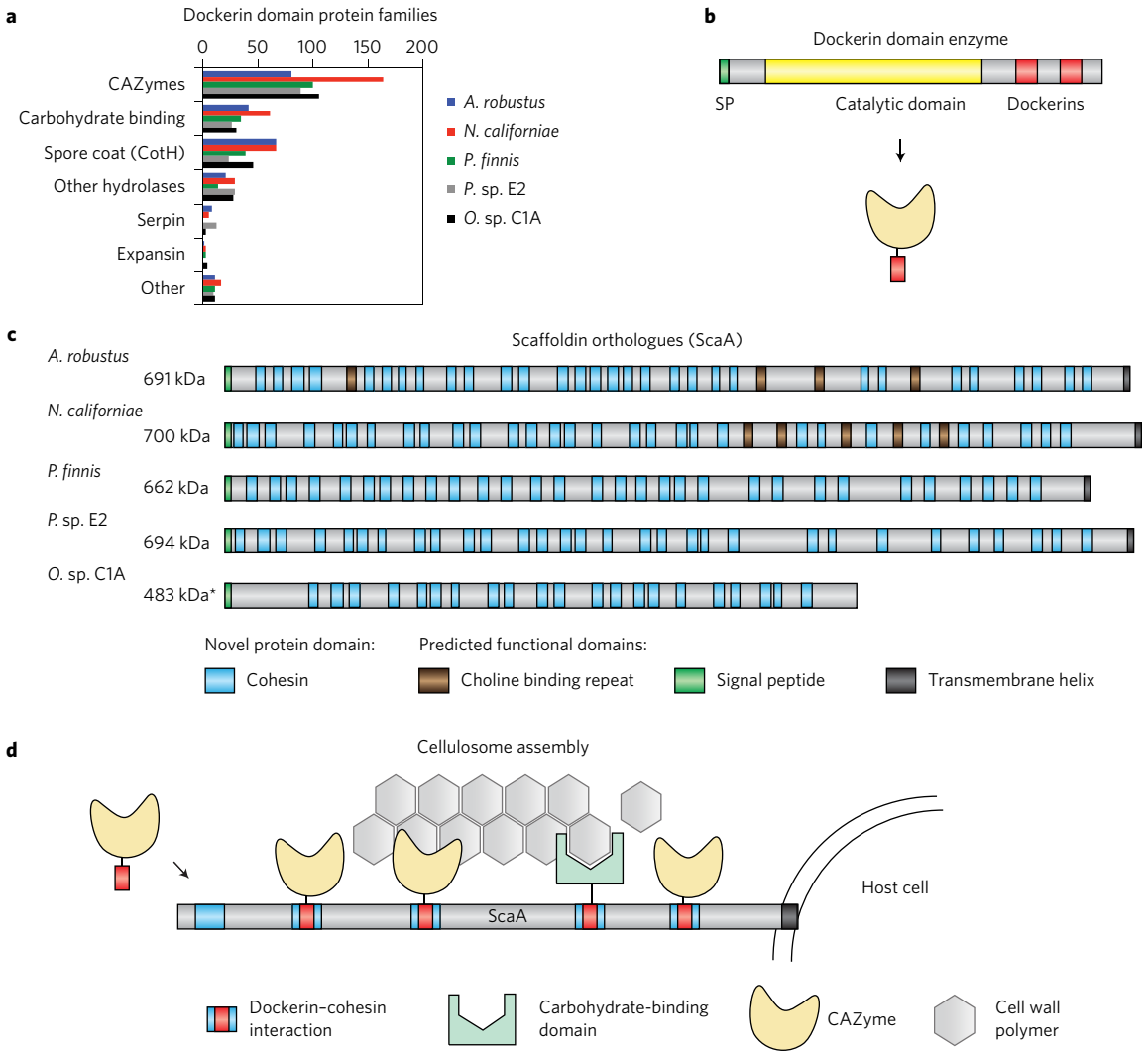

Figure 1 | Overview of gut fungal cellulosome components. a, Protein functional domains associated with non-catalytic dockerin domains. **b**, Schematic of a typical dockerin domain glycoside hydrolase (GH). SP, signal peptide. **c**, Schematic of large non-catalytic scaffoldin molecules in the cellulosomes of gut fungi. The predicted functional domains were determined by InterProScan 5. The extracellular domains (shaded grey) are decorated with interspersed repeating cohesin motifs (blue). The predicted N-terminal signal sequence and C-terminal membrane helix are in green and black, respectively. **d**, Cartoon model of gut fungal cellulosome assembly. *The *O.* sp. C1A *ScaA* gene model sequence is incomplete due to fragmented genome assembly.

presence of many glycosylated proteins (Supplementary Fig. 2). Each fraction was subjected to tandem mass spectrometry, and peptide sequences were mapped to their respective genomic and transcriptomic databases[7]. Many of the proteins associated with these complexes were identified as GHs and other plant cell wall degrading enzymes (Supplementary Table 3). Proteins found in the cellulosome fraction were particularly enriched with NCDDs, indicating modular complex formation. Unexpectedly, all fractions also contained very large uncharacterized proteins (hereafter named ScaA) with molecular weights (M_w) of ~700 kDa (Fig. 1c). These ScaA proteins share 32% sequence identity over at least 92% sequence length (E-value = 0.0) between fungal genera. ScaA orthologues were also detected in the only other sequenced gut fungal genomes, *Piromyces* sp. E2 (hereafter *P.* sp. E2) and *Orpinomyces* sp. C1A (hereafter *O* sp. C1A)[14], although the orthologue detected in *O* sp. C1A was incomplete, probably due to fragmented genome assembly (Fig. 1c).

Sequence analysis of ScaA proteins across all five sequenced genomes showed a predicted N-terminal signal sequence, followed by a large extracellular repeat-rich domain, and ending with a C-terminal membrane anchor (Fig. 1c). Some ScaA proteins also encode predicted choline binding repeats (CBRs), which are known to bind glucan in prokaryotic glucosyltransferases[15]. Thus, one possibility is that CBRs help mediate fungal cellulosome assembly, as many cellulosome proteins are glycosylated (Supplementary Fig. 2). Closer examination of the sequences revealed the presence of a repeating amino-acid sequence motif that is conserved among all ScaA orthologues and that occurs many times throughout these proteins (Fig. 2). This motif is 20–30 amino acids long, typically includes a Gly residue immediately followed by two large hydrophobic residues (most often Tyr residues) and two non-consecutive downstream Cys residues (Fig. 2).

Because ScaA proteins and orthologues are highly represented in secretome and cellulosome fractions from these diverse species of gut fungi, we postulated that they share a common role in these systems, possibly in DDP assembly. We hypothesized that ScaA proteins function as scaffolds, where the repeating motifs act as dockerin-binding cohesins. To investigate this, we recombinantly expressed fragments of the ScaA homologues in *Escherichia coli* and performed enzyme-linked immunosorbent assay (ELISA) using purified dockerin and anti-dockerin chemiluminescent secondary antibody. These results showed a strong dockerin

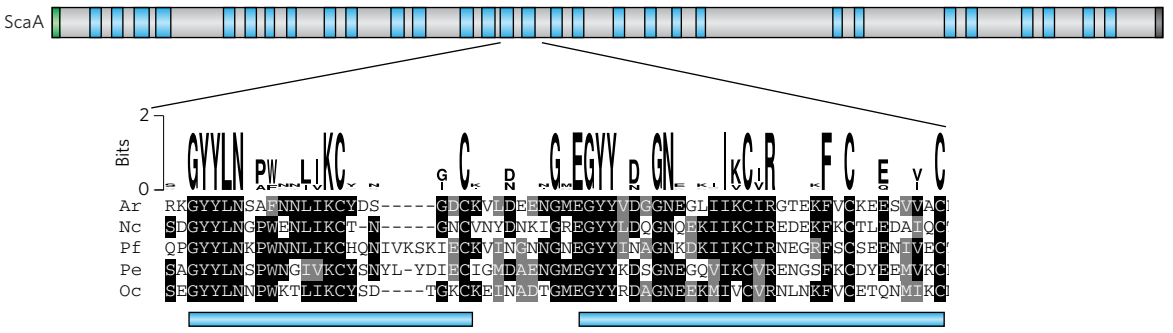

Figure 2 | Sequence analysis of repeating cohesin motifs. Multiple sequence alignment of the five ScaA orthologues reveals a repeating motif. A sequence logo plot that scores the conserved residues is overlaid with the alignment. Amino acid residues delimiting the repeating motif are highlighted by blue rectangles. Ar, *A. robustus*; Nc, *N. californiae*; Pf, *P. finnis*; Pe, *P.* sp. E2; Oc, *O.* sp. C1A.

binding signal in wells containing the scaffoldin fragment cells compared to those containing the empty vector control (Fig. 3a–c). As an additional control, a phenylalanine substitution to dockerin residue W28, previously identified to be critical for binding[16], showed significantly reduced binding activity (Fig. 3c). To determine the binding affinity of the dockerin–ScaA interaction, we purified *Piromyces* ScaA fragments and performed equilibrium analysis by surface plasmon resonance (SPR) against purified fungal dockerin. This analysis revealed that a single dockerin domain interacts with the scaffoldin fragment with an approximate dissociation constant (K_d) of 944 nM and a maximum response (R_{max}) of 91 RU (Fig. 3g; RU, response units). Additionally, the W28F dockerin mutant showed significantly reduced binding affinity (K_d = 3,466 nM, R_{max} = 61 RU). These binding affinities are very similar to those determined by an ELISA against purified native cellulosome (Fig. 3h), where the single dockerin binds with an affinity of 1,230 nM. Taken together, these results suggest that fungal scaffoldin proteins probably mediate the assembly of DDPs in fungal cellulosomes, as depicted by the schematic in Fig. 1d.

Although limited, previous studies have shown that fungal cellulosomes are quite divergent from their bacterial counterparts. For example, NCDDs occur as tandem repeats at the N and/or C terminus, with the most common form being a double tandem repeat (that is, double dockerin) at the C terminus (Supplementary Fig. 4). Although the functional role of this motif repetition is not known, it has been previously noted that double dockerins bind to native cellulosomes more efficiently than single domains[5]. Thus, we hypothesized that increasing the number of NCDDs from one to two could enhance binding affinity to the scaffoldin fragment. By ELISA, we found that the *Piromyces finnis* single dockerin domain had higher binding affinity than the double dockerin domain (Fig. 3c). However, by SPR, the double dockerin had a comparable K_d, but a higher R_{max} of 120 RU (Fig. 3g), suggesting that the double dockerin is indeed capable of binding to more sites on the ScaA fragment than the single dockerin. Thus, site specificity may be more subtly encoded in the different dockerin domains and cohesin repeats. The minimum sequence that defines a single cohesin remains to be determined, but it is clear from our study that fragments of the scaffoldin encoding as few as four repeats are sufficient for dockerin assembly. Additionally, we cannot rule out that additional binding factors (for example, glycosylation) found in native cellulosomes probably further modulate the fungal dockerin–cohesin interaction, which are lacking in this recombinant system.

It has previously been reported that dockerins are capable of binding to cellulosome fractions from other species of gut fungi, which is a marked departure from bacterial cellulosomes[4]. In agreement with this observation, a *Piromyces* dockerin is capable of binding to intact cellulosome fractions taken from *Anaeromyces* and *Neocallimastix* species (Supplementary Fig. 5). Thus, we tested whether this cross-species binding activity is encoded specifically within ScaA homologues. We purified single dockerin domains from all three genera of gut fungi and tested their ability to bind to all combinations of ScaA fragments. We observed binding for all combinations tested, and the binding signal was within standard error for almost all cases (Fig. 3d–f). Taken together, these results demonstrate that the fungal scaffoldin system is broadly conserved across the anaerobic fungal phylum, allowing for high interspecies infidelity. Therefore, it is not unreasonable to speculate that, in their native environments (for example, the dense microbial community of the herbivore rumen), fungal cellulosomes are a composite of enzymes from several species of gut fungi. This is in stark contrast to bacterial cellulosomes, which have high species specificity[17]. This promiscuity may confer a selective advantage of fungi over bacteria in these environments.

In addition to the ScaA orthologues, several more uncharacterized proteins that contain N-terminal secretion signals followed by stretches of repeating amino-acid sequence motifs similar to those found in ScaA were also detected via proteomic analysis (Supplementary Table 3). We hypothesized that these proteins also function as dockerin-binding scaffoldin proteins. We tested three of these putative scaffoldins (MycoCosm protein IDs: *Anaeromyces robustus* 296897, *Neocallimastix californiae* 673330 and *P. finnis* 124175) for dockerin binding activity and each tested positive over the empty vector control (Supplementary Fig. 6), suggesting that multiple scaffoldins probably exist in fungal cellulosomes. To comprehensively search for scaffoldin-like proteins throughout the genomes of these three organisms, we developed a hidden Markov model (HMM) based on the repeating motif from all six scaffoldins biochemically verified to interact with dockerins. We found 95 unique loci in the genomes of *A. robustus, P. finnis* and *N. californiae* that bear a signal peptide and at least 10 cohesin repeats (Supplementary Table 4 and Supplementary Fig. 3). Fewer loci (14) were detected in *P.* sp. E2 and *O.* sp. C1A due to fragmented genome assemblies (Supplementary Table 5). Significantly, no loci were found in prokaryotes (~2,000 genomes) and only one or two weak hits in other fungi (~400 genomes), demonstrating this HMM is highly specific to fungal scaffoldins. These results indicate that gut fungi probably produce multiple scaffoldins for cellulosome assembly and these scaffoldins represent a family of genes that is unique to the early branching anaerobic fungi.

Although fungal scaffoldins and their NCDD ligands are specific to gut fungi, many plant biomass-degrading enzymes that encode NCDDs are of bacterial origin, which has been noted previously for a limited subset of enzymes[14,18]. Indeed, all five gut fungal genomes sequenced to date have large numbers of genes that are

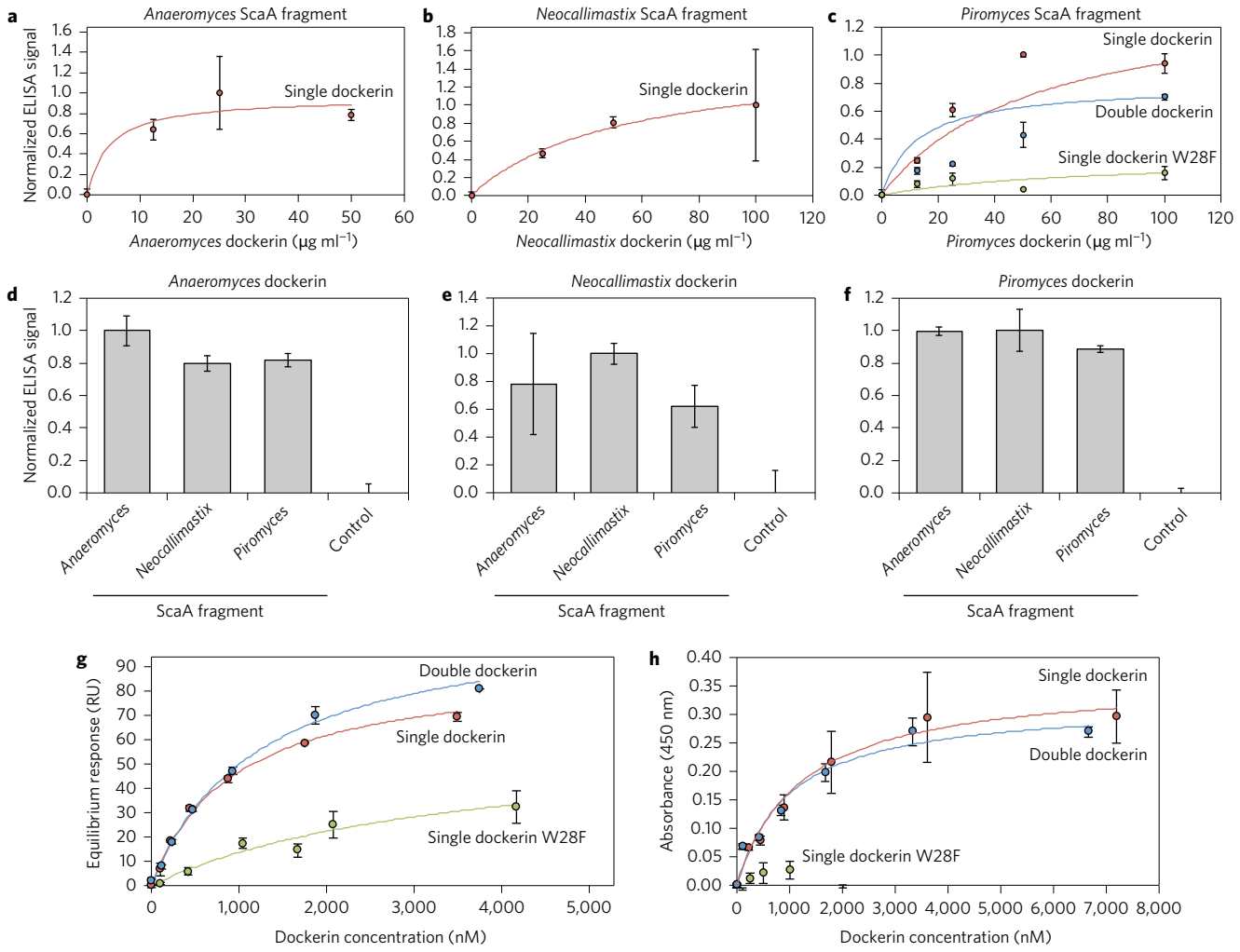

Figure 3 | Dockerins self-assemble onto scaffoldin fragments. a–c, Binding activity of purified dockerins to their endogenous ScaA fragments, as indicated. *E. coli* cells expressing ScaA fragments were lysed and coated on a 96-well ELISA plate and the binding activity of purified dockerin measured. Values were normalized after subtracting empty vector controls. **d–f**, Cross-species binding activity of purified dockerin to exogenous ScaA fragments, as indicated. Empty vector lysates were used as the control. **g**, Surface plasmon resonance of the dockerin–scaffoldin interaction. The binding activity of purified ScaA protein fragment from *P. finnis* was evaluated with a *P. finnis* single dockerin (wild-type and W28F) and a double dockerin, the same proteins as shown in **c**. Calculated $K_{d,app}$ (nM) and R_{max} (response units, RU): double dockerin, 1,230 ± 54, 111 ± 6.3; single dockerin, 944 ± 241, 91 ± 9; single dockerin W28F, 3,466 ± 3,649, 61 ± 50, respectively. Raw data traces for SPR are shown in Supplementary Fig. 8. **h**, Binding activity of purified dockerins to intact cellulosome measured by ELISA. The calculated $K_{d,app}$ (nM) values for double and single dockerin are 991 and 1,211, respectively. The single dockerin W28F yielded no $K_{d,app}$ due to negligible binding. All data points are presented as mean ± s.e.m of three technical replicates, with a line of best fit included where applicable.

more similar to bacterial than to eukaryotic genes (9–13%, Supplementary Fig. 7). We aligned 1,600 DDPs of the 5 anaerobic fungal genomes with 394 fungi currently deposited in JGI MycoCosm[19] (excluding Neocallimastigomycota), 1,774 bacteria and archaea in JGI Integrated Microbial Genomes (IMG)[20], 15 plants and green algae in JGI Phytozome[21], and 10 animals and 17 protists available on the JGI 'Tree of Life' (http://genome.jgi.doe.gov/). Of these DDPs, 644 aligned better with prokaryotic than eukaryotic proteins, and 117 aligned exclusively with prokaryotes. Conversely, only 22 aligned exclusively with fungi and 390 aligned better to fungi than to anything else. The remaining 417 DDPs did not align with anything. To determine whether this bacterial resemblance is the result of inter-kingdom horizontal gene transfer (HGT), we queried the domains that are fused to NCDDs to extract homologous sequences from the same bacterial and fungal genomes. When possible, we built phylogenetic trees of the domain sequences. Of 35 non-dockerin domains analysed, 10 (29%) passed our two criteria of (1) greater amino acid similarity to bacterial than to fungal sequences and (2) branching with bacterial rather than fungal sequences in the phylogenetic tree with >70% bootstrap support (Fig. 4a). The list of domains with an HGT signature includes nine CAZyme domains, as well as the spore coat domain (Supplementary Table 6). However, this analysis does not inform us as to the direction of any possible HGT events. Subjecting NCDDs to the same analysis showed that there are no similar sequences in IMG at all, suggesting that many DDPs may be fusions between native fungal and horizontally transferred bacterial components (Fig. 4b). Intriguingly, we found 12 fungal–bacterial homologue pairs where the bacterial protein is also a bacterial dockerin-domain protein. However, the sequence similarity between each pair of homologues encompasses only the catalytic domain and does not extend into the respective dockerin domains (Fig. 4c).

Over the past several decades, characterization of cellulosomes in fungi has been elusive, with multiple studies suggesting conflicting scaffolding schemes[5,22,23]. Here, next-generation sequencing combined with functional proteomics uncovered a family of genes that probably serve as scaffoldins in the cellulosomes of anaerobic fungi. The evidence for this is threefold. (1) Scaffoldins appear

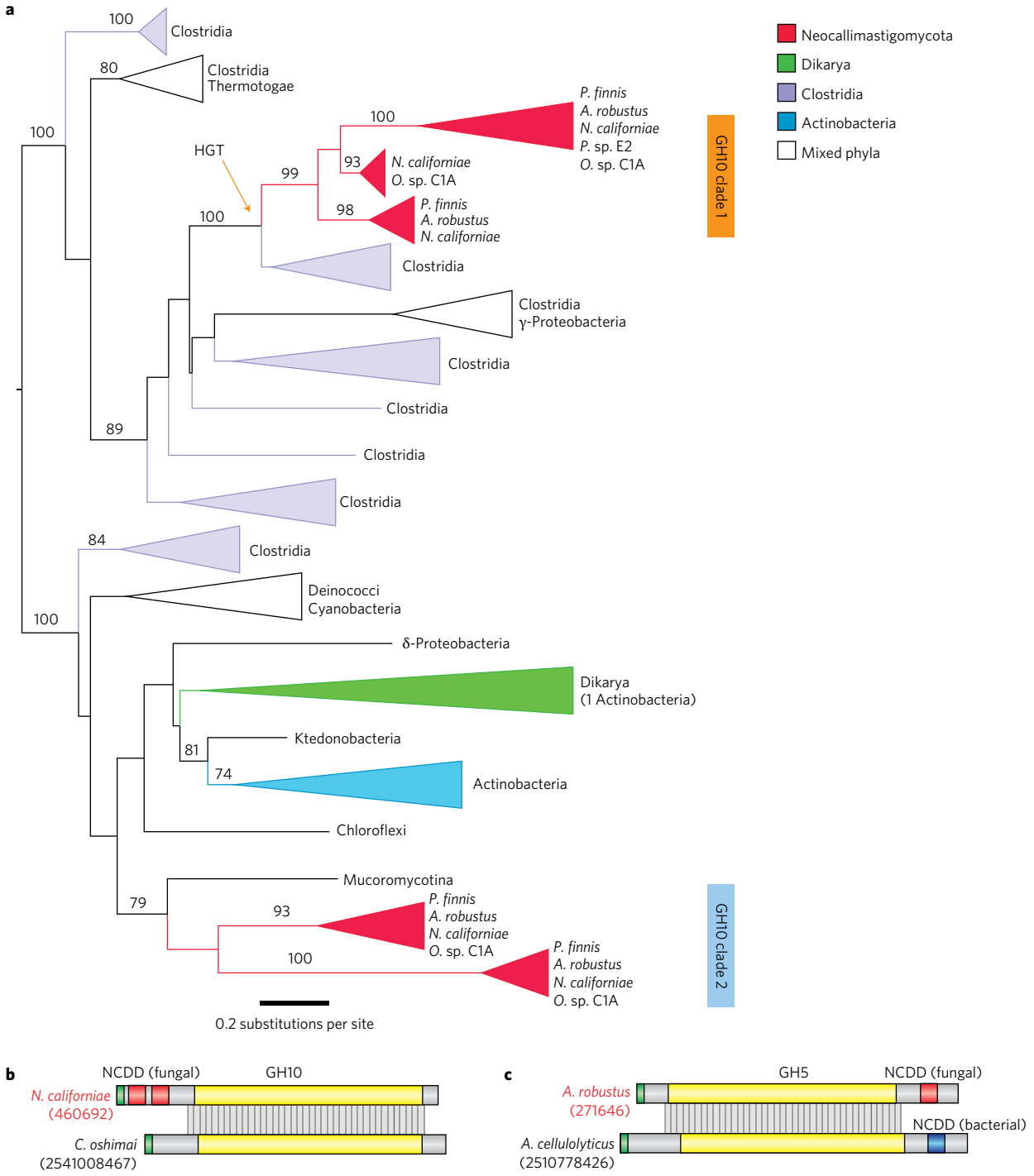

Figure 4 | CAZyme domains of fungal cellulosomes probably originated from bacteria by HGT. a, Example of a DDP non-NCDD Pfam domain with a HGT signature. The maximum likelihood tree is built from fungal and bacterial glycoside hydrolase 10 (GH10) sequences. The Neocallimastigomycota domains are from DDPs (red) and fall into two distinct clades 1 and 2. Clade 1 has the HGT signature but clade 2 does not. Nodes are labelled with bootstrap support where >70%. **b**, Example of a clade 1 Neocallimastigomycota GH10 DDP, MycoCosm ID Neosp1|460692 from *N. californiae*, compared to its closest bacterial homologue, IMG ID 2541008467 from *Caldicoprobacter oshimai*. Common elements are GH10 (yellow) and signal peptide (green). Only the fungal enzyme has NCDDs (red). The stretch of sequence similarity is indicated by vertical lines. **c**, Example of fungal and bacterial cellulases with similar cellulase domains but kingdom-specific cellulosome components. The fungal NCDD (red) has no sequence similarity to the bacterial NCDD (blue). The fungal protein is from *A. robustus* (MycoCosm ID Anasp1|271646), and the bacterial protein from *Acetivibrio cellulolyticus* (IMG ID 2510778426).

among the most represented proteins in supernatant and cellulosome fractions in three diverse isolates of gut fungi, and their amino-acid sequences encode hallmarks of a membrane-anchored scaffoldin molecule, including an N-terminal secretion motif, a C-terminal membrane anchor and a repeating amino-acid motif in between. These scaffoldins are encoded in all sequenced Neocallimastigomycota (representing four of the eight genera of gut fungi identified to date) and absent in other fungi. (2) Expression of repeat-containing scaffoldin fragments shows a robust interaction with purified dockerins by ELISA. (3) This dockerin–scaffoldin interaction is

biologically significant ($K_d \approx 0.9$ μM) as measured by SPR, whereas a mutated dockerin derivative significantly reduced binding activity. Taken together, the identification of a dockerin-binding protein scaffold from fungi opens the way for exploitation of this modular interaction for synthetic biology and substrate channelling. Finally, the powerful degradation activity of fungal cellulosomes is provided by the diverse functionality of their constituents, with 50 unique protein families of bacterial and fungal origin (Supplementary Table 2), and the assembly of these constituents onto scaffoldin molecules and into cellulosome-like complexes. Perhaps the most intriguing observation from this study is that fungal dockerins and their scaffoldin ligands have no sequence similarity to their bacterial counterparts. Thus, it is likely that the cellulosome-based strategy for plant cell wall degradation evolved in anaerobic gut fungi independently of bacteria. These observations suggest that co-localizing plant cell wall degrading enzymes at the cell surface is so important that nature has evolved cellulosomes on more than one occasion.

Methods

Strains, plasmids and growth conditions. All plasmids and strains used in this study are listed in Supplementary Table 7. *E. coli* Tuner (DE3) and *E. coli* BL21 (DE3) cells were used to express fungal proteins. DNA sequences encoding fungal dockerins were PCR-amplified from fungal cDNA libraries and cloned into the pET32a expression system (Addgene), which creates TrxA genetic fusions to promote protein solubility. An N-terminal Strep-tag (WSHPQFEK) was included on the forward primer during PCR amplification. DNA encoding ScaA fragments from *A. robustus* and *N. californiae* were cloned into pET32a. DNA encoding ScaA fragment from *P. finnis* was cloned into pET28a. Protein synthesis was induced when the cells reached an absorbance at 600 nm (A_{600}) of ~0.6 by adding 0.1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) to the medium. *E. coli* strains were routinely grown aerobically at 37 °C in lysogeny broth (LB) medium, and antibiotics were supplemented at the following concentrations: ampicillin (Amp, 100 μg ml^{-1}) and kanamycin (Kan, 50 μg ml^{-1}). Anaerobic fungal isolates were grown in Medium C, essentially as described elsewhere[6], using reed canary grass, corn stover or switch grass as a sole carbon source.

Transcriptome sequencing. Stranded cDNA libraries were generated using the Illumina TruSeq Stranded RNA LT kit. mRNA was purified from 1 μg of total RNA using magnetic beads containing poly-T oligos, fragmented and reverse-transcribed using random hexamers and SSII (Invitrogen), followed by second strand synthesis. The fragmented cDNA was treated with end-pair, A-tailing, adapter ligation and eight cycles of PCR. The prepared libraries were quantified using KAPA Biosystem's next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. The quantified libraries were then prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v.3 or v.4, and an Illumina cBot instrument to generate a clustered flowcell for sequencing. Sequencing of the flowcells was performed on the Illumina HiSeq 2000 or 2500 using HiSeq TruSeq SBS sequencing kits, following a 2 × 150 indexed run recipe. RNAseq reads were assembled *de novo* using Rnnotator v.3.4.0 (ref. 24) and used in genome annotation.

Genomic sequencing. Genomic DNA for *A. robustus*, *P. finnis* and *N. californiae* was isolated as described previously[25]. Genomes were sequenced using the PacBio platform and assembled together with Falcon (Pacific Biosciences), improved with FinisherSC[26], except for *N.californiae*, and polished with Quiver[27]. To prepare PacBio libraries, gDNA was treated with DNA damage repair mix followed by end repair and ligation of SMRT adapters using the PacBio SMRTbell Template Kit (PacBio). DNA was sheared to 10 kb fragments using the g-TUBE (Covaris), or templates were size-selected using the Sage Science BluePippin instrument with a 10 kb minimum cutoff. PacBio Sequencing primer was then annealed to the SMRTbell template library and sequencing polymerase was bound to it. The prepared SMRTbell template libraries were then sequenced on a Pacific Biosciences RSII sequencer using 4 h sequencing movie run times.

The genome of *P* sp. E2 was sequenced using a hybrid approach. Using a Newbler assembler[28], Sanger paired-end reads from five 2.2–6.5 kb insert size libraries were combined with Illumina data initially assembled with Velvet[29] and shredded into smaller fragments. The Newbler assembly was then further improved using GapResolution[30].

Genome annotation. All genomes were annotated using the JGI Annotation Pipeline and are available via the JGI fungal portal MycoCosm[19] (http://genome.jgi.doe.gov/neocallimastigomycota/). Genome assemblies and annotations were also deposited at GenBank under the following accession numbers: *A. robustus*, MCFG00000000; *N. californiae*, MCOG00000000; *P. finnis*, MCFH00000000.

Construction of a cohesin HMM. We built a HMM for a putative fungal cohesin domain based on multiple sequence alignments (MSAs) of six experimentally verified scaffoldin proteins from three Neocallimastigomycota ('Neo') species (Supplementary Fig. 4). Because scaffoldins are very large (1,000–6,000 amino acids) low-complexity proteins, we deployed an iterative optimizing method based on outputs of different MSA programs (MAFFT[31], MUSCLE[32] and CLUSTALW[33]). For a large number of local sub-alignments, a corresponding HMM was built using HMMER[34] and tested against a collection of 71 candidate scaffoldins previously identified by BLAST[35] and pattern searches. The HMM with the highest score was tested against 394 fungal genomes in MycoCosm (excluding Neo fungi) and did not yield any false positives. This HMM was then used to search all proteins translated from all gene predictions on the five Neo genomes.

Identification of Pfams with HGT signature. We developed an automatic pipeline for inferring potential HGT events and applied it to the non-NCDD domains of the DDPs. First, we collected all Neocallimastigomycota proteins with best BLASTp hits to bacteria (excluding top hits to other Neocallimastigomycota; minimum E-value of 1e-5), then excised each non-NCDD Pfam domain present in the Neocallimastigomycota ('Neo') DDP proteins and used that domain to extract the corresponding sequences from the proteins of 1,774 IMG genomes (one representative strain from each species) and 394 MycoCosm genomes (excluding Neo fungi). If the number of found domains exceeded 1,000, only the highest scoring 1,000 sequences were selected. Second, we aligned the non-NCDD domains and their selected homologues using MAFFT, with poorly aligned positions removed using trimAl[36]. For each Pfam, a phylogenetic tree was constructed using RAxML[37] with a PROTGAMMAAUTO model and 100 bootstrap replicates. Third, we identified those Pfams where (1) the IMG sequences had higher average Blastp scores with the Neo domains than did the MycoCosm sequences (excluding Neo sequences) and (2) the associated tree had nodes with bootstrap value >70% that displayed 'nestiness'. A node had 'nestiness' if a Neo leaf or subtree branched with a bacterial leaf or subtree and within an otherwise bacterial subtree, or vice versa (bacterial branches amongst Neo branches). Such Pfams were considered to display a 'HGT signature', suggesting that their inter-kingdom distribution might be explained by potential HGT events.

Protein purification and analysis. Cells expressing recombinant proteins were recovered after 16 h induction at 30 °C. Whole-cell lysate was prepared by centrifugation at 3,200*g* for 15 min in 50 ml conical tubes in a swinging bucket rotor and then cells were resuspended in 0.5 ml of 20 mM sodium phosphate, 300 mM sodium chloride, 10 mM imidazole (pH 7.4). Silica beads were added and the suspension was vortexed rigorously. The soluble supernatant was recovered by centrifugation and then target proteins encoding a 6xHis tag were purified by IMAC. Following elution of target proteins, the buffer was exchanged to PBS (pH 7.4) using Zeba desalting columns (Thermo Fisher Scientific). Protein concentration was measured using the BCA protein assay kit (Thermo Fisher Scientific).

Isolation of supernatant and cellulosome fractions. Supernatant and cellulosome fractions were collected between 72 and 96 h post inoculation. Cellulosomes were isolated essentially as described elsewhere[38]. Briefly, the vegetative growth was removed and 0.4% (wt/vol) SigmaCell type 50 (Sigma) was added to the supernatant and incubated with gentle agitation at 4 °C for 2 h. The cellulose was spun down, and washed once with 100 mM Tris-HCl (pH 7.5) containing 150 mM NaCl. The proteins were eluted by resuspending the cellulose in PBS (pH 7.4) for 1 h at room temperature. Secretome samples from *N. californiae, A. robustus* and *P. finnis* and cellulosome samples from *P. finnis* have been reported previously[7]. Cellulosome samples from *A. robustus* and *N. californiae* were new to this study.

Preparing samples for mass spectrometry. Secretome and cellulosome samples were buffer-exchanged to 50 mM ammonium bicarbonate using 3 kDa molecular weight cut off (MWCO) centrifugal filters (Millipore) according to the manufacturer's specifications. The secretome samples were treated with 8 M urea at 37 °C for 60 min and then washed four times with 8 M urea, 100 mM ammonium biocarbonate and then were buffer-exchanged to 100 mM ammonium bicarbonate in 500 μl 30 K MWCO centrifugal filter (Millipore). Cell pellet samples were transferred into four 2.0 ml BioPur tubes (Eppendorf) Approximately four 3 mm tungsten carbide TissueLyzer (Qiagen) beads were added and placed in pre-cooled TissueLyzer 24-position block and lysed on the TissueLyzer for 2 min at 30 oscillations per second. The samples were then centrifuged at 14,000*g*, 4 °C for 10 min, then the supernatant was transferred to a 5.0 ml cryovial for storage at −80 °C, and a final concentration of 7 M urea and 10 mM dithiothreitol (DTT) was added to each sample. The pellet was resuspended in 500 μl of 8 M urea, 5 mM DTT in 100 mM NH_4HCO_3 and transferred into a 15 ml conical centrifuge tube, leaving the beads behind. All of the samples were incubated at 60 °C for 30 min, then diluted eightfold with 100 mM buffer. A final concentration of 1 mM $CaCl_2$ was added. The protein concentrations for all samples were measured using the BCA protein assay kit (Thermo Fisher Scientific). Proteins in each fraction were digested with trypsin (1 unit trypsin per 50 units protein) and incubated at 37 °C for 3 h with shaking at 800 r.p.m. Samples were centrifuged at 5,525*g*, 4 °C for 15 min. The supernatant of each sample was cleaned for mass spectrometry (MS) analysis using a

100 mg ml^{-1} C18 solid phase extraction (SPE) column (Sigma Aldrich). Columns were conditioned with 3 ml of methanol and 2 ml of 0.1% trifluoroacetic acid (TFA). Samples were passaged through the columns and then the columns were washed with 4 ml of 95:5 H_2O:acetonitrile, 0.1% TFA. Collection tubes were placed under the dried columns and the peptides were eluted with 1 ml of 80:20 acetonitrile:H_2O, 0.1% TFA. The samples were concentrated in a speed-vac (ThermoFisher Scientific) to a volume of 50 µl. A BCA protein assay was performed on the samples, and the samples were diluted to 0.1 ug µl^{-1} and analysed via liquid chromatography-tandem mass spectrometry (LC-MS/MS).

Mass spectrometry. All data were collected on hybrid Velos linear ion trap coupled Orbitrap mass spectrometers (Thermo Electron) coupled to Waters NanoAcquity or Next-Gen 3 high-performance liquid chromatography systems (Waters Corporation) through 75 µm × 70 cm columns packed with Phenomenex Jupiter C-18 derivatized 3 µm silica beads (Phenomenex). Samples were loaded onto columns with 0.05% formic acid in water and eluted with 0.05% formic acid in acetonitrile over 100 min. Ten data-dependent MS/MS scans were recorded for each survey MS scan using a normalized collision energy of 35, an isolation width of 2.00 and a rolling exclusion window of +1.55/−0.55 Th lasting 60 s before previously fragmented signals were eligible for re-analysis. The MS/MS spectra from all LC-MS/MS data sets were converted to ASCII text (.dta format) using DeconMSn (http://www.ncbi.nlm.nih.gov/pubmed/18304935), which more precisely assigns the charge and parent mass values to an MS/MS spectrum. The data files were then interrogated using a target-decoy approach (http://www.ncbi.nlm.nih.gov/pubmed/20013364) using MSGFPlus (http://www.ncbi.nlm.nih.gov/pubmed/25358478), with a ±20 ppm parent mass tolerance, no specific digestion enzyme settings and a variable post-translational modification of oxidized methionine. All MS/MS search results for each data set were collated into tab-separated ASCII text files listing the best scoring identification for each spectrum. Collated search results were further combined into a single result file. These results were imported into a Microsoft SQL Server database. Results were filtered to 1% false discovery rate using an MSGF+ supplied *Q*-value that assesses reversed sequence decoy identifications for a given MSGF score across each data set. Using the protein references as a grouping term, unique peptides belonging to each protein were counted, as were all peptide-spectrum match (PSMs) belonging to all peptides for that protein (that is, a protein level observation count value). PSM observation counts were reported for each sample that was analysed. Cross-tabulation tables were created to enumerate protein-level PSM observations for each sample, allowing low-precision quantitative comparisons to be made. Protein sequences for each reported entry were subjected to BLAST analysis (version 2.2.28, ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.28/) using a combined collection of 2,784,909 fungal proteins reported in the Uniprot knowledgebase (ftp://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2014_09/knowledgebase/) and NCBI (ftp://ftp.ncbi.nih.gov/refseq/release/fungi/, combining all .faa files) as of September 2014. Results were imported into SQL Server and the highest-similarity identification (lowest E-value, highest percent similarity) extracted. As many of these matches resulted in an uncharacterized or hypothetical protein reference, the highest-similarity non-hypothetical entry (does NOT contain 'uncharacterized', 'predicted protein', 'hypothetical' or 'unplaced') was also extracted for each query, allowing for more useful biological inferences to be made. All BLAST results were imported into the Excel file and related to the cross-tabulated results to allow further investigation.

Size-exclusion chromatography. Intact cellulosomes were concentrated using five MWCO Amicon centrifugal filters (Milipore) and 100 µl was loaded onto a BioLogic DuoFlow System (Bio-Rad) with a HiPrep 16/60 Sephacryl S-500 High Resolution column (GE Healthcare). Fractions (6 ml) were collected and concentrated for enzyme activity assay and SDS–PAGE analysis.

Carboxymethyl cellulose assay for endoglucanase activity. The endoglucanase activity of cellulosome fractions was determined essentially as described elsewhere[39]. Briefly, 30 µl of cellulosome was added to 30 µl of 2% carboxymethyl cellulose in 0.1 M sodium acetate (pH 5.5). Reactions were performed at 39 °C for 24 h. A 60 µl volume of dinitrosalicylic acid (DNS) was added and the solution boiled at 95 °C for 5 min. Absorbance was measured at 540 nm. All reactions were performed in triplicate.

ELISA. *E. coli* lysates expressing scaffoldin fragments were diluted 1:2,000 in 0.05 M Na_2CO_3 buffer (pH 9.6) and 100 µl was coated on a 96-well microtitre ELISA plate at 4 °C overnight. Wells were then washed three times with 200 µl PBS (pH 7.4) and 100 µl of PBS containing 2% (wt/vol) BSA and 0.05% Tween-20 (vol/vol) was added and the plate incubated at 4 °C for 1 h. Purified dockerins were serially diluted in the same solvent (0–100 µg ml^{-1}) and were added to the plate and incubated at 4 °C for 1 h with gentle agitation. Wells were washed three times with PBS and then StepTactin (Bio-Rad), an horseradish peroxidase-conjugated secondary antibody against the Strep-tag, was diluted 1:5,000 in the same solvent and added to the plate and incubated at 4 °C for 1 h with gentle agitation. The wells were washed four times with PBS and signals were measured using TMB chromogen solution (Thermo Fisher Scientific) according to the manufacturer's instructions. All reactions were performed in triplicate and ELISA signals were normalized by total protein concentration of the lysate.

SPR. SPR analysis was performed using a BIACORE 3000 (GE Healthcare) equipped with a research-grade CM5 sensor chip. Pure scaffoldin fragment was immobilized using amine-coupling chemistry. The surfaces of flow cells 1 and 2 were activated for 7 min with a 1:1 mixture of 0.1 M NHS (*N*-hydroxysuccinimide) and 0.1 M EDC (3-(*N*,*N*-dimethylamino)propyl-*N*-ethylcarbodiimide) at a flow rate of 10 µl min^{-1}. Scaffoldin at a concentration of 10 µg ml^{-1} in 10 mM sodium acetate, pH 4.5, was immobilized at a density of 400 RU on flow cell 2, with flow cell 1 left blank as a reference surface. Both surfaces were blocked with a 7 min injection of 1 M ethanolamine, pH 8.5. To collect equilibrium binding data, pure single dockerin, single dockerin W28F and double dockerin suspended in 10 mM HEPES, 150 mM NaCl, 3 mM EDTA, 0.005% Tween-20, pH 7.4, were injected over the two flow cells for 10 min at a flow rate of 10 µl min^{-1}. The complex was allowed to associate for 600 s. The surfaces were regenerated with a 30 s injection of 10 mM glycine, pH 2.0 at a flow rate of 50 µl min^{-1}. Data were collected at a rate of 1 Hz. Nonlinear regression of the equilibrium response versus analyte concentration was performed using a custom MATLAB script. Analysis for each protein was performed in triplicate, with at least seven concentrations tested (one in technical duplicate). The purity of each protein was verified by SDS–PAGE prior to SPR analysis (Supplementary Fig. 9).

Data availability. Raw transcriptomic sequence data and transcriptomic profiles reported in this study have been deposited under BioProject accession no. PRJNA291757 (www.ncbi.nlm.nih.gov/bioproject/291757). Genomes are available via the JGI fungal portal MycoCosm[19] (http://genome.jgi.doe.gov/neocallimastigomycota/). Genome assemblies and annotations have been deposited at GenBank under the following accession numbers: *A. robustus*, MCFG00000000; *N. californiae*, MCOG00000000; *P. finnis*, MCFH00000000. Mass spectrometry proteomics data have been deposited at the ProteomeXchange Consortium via the PRIDE[40] partner repository with data set identifier PXD006325. All other data supporting the findings of this study are available from the corresponding author upon request.

Received 23 August 2016; accepted 25 April 2017; published 30 May 2017

References

1. Fontes, C. M. G. A. & Gilbert, H. J. Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates. *Annu. Rev. Biochem.* **79,** 655–681 (2010).
2. You, C., Myung, S. & Zhang, Y.-H. P. Facilitated substrate channeling in a self-assembled trifunctional enzyme complex. *Angew. Chem. Int. Ed.* **51,** 8787–8790 (2012).
3. Liu, F., Banta, S. & Chen, W. Functional assembly of a multi-enzyme methanol oxidation cascade on a surface-displayed trifunctional scaffold for enhanced NADH production. *Chem. Commun.* **49,** 3766–3768 (2013).
4. Fanutti, C. C., Ponyi, T. T., Black, G. W. G., Hazlewood, G. P. G. & Gilbert, H. J. H. The conserved noncatalytic 40-residue sequence in cellulases and hemicellulases from anaerobic fungi functions as a protein docking domain. *J. Biol. Chem.* **270,** 29314–29322 (1995).
5. Nagy, T. T. *et al.* Characterization of a double dockerin from the cellulosome of the anaerobic fungus *Piromyces equi*. *J. Mol. Biol.* **373,** 612–622 (2007).
6. Haitjema, C. H., Solomon, K. V., Henske, J. K., Theodorou, M. K. & O'Malley, M. A. Anaerobic gut fungi: advances in isolation, culture, and cellulolytic enzyme discovery for biofuel production. *Biotechnol. Bioeng.* **111,** 1471–1482 (2014).
7. Solomon, K. V. *et al.* Early-branching gut fungi possess a large, comprehensive array of biomass-degrading enzymes. *Science* **351,** 1192–1195 (2016).
8. Anderson, T. D. *et al.* Assembly of minicellulosomes on the surface of *Bacillus subtilis*. *Appl. Environ. Microb.* **77,** 4849–4858 (2011).
9. Tsai, S.-L., Oh, J., Singh, S., Chen, R. & Chen, W. Functional assembly of minicellulosomes on the *Saccharomyces cerevisiae* cell surface for cellulose hydrolysis and ethanol production. *Appl. Environ. Microb.* **75,** 6087–6093 (2009).
10. Tsai, S.-L., Dasilva, N. A. & Chen, W. Functional display of complex cellulosomes on the yeast surface via adaptive assembly. *ACS Synth. Biol.* **2,** 14–21 (2013).
11. Wilson, C. A. & Wood, T. M. The anaerobic fungus *Neocallimastix frontalis*—isolation and properties of a cellulosome-type enzyme fraction with the capacity to solubilize hydrogen-bond-ordered cellulose. *Appl. Microbiol. Biotechnol.* **37,** 125–129 (1992).
12. Nguyen, K. B. *et al.* Phosphorylation of spore coat proteins by a family of atypical protein kinases. *Proc. Natl Acad. Sci. USA* **113,** E3482–E3491 (2016).
13. Steenbakkers, P. J. M. *et al.* beta-Glucosidase in cellulosome of the anaerobic fungus *Piromyces* sp. strain E2 is a family 3 glycoside hydrolase. *Biochem. J.* **370,** 963–970 (2003).
14. Youssef, N. H. *et al.* The genome of the anaerobic fungus *Orpinomyces* sp. strain C1A reveals the unique evolutionary history of a remarkable plant biomass degrader. *Appl. Environ. Microb.* **79,** 4620–4634 (2013).
15. Shah, D. S., Joucla, G., Remaud-Simeon, M. & Russell, R. R. B. Conserved repeat motifs and glucan binding by glucansucrases of oral streptococci and *Leuconostoc mesenteroides*. *J. Bacteriol.* **186,** 8301–8308 (2004).

16. Raghothama, S. *et al.* Characterization of a cellulosome dockerin domain from the anaerobic fungus *Piromyces equi*. *Nat. Struct. Biol.* **8,** 775–778 (2001).
17. Bayer, E. A., Belaich, J.-P., Shoham, Y. & Lamed, R. The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annu. Rev. Microbiol.* **58,** 521–554 (2004).
18. Garcia-Vallve, S., Romeu, A. & Palau, J. Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. *Mol. Biol. Evol.* **17,** 352–361 (2000).
19. Grigoriev, I. V. *et al.* Mycocosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42,** D699–D704 (2014).
20. Markowitz, V. M. *et al.* IMG/m 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res.* **42,** D568–D573 (2014).
21. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40,** D1178–D1186 (2012).
22. Steenbakkers, P. J. P. *et al.* Noncatalytic docking domains of cellulosomes of anaerobic fungi. *J. Bacteriol.* **183,** 5325–5333 (2001).
23. Gilmore, S. P., Henske, J. K. & O'Malley, M. A. Driving biomass breakdown through engineered cellulosomes. *Bioengineered* **6,** 204–208 (2015).
24. Martin, J. *et al.* Rnnotator: an automated *de novo* transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics* **11,** 663 (2010).
25. Solomon, K. V., Henske, J. K., Theodorou, M. K. & O'Malley, M. A. Robust and effective methodologies for cryopreservation and DNA extraction from anaerobic gut fungi. *Anaerobe* **38,** 39–46 (2016).
26. Lam, K. K., LaButti, K., Khalak, A. & Tse, D. FinisherSC: a repeat-aware tool for upgrading *de novo* assembly using long reads. *Bioinformatics* **31,** 3207–3209 (2015).
27. Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10,** 563–569 (2013).
28. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005); corrigendum **441,** 120 (2006).
29. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
30. Trong, S. *et al.* Gap resolution: a software package for improving newbler genome assemblies. in *Proceedings of the 4th Annual Meeting on Sequencing Finishing, Analysis in the Future* 35 (2009).
31. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30,** 3059–3066 (2002).
32. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
33. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
34. Eddy, S. R. Multiple alignment using hidden Markov models. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3,** 114–120 (1995).
35. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
36. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25,** 1972–1973 (2009).
37. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
38. Ali, B. R. S. *et al.* Cellulases and hemicellulases of the anaerobic fungus *Piromyces* constitute a multiprotein cellulose-binding complex and are encoded by multigene families. *FEMS Microbiol. Lett.* **125,** 15–21 (1995).
39. Xiao, Z., Storms, R. & Tsang, A. Microplate-based carboxymethylcellulose assay for endoglucanase activity. *Anal. Biochem.* **342,** 176–178 (2005).
40. Vizcaino, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44,** D447–D456 (2016).

Acknowledgements

The authors acknowledge funding support from the Office of Science (BER), US Department of Energy (DE-SC0010352), the US Department of Agriculture (award no. 2011-67017-20459), the National Science Foundation (DGE 1144085) and the Institute for Collaborative Biotechnologies through grant no. W911NF-09-0001 from the US Army Research Office. A portion of this research was performed under the Facilities Integrating Collaborations for User Science (FICUS) exploratory effort and used resources at the DOE Joint Genome Institute and the Environmental Molecular Sciences Laboratory, which are DOE Office of Science User Facilities. Both facilities are sponsored by the Office of Biological and Environmental Research and operated under contract nos. DE-AC02-05CH11231 (JGI) and DE-AC05-76RL01830 (EMSL). The authors acknowledge support from the California NanoSystems Institute (CNSI), supported by the University of California, Santa Barbara, and the University of California, Office of the President. SPR data were generated in the UCSB and UCOP-supported Biological Nanostructures Laboratory within the California NanoSystems Institute. The authors thank P.J. Weimer (US Dairy Forage Research Center) for lignocellulosic substrates. B.H. acknowledges IDEX Aix-Marseille (Grant Microbio-E) and Agence Nationale de la Recherche (grant no. ANR-14-CE06-0020) for funding.

Author contributions

C.H.H., S.P.G. and M.A.O. planned the experiments. C.H.H. and R.D. performed ELISA and S.P.G. performed SPR experiments. C.H.H., S.P.G., A.K. and M.A.O. wrote the manuscript. H.M.B., S.O.P. and A.T.W. performed proteomic analyses. K.V.S. and J.K.H. prepared and analysed genomic samples for *N. californiae*, *P. finnis* and *A. robustus*. B.B., T.v.A. and J.H.P.H. prepared and analysed genomic samples for *Piromyces* sp. E2. Z.Z. and J.C. sequenced, K.L. assembled, and A.K., S.J.M. and A.A.S. annotated and analysed genomes. B.H. and M.H. analysed and classified carbohydrate-active enzymes. M.A.O., S.E.B., K.B. and I.V.G. coordinated genome projects at JGI.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.A.O.

How to cite this article: Haitjema, C. H. *et al.* A parts list for fungal cellulosomes revealed by comparative genomics. *Nat. Microbiol.* **2**, 17087 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Competing interests

The authors declare no competing financial interests.